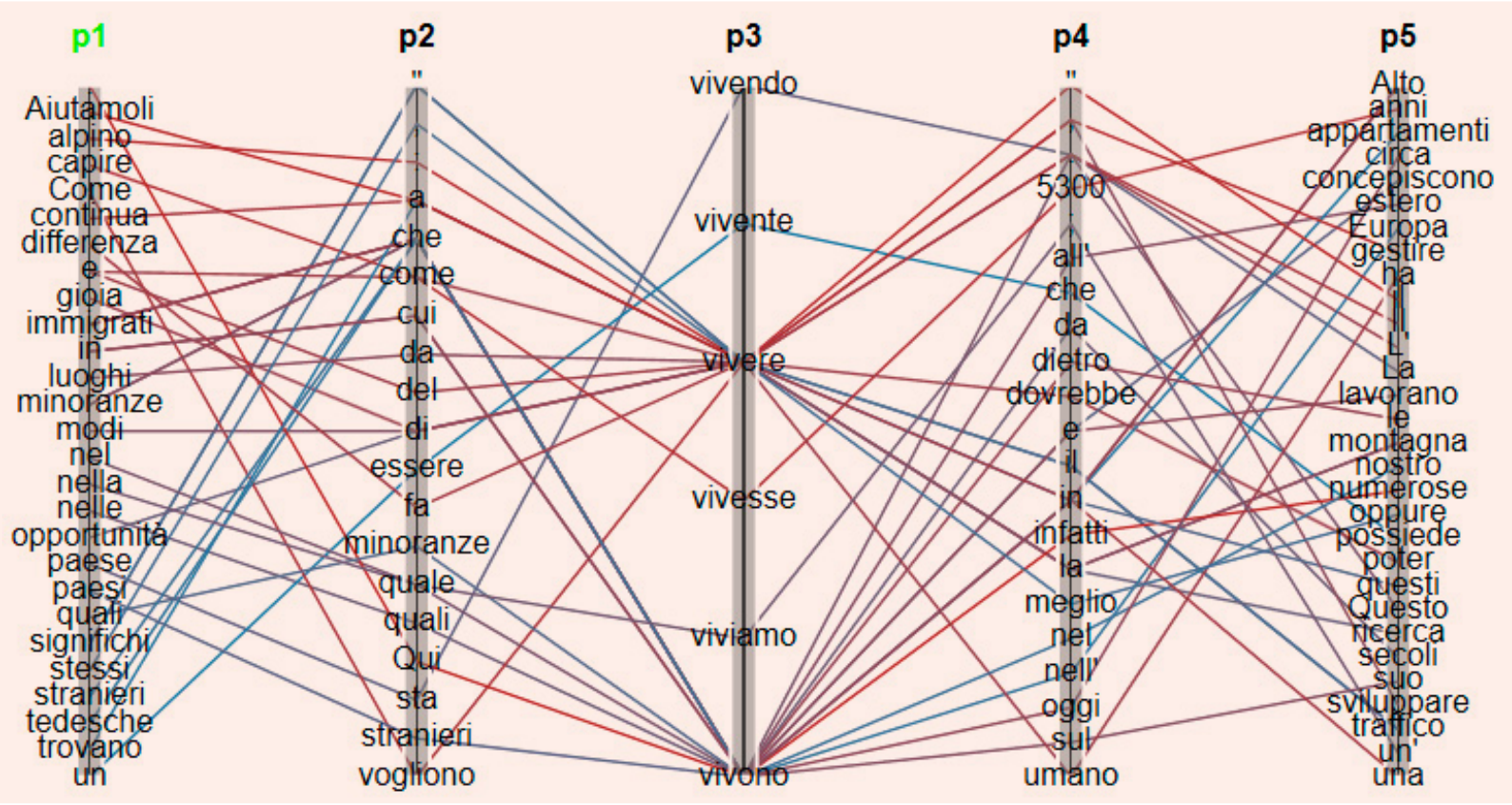
EURAC research

EURALEX

EURALEX 2014



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

Part 2

Index

Part 1 **Plenary Lectures** **23**

From Lexicography to Terminology: a Cline, not a Dichotomy 25
Thierry Fontenelle

Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries 47
Ulrich Heid

Using Mobile Bilingual Dictionaries in an EFL Class 63
Carla Marello

Meanings, Ideologies, and Learners' Dictionaries 85
Rosamund Moon

The Dictionary-Making Process **107**

The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary 109
Tressy Arts

Simple and Effective User Interface for the Dictionary Writing System 125
Kamil Barbierik, Zuzana Děngeová, Martina Holcová Habrová, Vladimír Jarý, Tomáš Liška, Michaela Lišková, Miroslav Virius

Totalitarian Dictionary of Czech 137
František Čermák

Dictionary of Abbreviations in Linguistics: Towards Defining Cognitive Aspects as Structural Elements of the Entry 145
Ivo Fabijancić

La definizione delle relazioni intra- e interlinguistiche nella costruzione dell'ontologia IMAGACT 159
Gloria Gagliardi

A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition 169
Ellert Thor Johannsson, Simonetta Battista

Making a Learner's Dictionary of Academic English 181
Diana Lea

The Danish Thesaurus: Problems and Perspectives 191
Sanni Nimb, Lars Trap-Jensen, Henrik Lorentzen

From a Dialect Dictionary to an Etymological One 201
Vilja Oja, Iris Metsmägi

Research on Dictionary Use **211**

Wörterbuchbenutzung: Ergebnisse einer Umfrage bei italienischen DaF-Lernern 213
Carolina Flinz

Translation, Cultural Adaptation and Preliminary Psychometric Evaluation of the English Version of "Strategy Inventory for Dictionary Use" (S.I.D.U) 225
Gavriilidou Zoe

The Authentic Voices of Dictionary Users – Viewing Comments on an Online Learner's Dictionary Before and After Revision 237
Ann-Kristin Hult

Mobile Lexicography: A Survey of the Mobile User Situation 249
Henrik Køhler Simonsen

Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht 263
Martina Nied Curcio

Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses 281
Sascha Wolfer, Alexander Koplenig, Peter Meyer, Carolin Müller-Spitzer

La performance dell'utente apprendente di italiano LS/L2 e la microstruttura dei dizionari: sussidi per lo sviluppo della Lessicografia Pedagogica 291
Angela Maria Tenório Zucchi

Lexicography and Language Technologies 303

ALIQUOT – Atlante della Lingua Italiana QUOTidiana 305
Michele Castellarin, Fabio Tosques

Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples 319
Paul Cook, Michael Rundell, Jey Han Lau, and Timothy Baldwin

Cross-linking Austrian dialectal Dictionaries through formalized Meanings 329
Thierry Declerck, Eveline Wandl-Vogt

Nutzung des DWDS-Wortprofils beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen 345
Jörg Didakowski, Nadja Radtke

Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies 355
Iztok Kosem, Polona Gantar, Nataša Logar, Simon Krek

A Corpus-based Dictionary of Polish Sign Language (PJM) 365
Jadwiga Linde-Usiekniewicz, Małgorzata Czajkowska-Kisil, Joanna Łacheta, Paweł Rutkowski

Laying the Foundations for a Diachronic Dictionary of Tunis Arabic: a First Glance at an Evolving New Language Resource 377
Karlheinz Mörth, Stephan Procházka, Ines Dallaji

BabelNet meets Lexicography: the Case of an Automatically-built Multilingual Encyclopedic Dictionary 389
Roberto Navigli

A Simple Platform for Defining Idiom Variation Matching Rules 399
Koichi Takeuchi, Ulrich Apel, Ray Miyata, Ryo Murayama, Ryoko Adachi, Wolfgang Fanderl, Iris Vogel, Kyo Kageura

Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon ... 405
Jana Šindlerová, Kateřina Veselovská, Jan Hajič jr.

Part 2 **Lexicography and Corpus Linguistics** ... **415**

Compatible Sketch Grammars for Comparable Corpora ... 417
Vladimír Benko

From GLÀFF to PsychoGLÀFF: a Large Psycholinguistics-oriented French Lexical Resource ... 431
Basilio Calderone, Nabil Hathout, Franck Sajous

RIDIRE. Corpus and Tools for the Acquisition of Italian L2 ... 447
Alessandro Panunzi, Emanuela Cresti, Lorenzo Gregori

Empirical Approaches to German Paronyms ... 463
Petra Storjohann, Ulrich Schnörch

Pragmatic Meaning in Lexicographical Description: Semantic Prosody on the Go ... 477
Mojca Šorli

Bi-and Multilingual Lexicography ... **493**

Linking a Dictionary to Other Open Data – Better Access to More Specific Information for the Users ... 495
Ulrich Apel

Bilingual Word Sketches: the *translate* Button ... 505
Vít Baisa, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý

Creating a Bilingual Italian-English Dictionary of Collocations ... 515
Barbara Berti, Laura Pinnavaia

La valencia del adjetivo en diccionarios bilingües alemán-español-alemán ... 525
Andreu Castell, Natàlia Català, María Bargalló

Esame storico dei "realia" nei dizionari bilingui italiano/ungheresi ... 537
Zsuzsanna Fábián

Quello che i dizionari possono fare: l'esempio dei Dizionari di Tedesco (Giacoma/Kolb – Zanichelli/Klett) ... 551
Luisa Giacoma

Bilingual Dictionary Drafting. The Example of German-Basque, a Medium-density Language Pair ... 563
David Lindemann, Iker Manterola, Rogelio Nazar, Iñaki San Vicente, Xabier Saralegi

Illustrative Examples and the Aspect of Culture: The Perspective of a Tshivenḓa Bilingual Dictionary ... 577
Munzhedzi James Mafela

Corpus, Parallélisme et Lexicographie Bilingue ... 587
Adriana Zavaglia, Gisele Galafacci

Lexicography for Specialised Languages, Technology and Terminography **599**

EcoLexicon 601
Pamela Faber, Miriam Buendía Castro

Experts and Terminologists: Exchanging Roles in the Elaboration of the Terminological Dictionary of the Brenner Base Tunnel (BBT) 609
Elena Chiocchetti, Natascia Ralli

Cloud Terminology Services Facilitate Specialised Lexicography Work 621
Tatiana Gornostay, Andrejs Vasiļjevs

Good Contexts for Translators–A First Account of the Cristal Project 631
Amélie Josselin-Leray, Cécile Fabre, Josette Rebeyrolle, Aurélie Picton, Emmanuel Planas

Kontextbasierte lexikalische Kontrolle von Anforderungsdokumenten 647
Jennifer Krisch

From Term Dynamics to Concept Dynamics: Term Variation and Multidimensionality in the Psychiatric Domain 657
Pilar León-Araúz, Arianne Reimerink

Bon usage vs. Fachliches: Fachsprache in der Geschichte der französischen Sprachpflege und Lexikographie 669
Martina Mayer

EU-Terminologie in den einsprachigen Wörterbüchern des Deutschen 685
Diana Stantcheva

Text Boxes as Lexicographic Device in LSP Dictionaries 697
Elsabé Taljard, Danie J. Prinsloo, Rufus H. Gouws

Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources 707
Izabella Thomas, Blandine Plaisantin Alecu, Bérenger Germain, Marie-Laure Betbeder

Station Sensunique: une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques (orientée Langues Contrôlées) 727
Izabella Thomas, Blandine Plaisantin Alecu, Bérenger Germain, Marie-Laure Betbeder

Lexicography of Lesser Used Languages **737**

Towards an Integrated E-Dictionary Application – The Case of an English to Zulu Dictionary of Possessives 739
Sonja Bosch, Gertrud Faaß

Zur (Vor-)Geschichte der saamischen Lexikografie: ein lateinisch-saamisches Wörterverzeichnis aus dem 17. Jahrhundert 749
Eino Koponen

Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses 767
Gabriele Langer, Susanne König, Silke Matthes

Dearcadh na nDéise – Representations of Gaeltacht na nDéise in Dineen's Bilingual Irish-English Dictionary (1927) 787
Chris Mulhall, Seamus ó'Diollúin

The *eLexicon Mediae et Infimae Latinitatis Polonorum.* The Electronic Dictionary of Polish Medieval Latin 793
Krzysztof Nowak

From DANTE to Dictionary: The New English-Irish Dictionary 807
Pádraig Ó Mianáin, Cathal Convery

User Support in e-Dictionaries for Complex Grammatical Structures in the Bantu Languages 819
Danie J. Prinsloo, Theo J.D. Bothma, Ulrich Heid

Concerning the Treatment of Co-existent Synonyms in Estonian Monolingual and Bilingual Dictionaries 829
Enn Veldi

Part 3 **Phraseology and Collocation** **837**

Unusual Phrases in English MLDs: Increasing User Friendliness 839
Stephen Coffey

Harvesting from One's Own Field: A Study in Collocational Resonance 855
Janet DeCesaris, Geoffrey Williams

The Use of Corpora in Bilingual Phraseography 867
Dmitrij Dobrovol'skij

Comparing Phraseologisms: Building a Corpus-Based Lexicographic Resource for Translators 885
Laura Giacomini

Lexical Variation within Phraseological Units 893
Tarja Riitta Heinonen

Prendere il toro per le corna o *lasciare una bocca amara*? – The Treatment of Tripartite Italian Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries 905
Chris Mulhall

Especialización y Prototipicidad en Binomios N y N 915
Ignacio Rodríguez Sánchez

Syntax and Semantics vs. Statistics for Italian Multiword Expressions: Empirical Prototypes and Extraction Strategies 927
Luigi Squillante

Historical Lexicography and Etymology **939**

Il DiVo (Dizionario dei Volgarizzamenti). Un archivio digitale integrato per lo studio del lessico di traduzione nell'italiano antico 941
Diego Dotto

Informatiser le Französisches etymologisches Wörterbuch: la nécessaire prise en compte de l'utilisateur 955
Pascale Renders, Esther Baiwir

A Morphological Historical Root Dictionary for Portuguese 967
João Paulo Silvestre, Alina Villalva

Lexicological Issues of Lexicographical Relevance **979**

What can Lexicography Gain from Studies of Loanword Perception and Adaptation? 981
Mirosław Bańko, Milena Hebal-Jezierska

Pejorative Language Use in the Satirical Journal "Die Fackel" as documented in the "Dictionary of Insults and Invectives" 993
Hanno Biber

The Presence of Gender Issues in Spanish Dictionaries 1001
Ana Costa Pérez

Reflexive Verbs in a Valency Lexicon: The Case of Czech Reflexive Morphemes 1007
Václava Kettnerová, Markéta Lopatková

Polysemous Models of Words and Their Representation in a Dictionary Entry 1025
Tinatin Margalitadze

One Lexicological Theory, two Lexicographical Models and the Pragmatemes 1039
Lena Papadopoulou

Analyzing Specialized Verbs in a French-Italian-English Medical Corpus: A Frame-based Methodology 1049
Anna Riccio

Neoclassical Formatives in Dictionaries 1059
Pius ten Hacken, Renáta Panocová

Reports on Lexicographical and Lexicological Projects **1073**

Revision and Digitisation of the Early Volumes of *Norsk Ordbok*: Lexicographical Challenges 1075
Sturla Berg-Olsen, Åse Wetås

A Dictionary Guide for Web Users 1087
Valeria Caruso, Anna De Meo

What a Multilingual Loanword Dictionary can be used for: Searching the *Dizionario di italianismi in francese, inglese, Tedesco* (DIFIT) 1099
Matthias Heinz, Anne-Kathrin Gärtig

The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian 1109
Jelena Kallas, Maria Tuulik, Margit Langemets

Die fremdsprachige Produktionssituation im Fokus eines onomasiologisch konzeptuell orientierten, zweisprachig-bilateralen Wörterbuches für das Sprachenpaar Deutsch - Spanisch: Theoretische und methodologische Grundlagen von DICONALE 1119
Meike Meliss

Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries 1135
Peter Meyer

At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project) 1145
Pavla Kochová, Zdeňka Opavská, Martina Holcová Habrová

Frame Semantics and Learner's Dictionaries: Frame Example Sections as a New Dictionary Feature 1153
Carolin Ostermann

Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology 1163
Alessandro Panunzi, Irene De Felice, Lorenzo Gregori, Stefano Jacoviello, Monica Monachini, Massimo Moneglia, Valeria Quochi, Irene Russo

Degrees of Synonymity as the Basis of a Network for German Communication Verbs in the Online Reference Work *Kommunikationsverben* in OWID 1171
Kristel Proost, Carolin Müller-Spitzer

Job-hunting in Italy: Building a glossary of "English-inspired" job titles 1187
Virginia Pulcini, Angela Andreani

A Small Dictionary of Life under Communist Totalitarian Rule (Czechoslovakia 1948-1989) 1203
Věra Schmiedtová

A Frequency Dictionary of Dutch 1211
Carole Tiberius, Tanneke Schoonheim, Adam Kilgarriff

The Corpus of the Croatian Church Slavonic Texts and the Current State of Affairs Concerning the Dictionary of the Croatian Redaction of Church Slavonic Compiling 1221
Vida Vukoja

Others **1237**

Considerations about Gender Symmetry in the Dictionary of Bavarian Dialects in Austria 1239
Isabella Flucher, Eveline Wandl-Vogt, Thierry Declerck

Advancing Search in the Algemeen Nederlands Woordenboek 1247
Carole Tiberius, Jan Niestadt, Lut Colman, Boudewijn van den Berg

Lexicography and Corpus Linguistics

Compatible Sketch Grammars for Comparable Corpora

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Comenius University in Bratislava, UNESCO Chair in Translation Studies
vladob@juls.savba.sk

Abstract

Our paper describes an on-going experiment aimed at creating a family of billion-token web corpora that could to a large extent deserve the designation "comparable": corpora are of the same size, data gathered by crawling the web at (approximately) the same time, containing similar web-specific domains, genres and registers, further pre-processed, filtered and deduplicated by the same tools, morphologically annotated by (possibly) the same tagger and made available via Sketch Engine. To overcome the problem of great differences in the existing sketch grammars for the respective languages, a set of "compatible" sketch grammars have been written that will aid contrastive linguistic research and bilingual lexicographic projects. The sketch grammars use a uniform set of rules for all word categories (parts of speech) and the resulting set of tables is displayed in a fixed order in all languages.
Keywords: comparable web corpora; sketch grammars; bilingual lexicography

1 Introduction

Ten years after its introduction to the lexicographic community at the Lorient Euralex Congress (Kilgarriff et al., 2004), Sketch Engine (*SkE*) has become a standard tool in numerous lexicographic projects, as well as in various areas of corpus-based linguistic research. Sketch grammars for corpora in many languages have been written (cf. References). Recently published open-source tools for efficient web crawling (Suchomel a Pomikálek, 2012) stimulate the building of very large web corpora, the analysis of which is hardly imaginable without a powerful summarisation machine such as *SkE*. Newly implemented *SkE* support for parallel and comparable corpora (Kovář, 2013) facilitate its use in the area of bilingual lexicography and contrastive linguistic research.

In bilingual and multilingual linguistic work with *SkE*, we often encounter the problem of sketch grammars defining the collocational profiles of a headword and its translation equivalent for the respective languages. Those sketch grammars have often been created for different purposes, having in mind different user requirements, with resulting word sketches being rather disparate, making its use for contrastive linguistics problematic. Our paper suggests an alternative approach to the creation of sketch grammars, within the framework of which the respective grammars can be made compatible for (almost) all languages.

2 The Aranea project

2.1 Why new corpora?

Besides our interest in testing the new corpus-building tools, the motive for starting a new corpus project was the lack of suitable corpora that could be used by students of foreign languages and translation studies at our university. The existing web corpora families that are available for download do not cover all the languages needed. As for corpora stored at the *SkE* web site[1], they (1) are not available for download, (2) are mostly too large for classroom use[2], and (3) have too different sketch grammars, which makes them difficult to use in a mixed-language classroom.

We expect that a set of corpora for several languages of equal size and built by a standardized methodology can not only be used for teaching purposes, but also in other areas of linguistic research (contrastive studies) and in lexicography (both mono- and bilingual).

2.2 The names

For our corpora, we have decided to use "language-neutral" Latin names denoting the language of the texts and the corpus size. The whole corpus family is called *Aranea*, and the respective members bear the appropriate language name, e.g. *Araneum Anglicum*, *Araneum Francogallicum*, *Araneum Russicum* for English, French, and Russian, respectively, etc.

Each corpus exists in several editions, differing by their sizes. The basic (medium-sized) version, *Maius* ("greater"), contains approximately 1.2 billion tokens (i.e., over 1 billion words). This size can be reached relatively quickly for all participating languages, and for the "large" ones with plenty of web data available, it usually takes just one or two days of download time. The 10% random sample of *Maius*, called *Minus* ("smaller"), is to be used for teaching purposes (e.g. for lessons in the framework of Corpus Linguistics programmes for students of foreign languages and translation studies). A 1% sample, *Minimum* ("minimal"), is not intended to be used directly by the end users, and is utilized in debugging the processing pipelines and tuning the sketch grammars. And lastly, the largest *Maximum* ("maximal") edition will contain as much data as can be downloaded from the web for the particular language, and its size is mostly determined by the configuration of the server.

1 http://www.sketchengine.co.uk/

2 According to our experience, the ideal corpus for teaching corpus linguistics is about BNC-sized, i.e. it contains some 100 million tokens. As it is not easy to prevent students from invoking search operations taking several minutes to evaluate, billion-plus token corpora proved to be quite unsuitable for teaching purposes.

2.3 Web crawling

The source data acquisition is being performed by means of *SpiderLing*[3], a web crawler optimized for collecting textual data from the web. The system contains an integrated character encoding (*chared.py*) and language recognition (*trigrams.py*) module, as well as a tool for boilerplate removal (jusText). The input seed URLs (some 1,000 for each language) have initially been harvested by BootCAT[4] (Baroni and Bernardini; 2004).

Several input parameters of the crawling process are to be set by the user, most notably the language name, a file containing sample text in the respective language (to produce a model for language recognition), a language similarity threshold (a value between 0 and 1; default 0.5), the number of parallel processes, and the crawling time.

In our processing, we usually crawled the web in 24-hour slots (the process could then be re-started) with all other values set to defaults. The only exception was crawling for Slovak and Czech, where we used 7-day slots, as the process was much slower here. The language similarity threshold also had to be changed in case of Slovak and Czech. As these languages are fairly similar, the trigram method did not seem to be able to distinguish between them sufficiently. We have therefore increased the similarity threshold value to 0.65 (saving many "good" documents, and causing many "wrong" ones to pass the filter) and removed the unwanted texts by subsequent filtration based on character frequencies .

2.4 Post-download processing

Besides the basic filtration aimed to remove texts with incorrect or misinterpreted character encoding, missing diacritics and texts with non-standard proportion of punctuation and uppercase characters, the main processing operation in this phase is tokenization. As the tokenization strategy has to be compatible with that of the corpus used to train the tagger, we decided to use the tokenizers supplied with Tree Tagger and TaKIPI for the respective languages. In the future, we would like to make use of the *unitok.py* universal tokenizing program developed at Masaryk University in Brno (Jakubíček; 2014).

2.5 Deduplication

The whole procedure (Benko; 2013) consists of three stages. The first stage detects near-duplicate documents by means of the Onion (Pomikálek; 2012) utility (similarity threshold 0.95), and the duplicate documents are deleted. The second stage deduplicates the remaining text at the paragraph level using the same procedure and settings. The tokens of the duplicate paragraphs, however, are not deleted but rather they are marked to make them "invisible" during corpus searches, while they can be displayed

3 http://nlp.fi.muni.cz/trac/spiderling
4 http://bootcat.sslmit.unibo.it/

as context at the boundary of non-duplicate and duplicate text. In the last stage, we make use of our own tool based on the fingerprint method (with ignoring punctuation, special graphics characters and digits) to deduplicate the text at the sentence level. The tokens of duplicate sentences are marked similarly to the previous stage. This last step can "clean up" the duplicities among the short segments that fail to be detected as duplicates by Onion.

As deduplication is beyond the scope of our paper, we only mention here that the process has typically removed some 20–45% of tokens in the *Maius* versions of our corpora

2.6 Morpho-syntactic annotation

For languages covered by the parameter files of Tree Tagger (Schmid; 1994), this tagger has been used to annotate the respective corpora. For Polish, the TaKIPI (Piasecki; 2007), and for Czech, the Morče (Hajič; 2004) taggers were used, respectively. The question of tools for tagging Hungarian and Ukrainian data has not been resolved yet.

2.7 Tagging-related filtration

To improve the precision of tag assignments, a series of pre- and post-tagging filters have been developed that fix issues introduced by Unicode encoding of the source text[5]. The filtration fixes known tagger issues for the respective languages, namely the misassigned tags for many punctuation and special graphic characters (that are often tagged as nouns, adjectives, or abbreviations, and sometimes even as verbs with subcategories). For some languages, an additional tag with masked subcategories for gender and number is created, that is later used by some rules within the respective sketch grammars.

2.8 Current state of the project

At present, eight language versions of the *Aranea* corpus family have been created, containing both *Maius* and *Minus* editions as follows (in chronological order): *Araneum Russicum* (Russian), *Araneum Francogallicum* (French), *Araneum Germanicum* (German), *Araneum Hispanicum* (Spanish), *Araneum Polonicum* (Polish), *Araneum Anglicum* (English), *Araneum Nederlandicum* (Dutch), and *Araneum Slovacum* (Slovak).

5 As an example we can point out the problem of the "apostrophe" character in French texts. As much as 8 different Unicode characters representing the apostrophe (with just two of them being "canonical") can be found in the texts collected from the web. As the Tree Tagger French parameter file originated in the pre-Unicode era, even one of the two "canonical" representations would not be processed (i.e., tokenized and lemmatized) properly without special measures, and tokens like "*l'*" and "*d'*", that belong to the most frequent ones, would be mistagged.

The crawling has also been done for *Araneum Bohemicum* (Czech). This data is now being pre-processed to be ready for annotation that will be performed by the Institute of Theoretical and Computational Linguistics at the Faculty of Arts of Charles University in Prague.[6]

The first stage of our project will be completed by *Araneum Hungaricum* (Hungarian), *Araneum Italicum* (Italian), and *Araneum Ukrainicum* (Ukrainian). With the exception of the last mentioned, we expect to complete the whole venture by the end of 2014.

For all of the languages mentioned, sketch grammars have been written and at least two rounds of testing have been performed for each corpus. The procedure involved is described in the following section.

3 Sketch grammars

A sketch grammar[7] is a set of rules based on the CQL (Corpus Query Language[8]) used by the Sketch Engine to generate the respective collocation profiles ("word sketches") for all lexical units (lemmas) in a corpus. The word sketches are pre-computed in advance, which makes the system user-friendly and very fast.

A sketch grammar rule consists of (1) an optional comment indicated by hash "#" character, (2) the rule type marked by an asterisk "*", (3) the rule name preceded by the equal sign "=", and (4) a list of CQL expressions. For example, a rule describing the relationship between two nouns (in English using the Penn Treebank tagset) might look as follows:

```
# noun followed by another noun
*DUAL
=modifier/modified
        2: [tag="NN.*"] 1: [tag="NN.*"]
```

The "1:" label denotes the "keyword", i.e. the lemma the word sketch is created for, and the "2:" label marks the lemma of the collocate. The "*DUAL" keyword indicates that the rule is to be used twice, the second time with swapped labels, i.e. exchanging the positions of the keyword and the collocate. The text following the slash "/" character will be used as a name for the second use of the rule.

In reality, the rules usually look slightly more complex to indicate that "intermediate" words may be present between a keyword and a collocate, or in the vicinity of them.

6 Besides Ukrainian, Czech is the only language within the *Aranea* project with no free tagging tool available.
7 https://www.sketchengine.co.uk/documentation/wiki/SkE/GrammarWriting
8 https://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying

3.1 What's in a name

Unlike Juliet Capulet[9], we believe that the name is often really important, and the sketch grammar rule name is exactly such a case. On one hand, it is the only component of the sketch grammar that is not predetermined, and thus can be "virtually anything". On the other hand, the name is the only clue for the user about the contents of the respective word sketch tables, and therefore should be as informative as possible. It has, however, to be very short as the name is displayed in the heading of the respective word sketch table within a only a limited space available. Rule names longer than 10–12 characters would increase the table widths, and the resulting word sketches could possibly not fit the screen.

Most sketch grammars used for corpora available at the *SkE* site follow the naming conventions introduced by A. Kilgarriff in the first English and French sketch grammars. These rule names are motivated syntactically, i.e. they denote the syntactic function of the collocate, with that of the keyword being implied. For example the rule name:

```
=modifier/modified
```

is representing two rule names with readings as follows: "collocate is a modifier of the keyword", and "collocate is modified by the keyword", respectively.

The syntactically motivated rules are transparent and user-friendly for description of basic relationships between subjects, object, modifiers/attributes, and verbs/predicates, but in more complex cases this strategy is not easily applicable. The nature of the problems can be observed in the Czech sketch grammar written by P. Smrž (Kilgarriff et al.; 2004). Some examples of rule names are as follows:

```
is_subj_of/has_subj
is_obj4_of/has_obj4
prec_prep
gen1/gen2
```

As it can be seen, it is not really easy for the user the figure out "who is who" in the keyword – collocate – syntactic function "puzzle". Moreover, rule names like "prec_verb" do not denote any syntactic functions but rather just describe collocational relationships.

There are two notable deviations from the "traditional" rule name conventions in the sketch grammars. In the grammar for the Slovenian FidaPLUS corpus[10], S. Krek (Krek; 2006) uses rule names containing (among other features) Slovenian "case questions". For example, the."*koga-česa*" name means

9 Juliet: *"What's in a name? that which we call a rose / By any other name would smell as sweet"* (William Shakespeare: Romeo and Juliet, Act II, Scene 2).

10 http://www.sketchengine.co.uk/documentation/wiki/Corpora/FidaPLUS

that only collocates of the keyword that are in genitive case are displayed, with the syntactic function of the collocate being implied.

The second notable exception is the sketch grammar written by P. Whilelock (2010) for the Oxford English Corpus[11] (OEC) where the rule names not only name the syntactic function, but also the PoS of the keyword and the collocate and their mutual position within the collocation. For example, the "V* ADJ" rule name stands for verb modified by an adjective, with askerisk indicating the keyword.

3.2 Sketch grammar for Slovak corpora

In our Institute, the *SkE* has been extensively used since autumn 2007 with several Slovak and Czech corpora. These corpora serve as a source of lexical evidence for our monolingual and bilingual lexicographic projects, as well as for other linguistic research activities.

The sketch grammar used in our *SkE* installation has been optimized for a lexicographic use, and differs from most "traditional" grammars for corpora stored at the *SkE* web site in several aspects:

- The rule names are not motivated syntactically (i.e., they do not indicate the syntactic relationship between the keyword and the collocate) but rather collocationally
- The right-hand or left-hand position of the collocate towards the keyword is indicated explicitly in the rule name
- The keyword's PoS in the rule is not specified, i.e., it covers any PoS
- Recall is preferred over precision
- The number of rules and the order of resulting tables is fixed
- The object names within the rules are governed by the following rules:
- The keyword is denoted by the X symbol
- The keyword's grammatical attributes (mostly in unary rules) are indicated by lowercase abbreviation, e.g., gen(X) indicates the genitive case of keyword
- The collocate's PoS is indicated by an abbreviation with a leading capital letter, e.g., Aj X indicates a left-hand adjective collocate
- Y indicates a collocate that is from any PoS category
- Z indicates a collocate from any PoS category not covered by the other "explicit" rules

3.3 Rule name summary

The core of our grammar consists of rules covering four basic autosemantic word classes. Taking into account our experience with early versions of the grammar, the rules for verbs (Vb X/X Vb) and adverbs (Av X/X Av) do not distinguish the left and right position of the respective collocate.

11 http://www.sketchengine.co.uk/documentation/wiki/Corpora/OEC

For nouns, two separate rules take into account the position of the collocate (Sb X, X Sb). Similar situations can be found with adjectives (Aj X; X Aj), prepositions (Pp X; X Pp) and for immediate autosemantic collocates (Y X; X Y). The "catch all" rules for the remaining word classes (Z X; X Z) cover mostly numerals and pronouns, as well as some synsemantic word classes.

The remaining two binary (symmetric) rules map the relationship of coordination, i.e., the situation where a keyword and a collocate with compatible morphological tags are separated by a comma (X/Y , X/Y) or a conjunction (X/Y Cj X/Y).

The four trinary rules cover relationships among a keyword, collocate, and preposition in different positions (Pp Y X, Pp X Y, Y Pp X, and X Pp Y).

Our set of rules is complemented by unary rules showing the frequency distribution of the keyword's forms according to grammatical categories and subcategories..

The compatible grammars

In creating sketch grammars for a group of languages, it is convenient not to use the "native" tagsets for the respective languages, but rather to use a common symbolic notation. This can be done, e.g., by means of a macro processor (such as m4). We have, however, decided to adopt a different approach and to create a simple universal tagset ("*Araneum Universal Tagset*" – AUT) similar to that of the *Universal PoS Tagset*[12] (UPT; Petrov et al., 2011), and to map all the respective tagsets into this tagset at the source vertical data level, i.e. to create a new layer of annotation. The *AUT* contains 11 tags for "traditional" part of speech categories, 7 additional tags for other elements, and one tag to indicate errors in the mapping process.

aTag	PoS
Dt	determiner/article
Nn	noun
Aj	adjective
Pn	pronoun
Nm	numeral
Vb	verb
Av	adverb

aTag	PoS
Pp	preposition
Cj	conjunction
Ij	interjection
Pt	particle
Ab	abbreviation/acronym
Sy	symbol
Nb	number

aTag	PoS
Xx	other (content word)
Xy	other other (function word)
Yy	unknown/foreign/alien
Zz	punctuation
Er	mapping error

Table 1: Araneum Universal Tagset (*AUT*).

The compatible sketch grammar using *AUT* consists of three sections. The first part (*AUT*-based) contains unary rules showing PoS category distribution for a particular lemma. The second part is

12 The AUT PoS tags for the eleven „traditional" word classes directly correspond with those of UPT, with the difference being just in the names as we wanted to keep the names of the PoS categories identical with those used in the sketch grammar rule names introduced before the UPT tagset has been published. The additional 7 categories accommodate information provided by the respective "native" tagsets that is being ignored by UPT. For example, the "Xx" (other: content word) tag is assigned to participles in Slovak that have a category of their own in the SNK Slovak tagset.

tagset-dependent and contains unary rules showing PoS subcategories provided by the respective tagset. Due to differences in the depth of the morpho-syntactic annotation, the number of subcategories varies among the languages. With verbs, e.g., we have just 5 subcategories for Spanish, while more than 20 for Polish. The final third section (*AUT*-based) covers the collocational relationships of the respective keyword by means of binary, symmetric and trinary rules.

The compatible sketch grammar is basically identical for all the languages with one important exception: the number of intermediate tokens between a keyword and a collocate is increased by one for languages having articles in their language system.

4 Discussion and conclusion

A collocationally-based sketch grammar has (against a traditional one) several advantages. It can symmetrically cover all relationships between keywords and collocates of all word classes (parts of speech). As the PoS category is not tested for the keyword, a word sketch can be created even in cases of incorrectly assigned tags. If the same (compatible) sketch grammar is used with corpora for two or more languages, the resulting word sketches can be conveniently used in contrastive linguistic research, as well as within bilingual lexicographic projects.

The disadvantage of our approach is that not all tables for some words represent linguistically relevant relationships, and they may contain a lot of noise. We believe, however, that having a fixed number of tables gives the user a clear overview, and he or she can easily ignore the irrelevant data.

In the Appendix, we present the word sketches for the lemma “without” created by means of compatible sketch grammars from four Aranea web corpora..

5 Further work

In the near future, we plan to carry out activities within several tracks. Firstly, we would like to improve the quality of the Aranea corpus data itself (by means of better filtration, normalization and deduplication), as well as its morpho-syntactic annotation by means of long-term evaluation of the resulting word sketches. Secondly, we want to include new languages into our Aranea corpus family and to write the respective corpus grammars, at least for the languages taught at Slovak universities. And finally, we plan to tune the global parameters of the compatible sketch grammars, as well as provide language-specific improvements so that the bilingual word sketches provide more relevant results.

6 References

Baroni, M. – Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In: *Proceedings of LREC'2004*. Lisbon: ELRA ,2004.

Benko, V. (2013). Data Deduplication in Slovak Corpora. In: *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning*. Katarína Gajdošová (Ed.), Adriána Žáková. Lüdenscheid: RAM-Verlag, 2013, pp. 27–39.

Hong, J. F. – Huang, C. R. (2007). Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research. URL: http://www.ling.sinica.edu.tw/eip/FILES/publish/2007.7.18.93102662.8243902.pdf.

Jakubíček, M. (2014). Personal communication.

Kilgarriff, A. et al. (2004). The Sketch Engine. In: G. Williams and S.Vessier (eds.), Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6–10, 2004. Lorient: Université de Bretagne-Sud, pp. 105–116.

Khokhlova, M. (2010). Building Russian Word Sketches as Models of Phrases. In: *Proc. EURALEX 2010, Leeuwarden,* July 2010.

Kovář, V. (2013). New features in the Sketch Engine interface. Part 1. In: *SKEW-4 Workshop*, Tallinn, October 2013. URL: https://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SKEW-4/Program/ske_interface_part1.pdf

Krek, S. – Kilgarriff, A. (2006). Slovene Word Sketches. In: *Proceedings of 5th Slovenian and 1st international Language Technologies Conference 2006*. October 9th – 10th 2006. Jožef Stefan Institute, Ljubljana, Slovenia

Macoveiciuc, M. – Kilgarriff, A. (2010). The RoWaC Corpus and Romanian Word Sketches. In: *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Edited by Dan Tufis and Corina Forascu. Romanian Academy.

Petrov, S. – Das, D. – McDonald, R. (2012). A Universal Part-of-Speech Tagset. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation* (*LREC'12*). Istanbul: ELRA, May 2012.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. TASK QUARTERLY 11, No 1–2, 151–167

Radziszewski, A. – Kilgarriff, A. – Lew, R. (2011). Polish Word Sketches In: Zygmunt Vetulani (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 5th Language & Technology Conference. Poznań : Fundacja Uniwersytetu im. A. Mickiewicza.

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Srdanović, E. I. – Erjavec. T. – Kilgarriff, A. (2008). A web corpus and word-sketches for Japanese. In: *Journal of Natural Language Processing* 15/2, 137– 159. (reprinted in Information and Media Technologies 3/3, 2008, 529– 551)

Suchomel, V. – Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In: *7th Web as Corpus Workshop* (*WAC-7*), Lyon, 2012.

Tiberius, C. – Kilgarriff, A. (2009). The Sketch Engine for Dutch with the ANW corpus. In: *Fons Verbhorum, Festschrift for Fons Moerdijk*. Instituut voor Nederlandse Lexicologie, Leiden, The Netherlands.

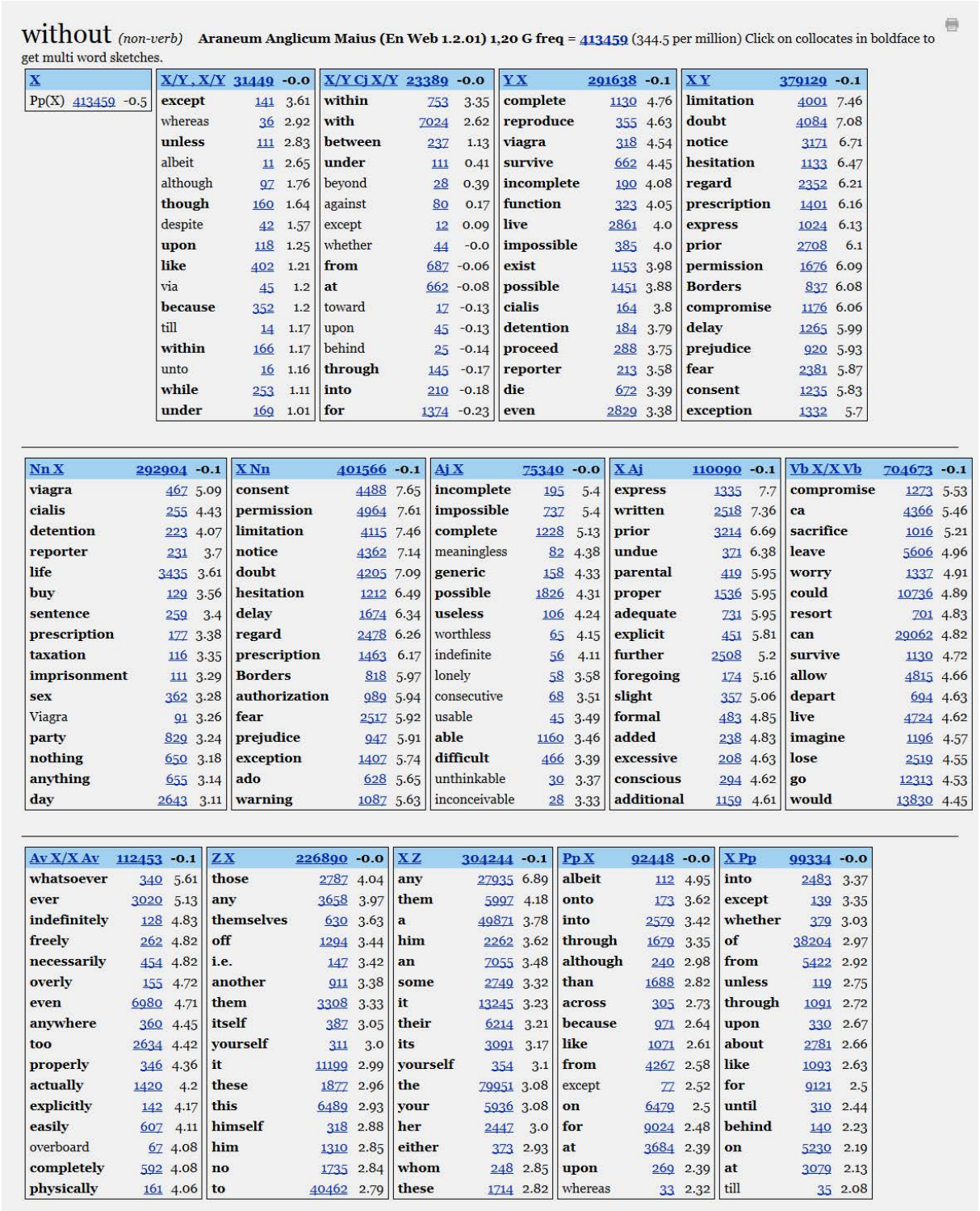
Whitelock, P. (2010). Personal communication.

Appendix

To demonstrate the compatible word sketches, we present screen shots for the preposition "without" in four languages (English, French, German, and Russian). Prepositions belong to word classes that are usually either not covered by the respective traditional sketch grammars at all, or that produce a limited number of output word sketch tables only.

For all languages involved, we can observe the typical binary collocations with noun and verbs. The collocations with adjectives usually form a multi-word expression that is not fully displayed in the word sketches, but many of those can be easily recognized even without going into the actual concordances.

Note: Due to the longer adjectives in Russian, the interesting table with verbal collocates did not fit onto the screen.



without *(non-verb)* **Araneum Anglicum Maius (En Web 1.2.01) 1,20 G freq** = 413459 (344.5 per million) Click on collocates in boldface to get multi word sketches.

X	
Pp(X) 413459	-0.5

X/Y , X/Y	31449	-0.0
except	141	3.61
whereas	36	2.92
unless	111	2.83
albeit	11	2.65
although	97	1.76
though	160	1.64
despite	42	1.57
upon	118	1.25
like	402	1.21
via	45	1.2
because	352	1.2
till	14	1.17
within	166	1.17
unto	16	1.16
while	253	1.11
under	169	1.01

X/Y Cj X/Y	23389	-0.0
within	753	3.35
with	7024	2.62
between	237	1.13
under	111	0.41
beyond	28	0.39
against	80	0.17
except	12	0.09
whether	44	-0.0
from	687	-0.06
at	662	-0.08
toward	17	-0.13
upon	45	-0.13
behind	25	-0.14
through	145	-0.17
into	210	-0.18
for	1374	-0.23

Y X	291638	-0.1
complete	1130	4.76
reproduce	355	4.63
viagra	318	4.54
survive	662	4.45
incomplete	190	4.08
function	323	4.05
live	2861	4.0
impossible	385	4.0
exist	1153	3.98
possible	1451	3.88
cialis	164	3.8
detention	184	3.79
proceed	288	3.75
reporter	213	3.58
die	672	3.39
even	2829	3.38

X Y	379129	-0.1
limitation	4001	7.46
doubt	4084	7.08
notice	3171	6.71
hesitation	1133	6.47
regard	2352	6.21
prescription	1401	6.16
express	1024	6.13
prior	2708	6.1
permission	1676	6.09
Borders	837	6.08
compromise	1176	6.06
delay	1265	5.99
prejudice	920	5.93
fear	2381	5.87
consent	1235	5.83
exception	1332	5.7

Nn X	292904	-0.1
viagra	467	5.09
cialis	255	4.43
detention	223	4.07
reporter	231	3.7
life	3435	3.61
buy	129	3.56
sentence	259	3.4
prescription	177	3.38
taxation	116	3.35
imprisonment	111	3.29
sex	362	3.28
Viagra	91	3.26
party	829	3.24
nothing	650	3.18
anything	655	3.14
day	2643	3.11

X Nn	401566	-0.1
consent	4488	7.65
permission	4964	7.61
limitation	4115	7.46
notice	4362	7.14
doubt	4205	7.09
hesitation	1212	6.49
delay	1674	6.34
regard	2478	6.26
prescription	1463	6.17
Borders	818	5.97
authorization	989	5.94
fear	2517	5.92
prejudice	947	5.91
exception	1407	5.74
ado	628	5.65
warning	1087	5.63

Aj X	75340	-0.0
incomplete	195	5.4
impossible	737	5.4
complete	1228	5.13
meaningless	82	4.38
generic	158	4.33
possible	1826	4.31
useless	106	4.24
worthless	65	4.15
indefinite	56	4.11
lonely	58	3.58
consecutive	68	3.51
usable	45	3.49
able	1160	3.46
difficult	466	3.39
unthinkable	30	3.37
inconceivable	28	3.33

X Aj	110090	-0.1
express	1335	7.7
written	2518	7.36
prior	3214	6.69
undue	371	6.38
parental	419	5.95
proper	1536	5.95
adequate	731	5.95
explicit	451	5.81
further	2508	5.2
foregoing	174	5.16
slight	357	5.06
formal	483	4.85
added	238	4.83
excessive	208	4.63
conscious	294	4.62
additional	1159	4.61

Vb X/X Vb	704673	-0.1
compromise	1273	5.53
ca	4366	5.46
sacrifice	1016	5.21
leave	5606	4.96
worry	1337	4.91
could	10736	4.89
resort	701	4.83
can	29062	4.82
survive	1130	4.72
allow	4815	4.66
depart	694	4.63
live	4724	4.62
imagine	1196	4.57
lose	2519	4.55
go	12313	4.53
would	13830	4.45

Av X/X Av	112453	-0.1
whatsoever	340	5.61
ever	3020	5.13
indefinitely	128	4.83
freely	262	4.82
necessarily	454	4.82
overly	155	4.72
even	6980	4.71
anywhere	360	4.45
too	2634	4.42
properly	346	4.36
actually	1420	4.2
explicitly	142	4.17
easily	607	4.11
overboard	67	4.08
completely	592	4.08
physically	161	4.06

Z X	226890	-0.0
those	2787	4.04
any	3658	3.97
themselves	630	3.63
off	1294	3.44
i.e.	147	3.42
another	911	3.38
them	3308	3.33
itself	387	3.05
yourself	311	3.0
it	11199	2.99
these	1877	2.96
this	6489	2.93
himself	318	2.88
him	1310	2.85
no	1735	2.84
to	40462	2.79

X Z	304244	-0.1
any	27935	6.89
them	5997	4.18
a	49871	3.78
him	2262	3.62
an	7055	3.48
some	2749	3.32
it	13245	3.23
their	6214	3.21
its	3091	3.17
yourself	354	3.1
the	79951	3.08
your	5936	3.08
her	2447	3.0
either	373	2.93
whom	248	2.85
these	1714	2.82

Pp X	92448	-0.0
albeit	112	4.95
onto	173	3.62
into	2579	3.42
through	1679	3.35
although	240	2.98
than	1688	2.82
across	305	2.73
because	971	2.64
like	1071	2.61
from	4267	2.58
except	77	2.52
on	6479	2.5
for	9024	2.48
at	3684	2.39
upon	269	2.39
whereas	33	2.32

X Pp	99334	-0.0
into	2483	3.37
except	139	3.35
whether	379	3.03
of	38204	2.97
from	5422	2.92
unless	119	2.75
through	1091	2.72
upon	330	2.67
about	2781	2.66
like	1093	2.63
for	9121	2.5
until	310	2.44
behind	140	2.23
on	5230	2.19
at	3079	2.13
till	35	2.08

ohne **Araneum Germanicum Maius (De Web 1.2.01) 1,20 G freq** = 592115 (493.4 per million) Click on collocates in boldface to get multi word sketches.

X	
Pp(X)	505192 3.0
Cj(X)	86923 0.9

X/Y , X/Y	34121	0.2
dass	1892	4.32
ob	270	3.32
oder	1064	3.25
d.h.	21	3.05
wie	659	2.85
und	5054	2.84
einschließlich	39	2.69
bzw.	54	2.23
inklusive	30	2.2
außer	45	2.17
weil	61	2.1
als	464	2.06
infolge	12	1.72
indem	12	1.6
aufs	20	1.43
sondern	55	1.37

X/Y Cj X/Y	41201	0.2
einschließlich	57	3.19
seitens	40	2.8
mit	8640	2.61
inklusive	34	2.33
zwischen	382	1.97
trotz	71	1.8
mittels	40	1.8
binnen	12	1.79
außer	30	1.55
samt	12	1.54
aufs	22	1.53
außerhalb	39	1.51
wider	11	1.47
ans	18	1.4
ob	71	1.38
gegen	269	1.36

Y X	323276	0.3
Kredit	2342	6.4
jederzeit	1222	5.36
Reporter	414	5.2
gänzlich	475	5.11
ganz	8310	4.75
Fahren	306	4.63
völlig	929	4.54
Fass	236	4.34
Abnehmen	236	4.32
Girokonto	288	4.28
Handy	539	4.26
Kreditkarte	348	4.24
Rechnung	486	4.13
Leben	2437	4.12
allerdings	1516	4.09
Pfanne	226	4.03

X Y	483395	0.4
vorherig	4624	7.62
Schufa	3481	7.57
Zweifel	3621	7.26
Rücksicht	2869	7.24
Weiteres	1810	6.82
Einschränkung	2002	6.49
Angabe	4455	6.46
gesondert	1607	6.35
Zustimmung	2145	6.3
Abzug	1372	6.19
Gewähr	1324	6.19
Problem	6810	6.14
Umweg	1192	6.11
Unterbrechung	1176	6.11
ausdrücklich	2164	6.1
Behinderung	1596	5.93

Nn X	311082	0.3
Kredit	2584	6.56
Publikation	1004	5.87
Reporter	415	5.26
Fahren	337	4.82
Rechnung	739	4.76
Handy	708	4.67
Girokonto	340	4.55
Kreditkarte	402	4.48
Abnehmen	247	4.43
Fass	239	4.41
Leben	2895	4.37
Angebot	1875	4.2
Tarif	439	4.19
Pfanne	242	4.17
Abmahnung	267	4.13
Baufinanzierung	197	4.05

X Nn	606564	0.5
Zustimmung	5620	7.5
Schufa	3453	7.29
Zweifel	4051	7.21
Ankündigung	3067	7.07
Rücksicht	3041	7.07
Aufwand	3653	6.93
Einschränkung	2510	6.59
Genehmigung	2522	6.55
Weiteres	1818	6.52
Angabe	4874	6.49
Grund|Gründen	3484	6.45
Einwilligung	1781	6.31
Problem	7956	6.31
Abzug	1758	6.29
Umweg	1353	6.01
Gewähr	1384	5.99

Aj X	89134	0.2
gedruckt	973	7.25
gänzlich	485	6.2
völlig	1022	4.96
anwaltlichen	83	4.73
gesamt	1581	4.62
erhoben	133	4.52
Prepaid	112	4.5
personenbezogen	248	4.5
komplett	820	4.38
selbstverständlich	433	4.38
viagra	61	4.27
vollkommen	226	4.26
undenkbar	64	4.24
kommerziell	153	4.22
verlinkten	97	4.19
berufsmäßig	52	4.19

X Aj	200309	0.5
vorherig	4938	8.37
gesondert	1704	7.27
ausdrücklich	3061	7.06
nennenswert	811	6.66
schriftlich	2599	6.55
lästig	639	6.18
erkennbar	856	6.09
störend	448	5.8
weit	11902	5.58
zusätzlich	3407	5.58
möglich	4803	5.48
ersichtlich	345	5.34
fremd	827	5.32
unnötig	449	5.28
aufwendig	523	5.21
finanziell	1189	5.19

Vb X/X Vb	
auskommen	
gestatten|stat	
funktioniere	
verändern	
verlaufen	
verlieren	
leben	
laufen	
zögern	
dürfen	
verlassen	
nachdenken	
bleiben	
können	
kündigen	
überstehen	

Av X/X Av	173826	0.4
jemals	724	6.31
jederzeit	1696	6.07
vorher	1598	5.93
jedoch	4386	5.15
allerdings	2735	5.01
ganz	9471	4.96
niemals	434	4.86
leider	1386	4.66
irgend	201	4.65
kaum	1261	4.63
freilich	204	4.62
also	3264	4.58
bisher	1131	4.56
auch	35534	4.56
sofort	1136	4.53
meistens	430	4.52

Z X	285861	0.3
man	8344	4.02
14	690	3.93
diese	1417	3.91
nicht	20169	3.85
wer	1280	3.83
keine	4041	3.74
daher	740	3.57
solche	1093	3.54
sie	6700	3.53
niemand	257	3.45
eine	37041	3.41
jemand	317	3.39
was	2380	3.32
er	5165	3.25
nichts	655	3.24
es	9643	3.23

X Z	356121	0.4
jegliche	4363	7.76
irgendwelche	1088	6.08
dabei	6094	5.38
irgendeine	556	4.95
allzu	476	4.74
dafür	1743	4.61
jede	5801	4.4
zu	52715	4.36
nichts	1282	4.16
keine	4669	3.94
solche	1384	3.85
jedwede	159	3.75
darauf	740	3.74
jemand	424	3.71
viele	4400	3.71
darüber	676	3.69

Pp X	67488	0.1
dank	140	3.13
pro	258	2.87
bei	5283	2.86
einschließlich	51	2.85
gegenüber	221	2.71
inklusive	49	2.7
trotz	135	2.67
nach	2766	2.58
innerhalb	209	2.49
für	7039	2.48
wegen	200	2.46
seit	561	2.46
zeit	20	2.46
von	8310	2.43
mittels	63	2.36
binnen	20	2.23

X Pp	111231	0.3
seitens	198	4.65
durchs	80	3.87
ans	124	3.79
auf	15175	3.71
außer	138	3.48
aufs	105	3.43
von	16501	3.42
durch	3718	3.34
binnen	56	3.32
über	3330	3.0
an	7594	2.92
innerhalb	277	2.84
per	225	2.79
in	22140	2.69
mit	8122	2.52
gegen	606	2.5

sans **Araneum Francogallicum Maius (Fr Web 1.2.02) 1,23 G freq** = 1023016 (829.5 per million) Click on collocates in boldface to get multi word sketches.

X		
Pp(X)	1012928	-0.6
Cj(X)	10088	-0.0

X/Y , X/Y	165883	-0.2
sauf	248	3.9
jusque	599	3.29
malgré	217	3.13
hormis	51	3.03
hors	147	3.01
avec	4598	2.94
depuis	869	2.89
chez	541	2.89
vers	626	2.88
pendant	439	2.83
sous	725	2.77
dès	262	2.63
à	18464	2.62
au	7379	2.59
devant	323	2.58
en	12377	2.58

X/Y Cj X/Y	89170	-0.1
avec	11553	4.27
hors	104	2.71
soit	77	2.17
malgré	78	1.8
parce	36	1.58
entre	519	1.56
sauf	41	1.52
en	5765	1.48
sous	284	1.45
pendant	162	1.45
contre	250	1.37
à	7766	1.37
dès	98	1.29
du	9242	1.28
jusque	137	1.22
de	23057	1.17

Y X	638024	-0.1
reporter	975	5.06
non	5532	4.86
c'est-à-dire	804	4.49
fonctionner	1169	4.48
organisme	1212	4.2
répéter	637	4.16
rester	3613	4.16
accepter	1380	4.15
presque	971	4.14
vivre	2985	4.14
consommer	669	4.14
interdire	864	4.01
mourir	943	3.98
vétérinaire	364	3.94
voiture	857	3.87
dérouler	816	3.85

X Y	878596	-0.1
doute	85421	9.99
cesse	20493	9.02
oublier	15100	7.48
fil	7755	7.07
précédent	6972	6.97
préavis	3701	6.96
autorisation	5545	6.84
autant	9143	6.81
relâche	3002	6.72
compter	12738	6.67
frontière	4720	6.61
conteste	2740	6.6
faille	3165	6.6
consentement	2514	6.28
moindre	3757	6.26
limite	4816	6.24

Nn X	526473	-0.1
reporter	920	5.16
organisme	1441	4.53
journée	1822	4.3
voiture	1075	4.28
nuit	1242	4.16
Internet	1353	4.07
réseau	2004	4.07
acceptation	367	4.02
amour	1131	3.91
jour	4405	3.88
médecin	739	3.88
licenciement	300	3.86
connexion	400	3.79
monde	3438	3.77
sexe	522	3.74
gens	1270	3.74

X Nn	969635	-0.1
doute	96148	10.11
cesse	20510	8.93
fil	7829	7.02
préavis	3832	6.89
autorisation	6013	6.88
précédent	6605	6.82
relâche	3004	6.59
faille	3200	6.49
conteste	2760	6.47
limite	5331	6.32
surprise	3308	6.29
consentement	2647	6.24
souci	3214	6.03
hésitation	2068	6.02
arrêt	3353	6.0
frontière	3229	5.98

Aj X	118195	-0.1
vétérinaire	260	5.06
impossible	736	5.01
tierce	139	4.55
modifiable	97	4.54
gratuit	747	4.34
accessible	517	4.29
possible	1898	4.26
utilisable	108	4.08
remboursable	69	3.93
sexuel	320	3.89
correct	133	3.85
partiel	173	3.8
réalisable	63	3.71
estre	81	3.7
immédiat	197	3.67
inconcevable	53	3.66

X Aj	167461	-0.1
lucratif	2851	8.2
préalable	4847	7.94
moindre	5688	7.7
expresse	1296	7.5
nul	2863	7.15
apparent	1106	6.93
frais	3179	6.59
frontière	1676	5.97
égal	1040	5.64
réel	2466	5.61
gras	748	5.6
fixe	640	5.5
excessif	420	5.44
supplémentaire	1195	5.42
valable	526	5.37
explicite	298	5.17

Vb X/X Vb	1329224	-0.1
oublier	16163	7.3
compter	14308	6.6
parler	12885	5.9
soucier	2432	5.7
tarder	2551	5.7
laisser	9961	5.6
attendre	7517	5.6
perdre	5087	5.3
pouvoir	54251	5.3
passer	13149	5.2
hésiter	2779	5.1
savoir	12005	5.1
rester	7627	5.0
vivre	6437	5.0
bouger	1713	5.0
regarder	3542	5.0

Av X/X Av	287079	-0.1
autant	10797	7.41
toutefois	3317	6.99
jamais	11860	6.7
trop	11522	6.58
forcément	1325	5.84
nécessairement	855	5.74
rien	6754	5.71
vraiment	3859	5.5
non	7636	5.4
priori	564	5.33
réellement	926	5.33
c'est-à-dire	989	5.31
exprès	387	5.23
presque	1466	5.05
préalablement	416	5.02
apparemment	419	4.92

Z X	526005	-0.1
ceci	1004	4.96
cela	6226	4.58
la|le	9272	4.47
la	2222	4.32
se	32382	4.28
un	77428	4.26
on	10142	4.24
elle	9775	4.24
il	25892	4.1
votre	5454	4.07
vous	12695	4.07
tu	2004	3.99
ce	38418	3.91
y	5689	3.85
leur	5436	3.81
qui	24786	3.8

X Z	554614	-0.1
aucun	36125	8.34
quoi	3522	5.73
OGM_	963	5.49
la|le	10965	4.71
la	2817	4.65
toi	1128	4.64
eux	1697	4.46
lequel	3094	4.41
lui	6975	4.33
y	7918	4.32
leur	7251	4.22
se	30956	4.21
me	6916	4.14
son	19423	4.14
que	19955	4.11
votre	5045	3.96

Pp X	258101	-0.1
devant	691	3.61
chez	870	3.52
vers	992	3.49
du	42461	3.48
à	31897	3.41
de	107821	3.4
pendant	677	3.39
après	1187	3.35
dans	11797	3.18
pour	11453	3.06
avec	4790	2.99
sur	7429	2.92
au	9114	2.9
derrière	177	2.85
contre	726	2.83
en	14810	2.83

X Pp	349901	-0.1
jusque	1966	4.88
pour	27339	4.31
avec	9295	3.94
pendant	1000	3.89
sur	13933	3.83
à	42038	3.8
depuis	1684	3.78
de	137573	3.75
au	15714	3.68
par	11444	3.67
devant	715	3.59
dans	15547	3.58
vers	1043	3.52
contre	1189	3.51
envers	242	3.43
quant	240	3.42

без

Araneum Russicum Maius (Ru Web 1.2.02) 1,22 G freq = 749749 (616.2 per million) Click on collocates in boldface to get multi word sketches.

X		
Pp(X)	749749	-2.3
nom(X)	6477	-0.0
gen(X)	737804	-1.3
dat(X)	1640	-0.0
acc(X)	1003	-0.0
loc(X)	2266	-0.0
ins(X)	559	-0.0

X/Y , X/Y	16565	-0.2
безо	15	3.46
вне	90	3.05
кроме	230	2.46
ради	54	1.86
посредством	18	1.21
вместо	39	0.99
путем	24	0.98
помимо	24	0.91
среди	74	0.65
выше	16	0.6
возле	11	0.56
от	782	0.4
для	989	0.35
из	811	0.33
у	552	0.27
внутри	22	0.27

X/Y Cj X/Y	9119	-0.1
вне	76	2.84
у	526	0.2
помимо	12	-0.07
вместо	16	-0.28
против	44	-0.28
вокруг	14	-0.74
для	445	-0.81
со	90	-0.84
до	163	-0.86
из-за	22	-0.91
от	294	-1.01
после	78	-1.19
среди	18	-1.39
из	237	-1.44
с	296	-2.38

Y X	424569	-0.3
обойтись	12297	8.55
обходиться	5964	8.15
остаться	14808	6.92
пропасть	2590	6.68
невозможный	2621	6.53
оставить	4941	6.24
невозможно	2174	6.07
немыслимый	954	5.93
оставлять	1678	5.37
пропавшими	485	5.2
практически	2233	5.14
оставаться	3850	5.12
прожить	741	4.98
жить	4890	4.89
вообще	2127	4.79
почти	1985	4.67

X Y	620883	-0.3
исключение	9151	7.69
сомнение	8080	7.63
весть	4171	7.49
попечения	2636	7.02
преувеличение	2256	6.76
особый	9784	6.71
труд	8844	6.61
лишний	4748	6.58
учет	6474	6.55
малое	1960	6.51
предварительный	2943	6.33
малейший	2032	6.25
присмотр	1582	6.25
посредник	1786	6.21
согласие	2584	6.19
разбор	1567	6.16

Nn X	364089	-0.3
оставление	423	5.12
кредит	1386	4.88
репортер	312	4.68
дым	377	4.51
участок	1432	4.4
кофе	432	4.26
чай	565	4.17
лицо	2825	4.14
секс	396	4.02
отпуск	412	4.01
наличные	219	3.98
квартира	1366	3.95
бой	586	3.93
жизнь	5441	3.92
столбик	187	3.91
заработок	387	3.84

X Nn	844256	-0.4
сомнение	9447	7.64
исключение	9375	7.52
весть	4173	7.12
труд	10846	6.8
согласие	4377	6.69
попечения	2737	6.66
преувеличение	2603	6.56
ведомо	2464	6.49
учет	6771	6.48
разрешение	4377	6.35
ограничение	3996	6.31
усилие	4131	6.25
малое	1968	6.12
потеря	3707	6.11
ущерб	2715	6.09
вмешательство	2359	6.03

Aj X	88968	-0.2
немыслимый	1008	7.55
пропавшими	471	7.31
невозможный	2767	7.25
пропавшим	176	5.94
худой	324	5.78
мыслимый	187	5.55
предпринимательский	249	5.14
минеральный	263	4.66
неполный	156	4.63
исковый	124	4.43
ровный	205	4.27
гладкий	146	4.12
земельный	290	3.92
апелляционный	83	3.91
длительный	397	3.85
Послеоперационный	39	3.82

X Aj	194940	-0.3
малейший	2061	7.24
предварительный	3438	7.19
видимый	1687	7.12
лишний	4940	7.08
особый	10017	6.96
посторонний	1570	6.9
уважительный	761	6.6
должный	1241	6.53
невозможный	1714	6.32
дополнительный	4464	6.13
излишний	669	5.86
надлежащий	563	5.67
немыслимый	397	5.53
хирургический	475	5.48
единый	2288	5.27
специальный	3111	5.13

Vb X/
обой
обход
остат
остав
проп
остав
прож
жить
остав
преде
смочь
мысл
позво
мочь
спра
выжи

Av X/X Av	100363	-0.2
невозможно	7192	8.32
немыслимо	439	6.99
нельзя	4436	6.65
извне	309	5.98
практически	3089	5.85
трудно	1261	5.83
сложно	797	5.79
вообще	3670	5.78
скучно	252	5.77
вовсе	983	5.53
бесплатно	507	5.37
можно	13945	5.32
почти	2600	5.25
желательно	325	4.96
возможно	826	4.93
тяжело	302	4.91

Z X	231029	-0.2
куда	1227	5.54
тут	1299	4.65
даже	4526	4.49
ж	324	4.4
никак	412	4.35
поэтому	1540	4.28
ведь	1216	4.15
теперь	1167	4.12
здесь	1347	4.05
не	40489	3.96
бы	4480	3.82
весь	8436	3.81
тоже	1295	3.73
туда	277	3.73
как-то	277	3.62
просто	1910	3.61

X Z	226728	-0.3
всякий	21391	8.49
какой-либо	10149	8.07
никуда	887	6.39
никак	1459	6.18
таковой	1174	6.03
оное	252	5.04
оный	281	4.87
чей-либо	190	4.68
какой	2738	3.89
ничто	1527	3.82
то	13929	3.79
они	13906	3.63
она	9102	3.56
он	15472	3.55
пять	539	3.45
ваш	2980	3.43

Pp X	70963	-0.1
ко	271	3.49
сверх	23	2.68
за	2747	2.45
на	12506	2.44
обо	39	2.41
до	1499	2.34
во	994	2.33
посреди	19	2.29
в	23004	2.25
через	483	2.21
при	1481	2.2
к	3292	2.16
из	2797	2.11
об	486	2.09
сквозь	29	2.09
под	609	2.06

X Pp	79547	-0.1
со	2113	3.69
над	724	3.57
для	6323	3.02
на	16431	2.83
к	4975	2.75
с	8158	2.4
от	3131	2.4
о	2583	2.28
из-за	206	2.24
об	527	2.21
свыше	40	2.21
около	223	2.2
до	1265	2.09
пред	27	2.07
между	375	2.04
в	19486	2.01

From GLÀFF to PsychoGLÀFF: a Large Psycholinguistics-oriented French Lexical Resource

Basilio Calderone, Nabil Hathout, Franck Sajous
CLLE-ERSS (CNRS&Université de Toulouse-Le Mirail)
{basilio.calderone, nabil.hathout, franck.sajous}@univ-tlse2.fr

Abstract

In this paper, we present two French lexical resources, GLÀFF and PsychoGLÀFF. The former, automatically extracted from the collaborative online dictionary Wiktionary, is a large-scale versatile lexicon exploitable in Natural Language Processing applications and linguistic studies. The latter, based on GLÀFF, is a lexicon specifically designed for psycholinguistic research.
GLÀFF, counting more than 1.4 million entries, features an unprecedented size. It reports lemmas, main syntactic categories, inflectional features and phonemic transcriptions. PsychoGLÀFF contains additional information related to formal aspects of the lexicon and its distribution. It contains about 340,000 entries (120,000 lemmas) that are corpora-attested. We explain how the resources have been created and compare them to other known resources in terms of coverage and quality. Regarding PsychoGLÀFF, the comparison shows that it has an exceptionally large repertoire while having a comparable quality.
Keywords: French lexicon; lexical resource for psycholinguistic studies; Wiktionary

1 Introduction

Lexical resources play an important role in psycholinguistics by providing researchers with a set of experimentally relevant corpus information concerning words and, to a lesser extent, their sub-lexical components. In particular, psycholinguists working on lexical access need to manipulate a set of formal properties of words, such as syllabification, phonemic transcription, lemmas, inflected forms or orthographic/phonological neighborhood (i.e., the number of words differing from the target word by only one character/phoneme). Word frequency is possibly the most crucial information to be accounted for in psycholinguistic studies, and it is generally provided for either wordforms, lemmas, or both. The most well-known resource for English, German and Dutch is probably CELEX (Baayen et al. 1995). Many other languages, including French, lack a similar resource.
Some freely available French morphological lexicons, such as Lefff (Clément et al. 2004) and Morphalou (Romary et al. 2004), contain inflected forms, lemmas and morphosyntactic tags. These resources, designed in the first place for natural language processing (NLP) or lexicography do not include, ho-

wever, phonemic transcriptions that are necessary to set up psycholinguistics experiments, for extensive morphology and in the design of tools such as phonetizers. One noticeable exception is Lexique (New 2006), a free lexicon quite popular in psycholinguistics. This lexicon includes phonemic transcriptions, word frequencies and various features relevant to this field. However it has a limited coverage, especially in terms of inflected forms. All other resources that have both exploitable coverage and phonemic transcriptions, such as BDLex (Pérennou and de Calmès 1987), ILPho (Boula De Mareuil et al. 2000) or GlobalPhone (Schultz et al. 2013) are not free. Besides their cost, derivative works cannot be redistributed, which constitutes an impediment for collaborative research. As of today, no French lexicon meets all following requirements: free license, wide coverage, phonemic transcriptions and word frequencies.

In this article, we present a psycholinguistics-oriented resource based on *Wiktionnaire*,[1] the French edition of Wiktionary. In a previous work (Sajous et al. 2013a), we automatically extracted GLÀFF, *"a Large Versatile French Lexicon"*. This large-scale resource contains, for each entry, inflectional and phonemic information. PsychoGLÀFF is a new step leveraging Wiktionnaire's content.[2] Grounded on GLÀFF, PsychoGLÀFF is a lexicon that contains additional features specifically designed for meeting psycholinguistic needs.

The paper is organized as follows: Section 2 gives an overview of Wiktionnaire and its features relevant to lexical resources building. GLÀFF is then described in Section 3. Finally, we present in Section 4 PsychoGLÀFF, a lexicon designed for psycholinguistics use and compare it to Lexique in terms of coverage and word frequency. Conclusions and future directions of work are discussed in Section 5.

2 Wiktionnaire as a source of lexical knowledge

Wiktionary is a free multilingual dictionary available online. As its mother project Wikipedia, Wiktionary is based on the wiki paradigm: every internet user may contribute by adding content or modifying existing one. Launched in 2003, the Wiktionary project boasts, ten years later, more than two million entries for its French language edition, the Wiktionnaire. The impressive size of its headword list has to be tempered: inflected forms, discussion pages and, more surprisingly, *"pages describing in French words from other languages"* are counted as regular entries. However, once these latter entries excluded, Wiktionnaire still accounts for 1.4 million entries (186,000 lemmas).

While Wikipedia has been extensively used in various disciplines, its lexicographic counterpart seems to have received less attention from the scientific community. Wiktionary was first used in NLP by Zesch et al. (2008) to compute semantic relatedness. Its potential as an electronic lexicon was studied for the first time by Navarro et al. (2009) for French and English synonymy mining. Along the same line of research, Anton Perez et al. (2011) realized the integration of the Portuguese edition of

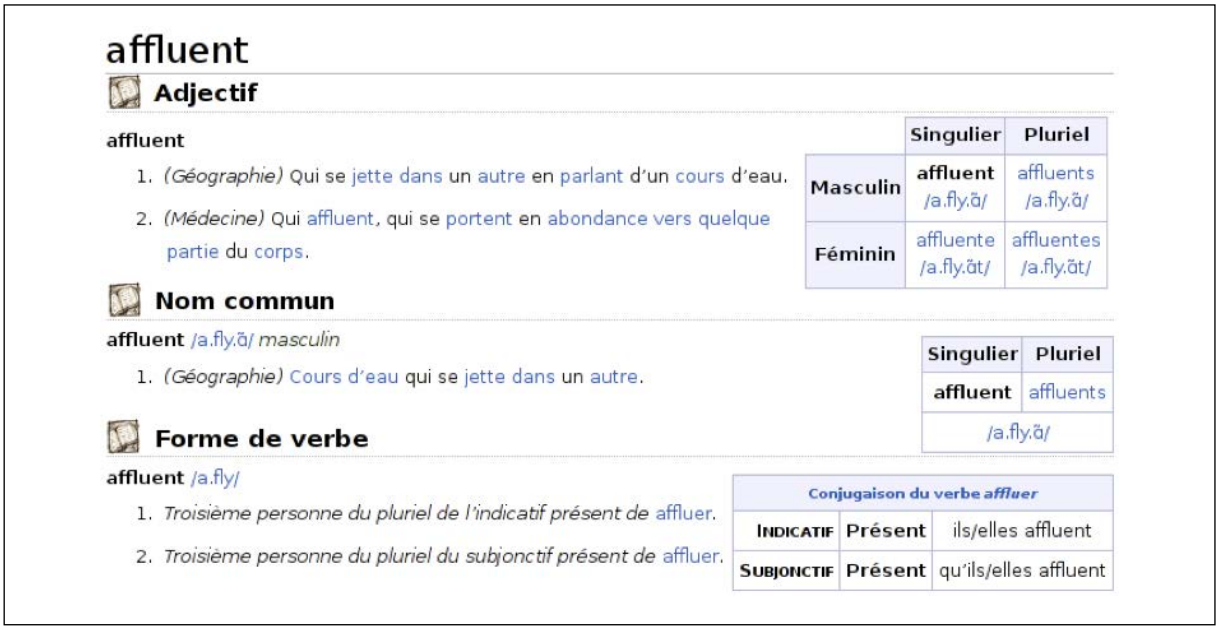
1 http://fr.wiktionary.org

2 GLÀFF and PsychoGLÀFF are freely available from the REDAC website: http://redac.univ-tlse2.fr/lexicons/

Wiktionary in the ontology Onto.PT (Gonçalo Oliveira and Gomes 2010). Serasset (2012) designed Dbnary, an open-source resource containing "easily extractable entries". The French subpart of this resource contains 260,467 entries. Works led by Meyer and Gurevych (2012) and Gurevych et al. (2012) resulted in German and English ontologies based on Wiktionary. Sajous et al. (2010) has made available a structured XML version of this lexicon for French and English, called WiktionaryX.[3]

Although Wiktionnaire presents interesting features (unprecedented coverage, definitions, phonemic transcriptions, semantic relations, translations, free license),[4] the information it contains is difficult to extract. This probably explains the relatively small number of works using it. Wiktionnaire, as other Wiktionary's language editions, is released as an "XML dump", where XML only marks the macrostructure. The microstructure is encoded in a format called *wikicode*, inherent in the content management system *MediaWiki*. This format has no formally defined syntax, evolves over time, and is not stable from one language edition to another. This underspecified syntax makes therefore the automatic information extraction from the collaborative dictionary uneasy: multiple deviations from a "prototypical article" should be expected, as well as missing information, redundancy and inconsistency.

Figure 1 shows the entry *affluent* (adjective and noun 'affluent', and two inflection forms of the verb *affluer* 'to flow into/to pour in') as it is visible in Wiktionnaire. The corresponding wikicode of this article is shown in Figure 2. Inflected forms may appear in the article related to their lemma (as it is the case in Figure 1). They may also have a dedicated page (cf. Figures 3 and 4).


affluent

Adjectif

affluent

1. (Géographie) Qui se jette dans un autre en parlant d'un cours d'eau.
2. (Médecine) Qui affluent, qui se portent en abondance vers quelque partie du corps.

	Singulier	Pluriel
Masculin	affluent /a.fly.ɑ̃/	affluents /a.fly.ɑ̃/
Féminin	affluente /a.fly.ɑ̃t/	affluentes /a.fly.ɑ̃t/

Nom commun

affluent /a.fly.ɑ̃/ masculin

1. (Géographie) Cours d'eau qui se jette dans un autre.

Singulier	Pluriel
affluent	affluents
/a.fly.ɑ̃/	

Forme de verbe

affluent /a.fly/

1. Troisième personne du pluriel de l'indicatif présent de affluer.
2. Troisième personne du pluriel du subjonctif présent de affluer.

Conjugaison du verbe *affluer*		
Indicatif	Présent	ils/elles affluent
Subjonctif	Présent	qu'ils/elles affluent

Figure 1: Article of the word *'affluent'* in Wiktionnaire.

3 WiktionaryX, is an XML version of the English and French editions of Wiktionary, freely available at http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html

4 For a more comprehensive description of the Wiktionnaire, see (Navarro et al. 2009) and (Sajous et al. 2010; 2013b).

The table of the adjective inflected forms (top-right in Figure 1) is not explicitly present in the wikicode, but is generated by the template {{fr-accord-cons|a.fly.ɑ̃|t}} (cf. Figure 2). There are hundreds of similar patterns in the wikicode. An example of the non-systematic wikicode's format and resulting article's layout can be seen in Figure 3: unlike the template {{f}} that defines the feminine gender of the form, there is no template specifying the grammatical number. The number can only be extracted by parsing the definition *"Féminin singulier"*. The heterogeneity of the wikicode also concerns the phonemic transcriptions: they occur sometimes in the *Ligne de forme* (the line following the part of speech heading), as in Figure 3 for *'affluente'*, and sometimes, on the contrary, they are specified in a separate *"Prononciation"* section as in Figure 4 for *'affluentes'*.

To build GLÀFF and PsychoGLÀFF, we automatically extracted the inflected forms and lemmas in their dedicated pages, and detected the inflection templates. We also identified the phonemic transcriptions wherever they occur. We finally parsed the conjugation tables (cf. Figure 5). We thus collected as much (possibly redundant) information as possible and applied some heuristics to automatically detect major inconsistencies.

```
{{-adj-|fr}}
{{fr-accord-cons|a.fly.ɑ̃|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]]
d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ɑ̃}}

{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui|}}
'''affluent''' {{pron|a.fly|fr}}
# ''3ème pers. du pluriel de l'indicatif présent de'' [[affluer]].
# ''3ème pers. du pluriel du subjonctif présent de'' [[affluer]].

{{-pron-}}
{| class="wikitable"
| Adjectif et nom commun
* {{pron-rég|France|ɛ̃.n_a.fly.ɑ̃|titre=un affluent}}
|-
| Forme du verber affluer
* {{pron-rég|France (Île-de-France)|a.fly}}
```

Figure 2: Wikicode of the article *'affluent'*.

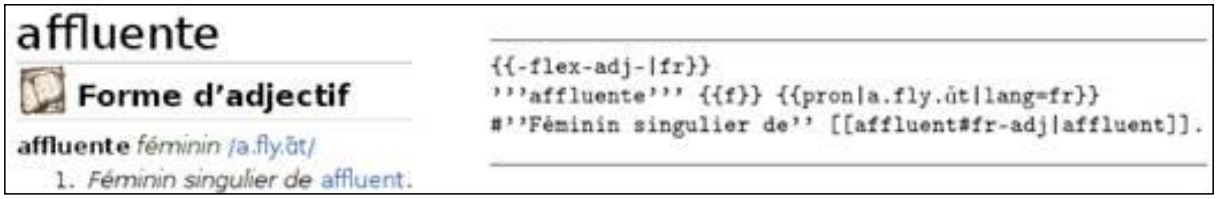


Figure 3: Article and wikicode of 'affluente'.

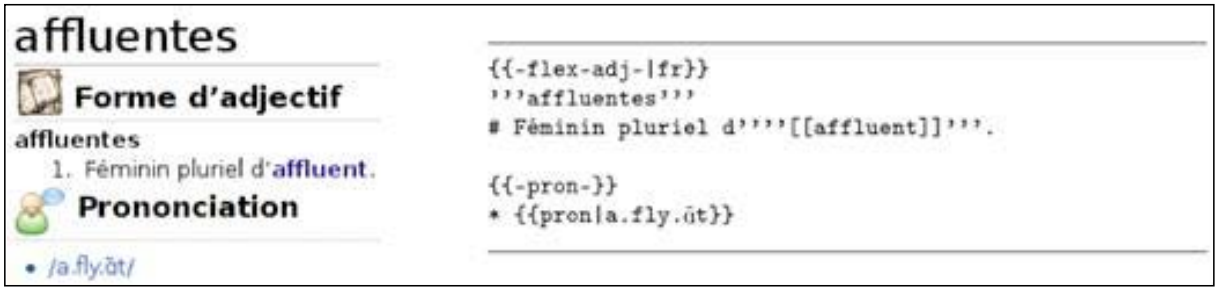
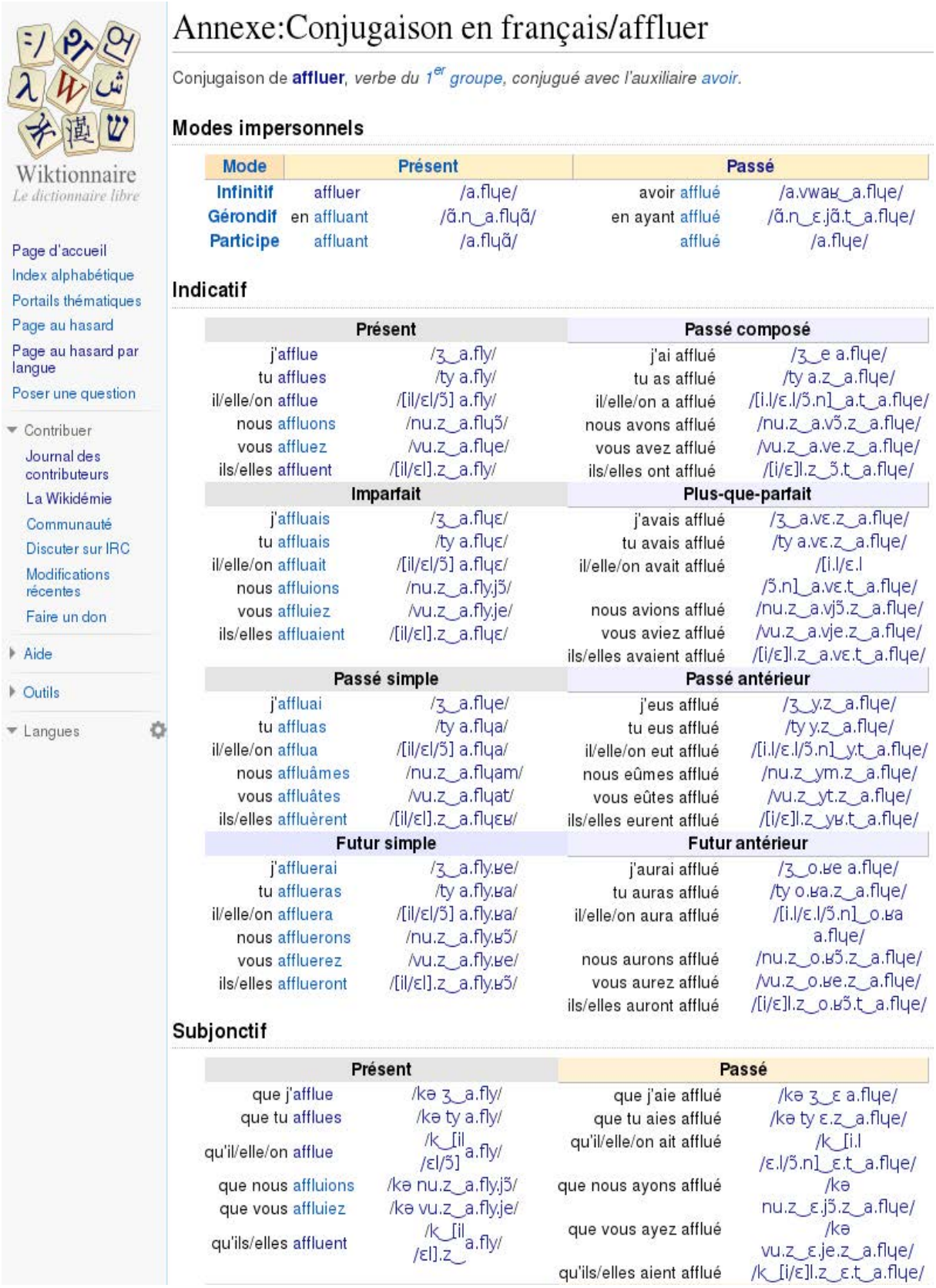


Figure 4: Article and wikicode of 'affluentes'.



Wiktionnaire
Le dictionnaire libre

Page d'accueil
Index alphabétique
Portails thématiques
Page au hasard
Page au hasard par langue
Poser une question

Contribuer
Journal des contributeurs
La Wikidémie
Communauté
Discuter sur IRC
Modifications récentes
Faire un don

Aide

Outils

Langues

Annexe:Conjugaison en français/affluer

Conjugaison de **affluer**, *verbe du 1er groupe, conjugué avec l'auxiliaire avoir.*

Modes impersonnels

Mode	Présent		Passé	
Infinitif	affluer	/a.flye/	avoir afflué	/a.vwaʁ_a.flye/
Gérondif	en affluant	/ɑ̃.n_a.flyɑ̃/	en ayant afflué	/ɑ̃.n_ɛ.jɑ̃.t_a.flye/
Participe	affluant	/a.flyɑ̃/	afflué	/a.flye/

Indicatif

Présent		Passé composé	
j'afflue	/ʒ_a.fly/	j'ai afflué	/ʒ_e a.flye/
tu affluues	/ty a.fly/	tu as afflué	/ty a.z_a.flye/
il/elle/on afflue	/[il/ɛl/õ] a.fly/	il/elle/on a afflué	/[i.l/ɛ.l/õ.n]_a.t_a.flye/
nous affluons	/nu.z_a.flyõ/	nous avons afflué	/nu.z_a.võ.z_a.flye/
vous affluez	/vu.z_a.flye/	vous avez afflué	/vu.z_a.ve.z_a.flye/
ils/elles affluent	/[il/ɛl].z_a.fly/	ils/elles ont afflué	/[i/ɛ]l.z_õ.t_a.flye/
Imparfait		**Plus-que-parfait**	
j'affluais	/ʒ_a.flyɛ/	j'avais afflué	/ʒ_a.vɛ.z_a.flye/
tu affluais	/ty a.flyɛ/	tu avais afflué	/ty a.vɛ.z_a.flye/
il/elle/on affluait	/[il/ɛl/õ] a.flyɛ/	il/elle/on avait afflué	/[i.l/ɛ.l /õ.n]_a.vɛ.t_a.flye/
nous affluions	/nu.z_a.fly.jõ/	nous avions afflué	/nu.z_a.vjõ.z_a.flye/
vous affluiez	/vu.z_a.fly.je/	vous aviez afflué	/vu.z_a.vje.z_a.flye/
ils/elles affluaient	/[il/ɛl].z_a.flyɛ/	ils/elles avaient afflué	/[i/ɛ]l.z_a.vɛ.t_a.flye/
Passé simple		**Passé antérieur**	
j'affluai	/ʒ_a.flye/	j'eus afflué	/ʒ_y.z_a.flye/
tu affluas	/ty a.flya/	tu eus afflué	/ty y.z_a.flye/
il/elle/on afflua	/[il/ɛl/õ] a.flya/	il/elle/on eut afflué	/[i.l/ɛ.l/õ.n]_y.t_a.flye/
nous affluâmes	/nu.z_a.flyam/	nous eûmes afflué	/nu.z_ym.z_a.flye/
vous affluâtes	/vu.z_a.flyat/	vous eûtes afflué	/vu.z_yt.z_a.flye/
ils/elles affluèrent	/[il/ɛl].z_a.flyɛʁ/	ils/elles eurent afflué	/[i/ɛ]l.z_yʁ.t_a.flye/
Futur simple		**Futur antérieur**	
j'affluerai	/ʒ_a.fly.ʁe/	j'aurai afflué	/ʒ_o.ʁe a.flye/
tu affluueras	/ty a.fly.ʁa/	tu auras afflué	/ty o.ʁa.z_a.flye/
il/elle/on affluera	/[il/ɛl/õ] a.fly.ʁa/	il/elle/on aura afflué	/[i.l/ɛ.l/õ.n]_o.ʁa a.flye/
nous affluerons	/nu.z_a.fly.ʁõ/	nous aurons afflué	/nu.z_o.ʁõ.z_a.flye/
vous affluerez	/vu.z_a.fly.ʁe/	vous aurez afflué	/vu.z_o.ʁe.z_a.flye/
ils/elles affluueront	/[il/ɛl].z_a.fly.ʁõ/	ils/elles auront afflué	/[i/ɛ]l.z_o.ʁõ.t_a.flye/

Subjonctif

Présent		Passé	
que j'afflue	/kə ʒ_a.fly/	que j'aie afflué	/kə ʒ_ɛ a.flye/
que tu affluues	/kə ty a.fly/	que tu aies affluué	/kə ty ɛ.z_a.flye/
qu'il/elle/on afflue	/k_[il /ɛl/õ] a.fly/	qu'il/elle/on ait affluué	/k_[i.l /ɛ.l/õ.n]_ɛ.t_a.flye/
que nous affluions	/kə nu.z_a.fly.jõ/	que nous ayons affluué	/kə nu.z_ɛ.jõ.z_a.flye/
que vous affluiez	/kə vu.z_a.fly.je/	que vous ayez affluué	/kə vu.z_ɛ.je.z_a.flye/
qu'ils/elles affluent	/k_[il /ɛl].z_a.fly/	qu'ils/elles aient affluué	/k_[i/ɛ]l.z_ɛ.t_a.flye/

Figure 5: Conjugation table of the verb affluer (extract).

3 GLÀFF

In this section, we summarize some relevant characteristics of GLÀFF, first introduced in (Sajous et al. 2013a), from which PsychoGLÀFF is derived. The latest version of GLÀFF includes nouns, verbs, adjectives, adverbs, and function words. As can be seen in Figure 6, GLÀFF specifies for each entry:

- the wordform;
- the lemma;
- the part of speech and morphosyntactic features in GRACE format (Rajman et al. 1997);
- the phonological transcription(s) (when specified in Wiktionnaire) in IPA and in SAMPA with syllable boundaries.

```
affluent|Ncms|affluent|a.fly.ɑ̃|a.fly.A~
affluent|Afpms|affluent|a.fly.ɑ̃|a.fly.A~
affluents|Afpmp|affluent|a.fly.ɑ̃|a.fly.A~
affluents|Ncmp|affluent|a.fly.ɑ̃|a.fly.A~
affluent|Vmip3p-|affluer|a.fly|a.fly
affluent|Vmsp3p-|affluer|a.fly|a.fly
```

Figure 6: Extract of GLÀFF.

3.1 Coverage

GLÀFF differs from the lexicons currently used in NLP and psycholinguistics by its exceptional size. Table 1 shows the number of inflected forms and lemmas for simple words (only letters) and non-simple words (containing spaces, dashes or digits) in five different French lexicons. GLÀFF contains from 3 to 4 times more tokens and from 3 to 9 times more inflected forms than the other lexicons.

	Categorized inflected forms			**Categorized lemmas**		
	Simple	Non simple	Total	Simple	Non simple	Total
Lexique	147,912	4,696	152,608	46,649	3,770	50,419
BDLex	431,992	4,360	436,352	47,314	1,792	49,106
Lefff	466,668	3,829	470,497	54,214	2,303	56,517
Morpha-lou	524,179	49	524,228	65,170	7	65,177
GLÀFF	1,401,578	24,270	1,425,848	172,616	13,466	186,082

Table 1: Size of five French lexicons (counting only nouns, verbs, adjectives and adverbs).

Table 2 reports the intersection of GLÀFF with the other lexicons. We observe that the magnitude of the intersection depends on the size of the lexicons: the bigger a lexicon, the larger its intersection

with the other ones. Three groupings emerge: Lexique has the smallest coverage, only containing 9% of GLÀFF and 22% to 26% of the entries of the other lexicons. BDLex, Lefff and Morphalou cover 76% to 80% of Lexique and about 30% of GLÀFF. Finally GLÀFF is clearly on top with coverage of 85% to 93%. In total, its coverage is 5% to 65% higher than the other lexicons.

	Lexique	**BDLex**	**Lefff**	**Morphalou**	**GLÀFF**
Lexique	-	26.03	25.20	22.46	8.95
BDLex	76.02	-	79.87	70.40	28.75
Lefff	79.50	86.28	-	72.32	30.04
Morphalou	79.58	85.43	81.24	-	32.03
GLÀFF	84.83	93.26	90.23	85.66	-

Table 2: Intersection of five French lexicons (% of the categorized inflected forms).

Size is a crucial aspect of the lexicons used for research in derivational and inflectional morphology or, more generally, in the development of NLP tools such as morphosyntactic taggers and parsers. In order to asses that GLÀFF's largest size is actually useful, we compared the five lexicons with the vocabulary of four corpora of various types. Frantext 20^{e} is constituted by 515 novels of 20th century French literature containing 30 million words. LM10 is a 200 million word corpus made up of the archives of the newspaper Le Monde from 1991 to 2000. The third corpus, containing 260 million words, consists of the articles from the French Wikipedia. Finally, FrWaC (Baroni et al. 2009) is a 1.6 billion words corpus of French web pages. Table 3 shows the coverage of the five lexicons with respect to the four corpora.

Threshold: frequency ≥		1	2	5	10	100	1000
Frantext	#forms	145,437	95,189	61,813	43,919	10,767	1,376
	Lexique	66.76	84.35	94.00	96.91	99.15	99.27
	BDLex	70.86	84.69	92.47	95.74	99.12	99.20
	Lefff	71.89	85.63	93.21	96.21	99.08	98.90
	Morphalou	73.93	86.66	93.29	96.00	98.48	97.09
	GLÀFF	76.92	88.57	94.54	96.72	98.77	98.76
LM10	#forms	300,606	172,036	106,470	77,936	29,388	83.21
	Lexique	29.59	47.28	65.23	76.31	93.81	98.58
	BDLex	37.77	55.79	71.76	80.93	95.53	98.69
	Lefff	39.64	58.22	74.33	83.20	95.99	98.90
	Morphalou	39.06	56.82	71.92	80.32	93.27	97.48
	GLÀFF	45.24	63.83	78.63	86.23	96.46	98.68
Wikipedia	#forms	953,920	435,031	216,210	136,531	35,621	7,956
	Lexique	9.13	18.27	31.52	43.03	78.58	95.72
	BDLex	12.29	22.89	36.80	48.04	79.39	95.33
	Lefff	12.88	23.94	38.26	49.65	80.57	95.71
	Morphalou	13.05	23.96	37.87	48.87	78.74	94.16
	GLÀFF	16.42	29.00	44.13	55.45	83.21	96.10
FrWaC	#forms	1,624,620	846,019	410,382	255,718	74,745	22,100
	Lexique	5.83	10.85	20.84	30.81	66.00	89.47
	BDLex	9.36	15.85	27.28	37.48	69.61	90.03
	Lefff	9.85	16.67	28.57	39.16	71.61	91.16
	Morphalou	10.09	16.89	28.53	38.68	69.36	88.51
	GLÀFF	13.13	21.13	34.29	45.35	76.39	92.76

Table 3: Lexicon/corpus coverage (% of non-categorized inflected forms).

The vocabulary is restricted to the forms that occur at least once, 2, 5, 10, 100 and 1000 times. The ranking of the corpora by coverage is the same for the five lexicons. Although their size affects the order, their nature is also crucial. For example, FrWaC being a collection of web pages, it contains a large number of "noisy" forms (foreign words, missing or extra spaces, missing diacritics, random spelling, etc.). Again, we see the division of lexicons into three groups. BDLex, Lefff and Morphalou have a quite close coverage. Except for Frantext 20^{e}, Lexique has the smallest coverage. GLÀFF has the largest coverage for all corpora, except for LM10 at the 1000 threshold where it is surpassed by Lefff by 0.2%. The best coverage of Lexique for the Frantext 20^{e} corpus, above the 10 threshold, while it has the weakest coverage in all other cases, is explained by the design of its vocabulary, extracted from this corpus. For the other corpora and up to the 100 threshold, the size of GLÀFF explains its larger coverage with re-

spect to the other lexicons (at the threshold 1, 14% to 53% larger for LM10 and 30% to 120% larger for FrWaC; at the threshold 10, 4% to 16% for LM10 and 15% to 47% for FrWaC). NLP tools that integrate GLÀFF should therefore offer an improved performance in the treatment of these corpora. In a qualitative study described in (Sajous et al. 2014), we observed that GLÀFF specific entries contains not only rare neologisms, but also very common words such as *attractivité* 'attractivity', *brevetabilité* 'patentability', *diabolisation* 'demonization', *employabilité* 'employability', *homophobie* 'homophobia', *hébergeur* 'host', fatwa, *institutionnellement* 'institutionally', *anticorruption* 'anti-corruption', etc. missing from the other lexicons.

3.2 Phonemic transcriptions

GLÀFF provides a phonemic transcription for about 90% of the entries. We evaluated the consistency of these transcriptions with respect to those of BDLex and Lexique (after conversion into IPA encoding).

Tables 4a to 4c report the top ten variations between pairs from the three lexicons. We only considered one phoneme differences, ignoring syllabification. The differences in transcriptions between GLÀFF and the other two lexicons are comparable to the differences observed between BDLex and Lexique. In particular, these differences are mostly due to the distinctions between the mid vowels, i.e. the front-mid vowels: [e] (close-mid) vs. [ɛ] (open-mid) and the back-mid vowels: [o] (close-mid) vs. [ɔ] (open-mid). This alternation is a well-known aspect of French phonology resulting from diatopic variations (North vs. South), as described in (Detey et al. 2010). Such expected oppositions account for about 91% of the divergences between BDLex and Lexique. Table 5 reports the percentage of identical phonological transcriptions shared by the lexicons and the percentage of the 'comparable' phonological transcriptions, i.e. disregarding the distinction between close-mid and open-mid vowels. GLÀFF and Lexique give identical transcriptions for 79.5% of entries whereas the percentage between GLÀFF and BDLex is lower, at 61.7%. Table 5 also reports the results of the comparison of syllabification in the three lexicons (performed on the basis of identical transcriptions only). This comparison shows that the three lexicons are quite similar with respect to syllabification (98%).

Comparing GLÀFF with the major resources that contain the same type of information clearly shows that the overall quality of the lexicon is quite satisfactory and is in all respect comparable to those of these resources.

Oper.	Phonemes	%	∑ %
r	ɛ/e	48.18	48.18
r	ɔ/o	32.17	80.36
r	o/ɔ	11.02	91.37
r	y/ɥ	1.83	93.21
r	ə/ø	1.44	94.64
r	ə/œ	1.39	96.03
r	u/w	0.84	96.87
r	b/p	0.73	97.61
r	s/z	0.51	98.12
d	j	0.25	98.37

(a) BDLex/Lexique

Oper.	Phonemes	%	∑ %
r	ɔ/o	60.03	60.03
i	ə	14.18	74.21
r	e/ɛ	6.90	81.11
r	ɛ/e	4.98	86.09
r	ɑ/a	4.92	91.01
r	s/z	1.25	92.26
r	ə/ø	0.91	93.17
r	œ/ø	0.47	93.64
i	i	0.42	94.06
r	o/ɔ	0.38	94.44

(b) GLÀFF/Lexique

Oper.	Phonemes	%	∑ %
r	e/ɛ	66.46	66.46
r	ɔ/o	10.58	77.05
i	ə	5.90	82.96
r	o/ɔ	4.36	87.32
r	ɑ/a	3.84	91.17
r	ɥ/y	1.61	92.78
r	œ/ə	1.09	93.88
r	ø/ə	0.86	94.74
i	i	0.84	95.58
r	w/u	0.79	96.38

(c) GLÀFF/BDLex

Table 4: The 10 most frequent differences in phonemic transcriptions (Operations: r = replacement, i = insertion, d = deletion).

			Phonemic transcriptions		Syllabification
Lexicons		Intersection	Identical	Comparable	Identical
BDLex	Lexique	112,439	58.31	96.88	98.92
GLÀFF	Lexique	123,630	79.50	97.81	98.48
GLÀFF	BDLex	396,114	61.72	96.88	98.30

Table 5: Inter-lexicon agreement: phonemic transcriptions and syllabification.

4 From GLÀFF to PsychoGLÀFF

4.1 Overview

Our goal in creating PsychoGLÀFF is to provide psycholinguists with a set of features related to the formal aspects of the lexicon entries. For this purpose, we selected from GLÀFF only forms having non-zero frequency in at least one of the corpora mentioned in section 3.1. This means that PsychoGLÀFF only contains lexical entries attested in the corpora, amounting to about 340,000 forms for 120,000 lemmas.

In addition to GLÀFF's features, PsychoGLÀFF includes the following information for each entry:

- the absolute and relative frequencies of the wordform and of the lemma in the aforementioned French corpora (Frantext 20^{e}, LM10 and FrWac);

- the length of the wordform (number of characters);
- the length of the phonological transcription(s) (number of phonemes);
- the syllabification and the CV structure of the wordform;
- the number of syllables;
- the geometric mean of the conditional character probabilities of bigrams, which calculates the probability of the bigram occurring given the preceding bigram;
- the geometric mean of the conditional character probabilities of trigrams, which calculates the probability of the trigram occurring given the preceding trigram;
- the geometric mean of the conditional character probabilities of 4-grams, which calculates the probability of the 4-gram occurring given the preceding 4-gram;
- the geometric mean of the conditional phoneme probabilities respectively calculated for bigrams, trigrams and 4-grams.
- the size of the orthographic neighbourhood, i.e. the number of words in the lexicon differing by one character (via deletion, insertion, or substitution);
- the size of the phonological neighborhood, i.e. the number of words differing from the phonological transcription by one phoneme (via deletion, insertion, or substitution);
- the size of the ratio between the number of consonants and syllables composing the phonological form. This score is meant to provide an estimate of the 'syllabic complexity' of the form.

The *n*-gram conditional probability represents a measure of phonotactic occurrence, defining the likelihood of occurrence of *n*-grams in French. This kind of measure is expected to be particularly helpful for the design of experimental stimuli in lexical access experiments (Storkel and Hoover 2011).

4.2 Comparison with Lexique

We compare hereafter, in terms of coverage and word frequencies, PsychoGLÀFF and Lexique, the most frequently used French lexicon in psycholinguistics.

Being directly extracted from GLÀFF, PsychoGLAFF stands out with respect to the lexicons currently used in psycholinguistics mostly for its size, as it counts 337,572 entries. Table 6 shows the number of inflected forms and lemmas of Lexique and PsychoGLAFF. The relative coverage of these lexicons is reported in Table 7.

	Categorized inflected forms	Categorized lemmas
Lexique	153,934	50,419
PsychoGLÀFF	337,572	121,021

Table 6: Size of Lexicons (restricted to nouns, verbs, adjectives and adverbs)

	Lexique	PsychoGLÀFF
Lexique		36.1 %
PsychoGLÀFF	78.9 %	

Table 7: Lexicons relative coverage (% of categorized inflected forms)

We observe that PsychoGLÀFF is more than twice larger than Lexique (for both inflected forms and lemmas) and has a total coverage of about 79% with respect to Lexique (which covers only 36.1% of the inflected forms of PsychoGLÀFF). PsychoGLÀFF reports the absolute and relative frequencies for its wordforms and lemmas. Frequencies are calculated on the basis of three stylistically different corpora of written French: the abovementioned Frantext20^{e}, LM10 and FrWaC (literature, newspaper and web corpora). Lexique reports word frequency estimates too. It exploits two smaller corpora: a) a written corpus made up of 218 books from Frantext 20^{e}; b) a corpus of French subtitles for 9,474 movies and television series, assumed to be more representative of spoken French.

While GLÀFF and PsychoGLÀFF frequencies are exclusively based on written French, Lexique mixes together spoken-like and written resources for the calculation of wordform and lemma frequencies. Although the corpora used by the two lexicons have very different sizes, we attempted a comparison of the PsychoGLÀFF's frequencies with respect to the frequencies reported in Lexique, looking only at the intersection of the two lexicons.

Table 8 reports the correlation between the normalized frequencies of wordforms in PsychoGLÀFF (separately for Frantext, LM10 and FrWac) and Lexique (separately for books and movie subtitles). The data were normalized by one million words. It is not surprising that the correlation between Frantext's frequencies and Lexique's book frequencies is quite high (Pearson's coefficient $\rho = .81$), the latter being a sub-corpus of Frantext 20^{e}. Although PsychoGLÀFF frequencies are based exclusively on written corpora, we found a statistically significant correlation $\rho = .68$ between Lexique's subtitle frequencies and the Frantext frequencies (the value slightly decreases for the subtitles/FrWac correlation). This seems to indicate that the lexical coverage of PsychoGLÀFF, though based on written sources, is comparable to a relevant extent to the coverage of corpora specifically devoted to spoken French.

		Lexique	
		Subtitles	Books
PsychoGLÀFF	Frantext	.68	.81
	LM10	.62	.59
	FrWac	.67	.62

Table 8: PsychoGLÀFF/Lexique correlations with respect their normalized frequency values.

An additional property of PsychoGLÀFF worth noting to is the presence of infrequent lexical items. This feature clearly derives from the nature of Wiktionnaire: being an online dictionary, it has not to conform to the same size constraints as printed ones. Bootstrapped by importation of articles from public domain dictionaries, it contains dated entries. Finally, being crowdsourced, it is regularly updated and contains general-domain neologisms, as well as subculture vocabulary and technical terms (Sajous et al. 2014). As a consequence, PsychoGLÀFF contains a large number of specific entries.

The bigger the corpus, the more low-frequency lexical items are likely to be included (while the size of the corpus is not likely to have a strong impact on the number of those words that are frequent or

very frequent in a language, representing to a certain extent the 'essential lexicon' of that language). The Figure 7 illustrates this point by showing the distribution of different frequency intervals for both Lexique's and PsychoGLÀFF's sub-corpora. A normalized frequency range of 10.01 or more corresponds to very high words frequency and is situated at the right edge of the graph. A frequency range of 0.01-0.1 corresponds to very low words frequency and is situated at the left edge of the figure.
Six intermediate ranges capture the frequency differences of the entire lexicon. The figure shows that the distribution of the frequency intervals is approximately the same for Lexique and PsychoGLÀFF, with the significant exception of the least frequent word class (< 0.2), for which the number of lexical items in PsychoGLÀFF is almost twice as large as that of Lexique. At the same time, PsychoGLÀFF contains many words of ordinary usage that are absent from Lexique, such as *acceptabilité* 'acceptability', *centralité* 'centrality', *Saturne* 'Saturn', etc. In this sense PsychoGLÀFF offers a much larger lexical repertoire not only in terms of tokens, but also in terms of types, which represents a particularly interesting feature for psycholinguistic studies and corpus investigations.

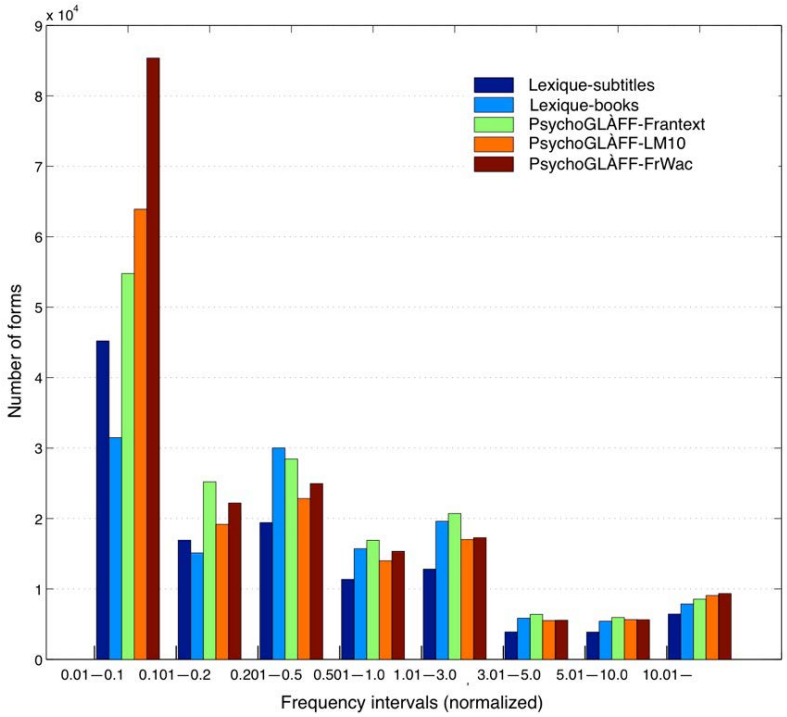


Figure 7: Distribution of forms with respect to their corpus frequency.

5 Conclusions and future directions

This paper presents a first version of PsychoGLÀFF, a large lexicon designed for psycholinguistic experimentation. PsychoGLÀFF was built on top of the inflectional and phonemic lexicon GLÀFF, itself acquired from Wiktionnaire, the French edition of the collaborative dictionary Wiktionary. In particular, PsychoGLÀFF contains the subset of GLÀFF's corpora-attested entries. This resource complements

the inflectional and phonological information present in GLÀFF with features needed for experimental material calibration including frequency, lexical neighborhood, syllabic complexity and phonotactic likelihood.

Like GLÀFF, PsychoGLÀFF is characterized by an exceptional coverage, much higher than those of comparable resources as Lexique on one hand and Morphalou and Lefff on the other. We also show that the "primary" information (parts of speech, phonemic transcriptions, frequency) of PsychoGLÀFF and GLÀFF has a satisfactory quality. PsychoGLÀFF is a free resource distributed under a copylefted license and is available to all psycholinguistic researchers working on French. We hope that it will soon be adopted by this community whose feedback will allow us to improve the resource and appropriately respond to its needs.

In the near future, we plan to improve PsychoGLÀFF on several aspects. One of them will concern the description completeness and consistency of the lexicon. An online interface, comparable to GLÀFFOLI (the GLÀFF OnLine Interface) will also be developed, that will enable users to query the lexicon and develop experimental material interactively.

6 References

Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, pp. 703–717, Lisboa, Portugal.

Baayen, R. H., Piepenbrock, R. and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM).* Philadelphia: Linguistic Data Consortium.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3): 209-226.

Boula De Mareuil, P., Yvon, F., D'Alessandro, C., Aubergé, V., Vaissière, J., and Amelot, A. (2000). A French Phonetic Lexicon with variants for Speech and Language Processing. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 273–276, Athens, Greece.

Clément, L., Lang, B. and Sagot. B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004),* pp. 1841–1844, Lisboa, Portugal.

Detey, S., Durand, J., Laks, B., and Lyche, C. (2010). *Les variétés du français parlé dans l'espace francophone*. Paris: Ophrys.

Gonçalo Oliveira, H. and Gomes, P. (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium*, pp. 199–211. IOS Press.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer C. M. and Wirth C. (2012). UBY - A Large-Scale Unified Lexical- Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France.

Meyer, M. C. and Gurevych, I. (2012). *OntoWiktionary - Constructing an Ontology from the Collaborative Online Dictionary Wiktionary*. In Semi-Automatic Ontology Development: Processes and Resources, chapter 6, pp 131–161. IGI Global.

New, B. (2006). Lexique 3: Une nouvelle base de données lexicales. *Actes de la 13e Conférence Traitement Automatique des Langues Naturelles (TALN 2006),* Louvain-la-Neuve, Belgium.

New, B., Brysbaert, M., Veronis, J. and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. In *Applied Psycholinguistics* (28): 661–677.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I. , Magistry, P. and Huang, C. R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the ACL Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. ACL-IJCNLP 2009, Singapore.

Pérennou, G. and De Calmès, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. In *Proceedings of ECST 1987*, pp. 1393–1396, Edinburgh, UK.

Rajman, M., Lecomte, J., and Paroubek, P. (1997). *Format de description lexicale pour le francçais. Partie 2 : Description morpho-syntaxique*. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

Romary, L., Salmon-Alt, S. and Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop on Electronic Dictionaries*, Geneva, Swiss.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., et Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Col- laboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S. (eds), *Advances in Natural Language Processing*, vol. 6233 of LNCS, 332–344. Springer.

Sajous, F., Hathout, N. and Calderone, B. (2013a). *GLÀFF, un Gros Lexique À tout Faire du Français*. Actes de la 20e Conference sur le Traitement Automatique des Langues Naturelles (TALN 2013), pp. 285-298, Les Sables d'Olonne, France.

Sajous, F., Navarro, E. Gaume, B. Prévot, L. and Chudy, Y. (2013b). Semi-automatic enrichment of crowdsourced synonymy networks: the WISIGOTH system applied to Wiktionary. *Language Resources and Evaluation*, 47(1): 63-96.

Sajous, F., Hathout, N. and Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Etudes et réalisations fondées sur le dictionnaire collaboratif. *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, Germany.

Schultz, T., Vu, N. T. and Schlippe, T. (2013). GlobalPhone: A multilingual text & speech database in 20 languages. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*, pp. 8126–8130, Vancouver, Canada.

Sérasset, G., (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Storkel, H. L. and Hoover, J. R. (2011). The influence of part-word phonotactic probability/neighborhood density on word learning by preschool children varying in expressive vocabulary. In *Journal of Child Language*, 38, 628-643

Zesch, T., Müller, C. and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

RIDIRE. Corpus and Tools for the Acquisition of Italian L2

Alessandro Panunzi, Emanuela Cresti, Lorenzo Gregori
University of Florence
alessandro.panunzi@unifi.it, elicresti@unifi.it, lorenzo.gregori@unifi.it

Abstract

This paper introduces the RIDIRE corpus, built by means of an open source tool (RIDIRE-CPI) for creating specifically designed web corpora through a targeted crawling strategy. The RIDIRE-CPI architecture combines existing open source tools with specifically developed modules, comprising a robust crawler, a user friendly web interface, several conversion and cleaning tools, an anti-duplicate filter, a language guesser, and a PoS-tagger. The RIDIRE corpus is a balanced Italian web corpus (1.5 billion tokens) designed for enhancing the study of Italian as a second language, while also being exploitable for lexicographic purposes. The targeted crawling was performed through content selection, metadata assignment, and validation procedures. These features allowed the construction of a large corpus with a specific design, covering a variety of language usage domains (News, Business, Administration and Legislation, Literature, Fiction, Design, Cookery, Sport, Tourism, Religion, Fine Arts, Cinema, Music). The RIDIRE query system allows research to be carried out on the whole corpus itself or on the sub-corpora. Specifically, available queries comprehend all the functions usually exploited in corpus-based lexicography: frequency lists, concordances and patterns, collocations, Sketches, and Sketch Differences.

Keywords: Corpus linguistics; Terminology; Collocations

1 Introduction

RIDIRE (acronym for *RIsorse Dinamiche dell'Italiano in Rete*, "Italian Dynamic Resources Online"; Moneglia & Paladini 2010) is a project which produced a large Italian language corpus, and an open-source tool for web corpora building and processing, named RIDIRE-CPI (Panunzi et al. 2012). The corpus - of 1.5 billion tokens - was built using web-crawling techniques and exploited the Italian content of the Internet. The corpus is now available online and is integrated with computational tools for the exploitation of vast corpora to enhance language usage in L2 Italian learners. RIDIRE is designed for use by both teachers and learners, who will be able to profit from access to a database of representative texts which characterize Italian culture. The database collects a massive amount of freely available content, covering a selection of domains which are relevant to Italian identity: law, religion, politics, literature, trade, administration, information, design, food, fashion. To reach this goal, a distributed

crawling infrastructure was created and a targeted crawling strategy pursued. This document will summarize the corpus design for the resource as well as the crawling techniques and processing tools used for deriving language corpora from the web. Also presented are examples of queries that are relevant for both learners and lexicographers.

Figure 1: The RIDIRE resource home page.

2 Corpus Design Strategy

Different kinds of projects have been carried out to exploit the language data populating the web (Kilgarriff & Greffenstette 2003, Sharoff 2006). Among these, the WaCky initiative (Baroni et al. 2009) and the Italian web corpus ItWaC are important antecedents. More recently a new generation of web corpora have been created and processed with boilerplate cleaning and de-duplication tools and are available through Sketch Engine for a large number of languages (Kilgarriff et al. 2004); these are identified through their target size as the TenTen collection: 10 billion word corpora (10^{10}). Such initiatives resulted in the development of dedicated software for crawling (Heritrix), text-processing, cleaning, and the large-scale use of existing technologies for morpho-syntactic annotation (TreeTagger) and online corpus querying (CQPweb). These technologies have been used in RIDIRE and adapted to its specific goals.

The RIDIRE project aimed to build an online database representative of a wide and significant Italian language universe which would have value for sourcing information on the use of Italian in various aspects of life and culture, for linguistic/lexicographic researches, and for didactic purposes. To build such a resource involved two corpus design requirements which did not characterize the web corpora collected in previous initiatives: a) the selection of linguistic resources which document the main domains of usage (life and culture); b) the enrichment of the resource with metadata which enables a perspicuous querying of the database in each specific domain.

The collection focuses on two sets of non-hierarchically structured domains, selected for their pragmatic relevance to the use of the Italian language. The first set is constituted by general non-semantic fields, in which language characterizes its function (up to 400 million words for each domain):

- News
- Business
- Administration and Legislation

The second consists of semantic fields in which Italian excellence is largely recognized (up to 100 million words for each domain):

- Literature
- Fiction
- Design
- Cookery
- Sport
- Tourism
- Religion
- Fine Arts
- Cinema
- Music

The possibility for learners to find specific information on the language usage characterizing a domain should enhance their ability to find the right expressions for it. From a lexicographic point of view, the presence of different domains allows the derivation of specific uses of a word and the description of its semantic variation across the different domains of language use. Table 1 and Figure 2 show the structure of the corpus and the quantitative measures for each domain.

DOMAINS	# WEBSITES	# PAGES	# TOKENS	# WORDS
Functional (total)	189	976,460	854,388,230	747,268,841
Information	27	550,169	216,431,868	186,577,769
Economics and Business	123	226,535	179,710,476	161,377,152
Administration and Law	39	199,756	458,245,886	399,313,920
Semantic (total)	816	907,374	660,243,564	566,229,119
Sport	49	138,235	98,172,470	82,695,548
Architecture and Design	142	136,725	93,822,675	81,235,939
Cooking	20	123,376	52,784,045	45,523,096
Cinema	25	122,850	51,466,145	44,370,692
Music	195	113,015	12,906,213	106,287,283
Fashion	103	74,584	24,645,980	21,690,140
Visual Arts	118	70,601	56,517,442	48,929,903
Religion	51	66,053	72,454,492	62,291,806
Literature and Theatre	113	61,935	85,474,102	73,204,712
Total	**2,010**	**3,767,668**	**1,514,631,794**	**1,313,497,960**

Table 1: Number of crawled websites, pages, tokens and words per domain.

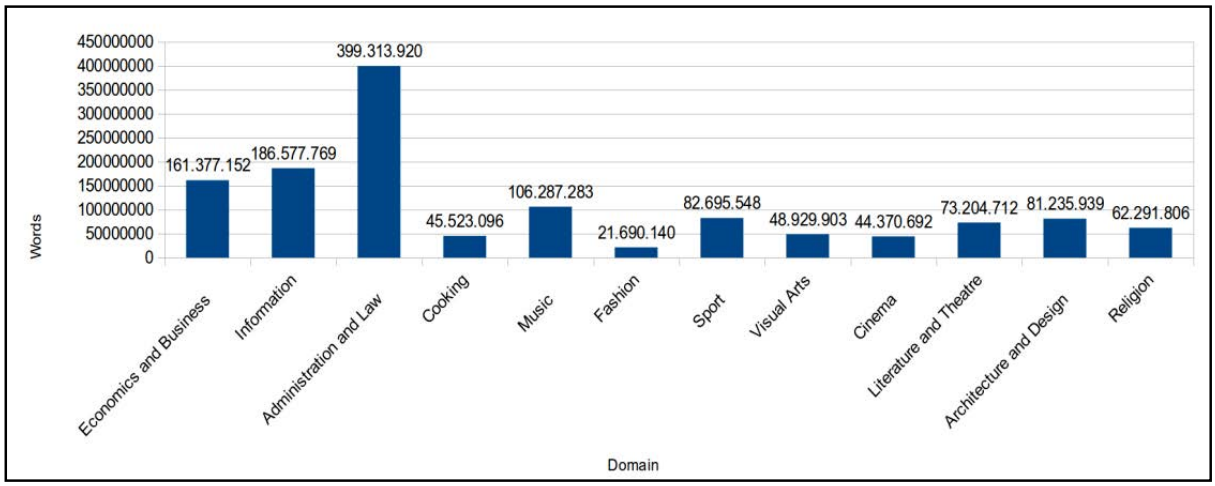


Figure 2: Words per Domain chart of RIDIRE Corpus.

3 The Crawling Infrastructure

The gathering of specific linguistic data for each sub-corpus requires a targeted crawling strategy performed by different teams of experts. The tool developed within the RIDIRE project for the crawling and the processing of the web resources (RIDIRE-CPI) is now open source and its user-friendly web interface is specifically intended to allow collaboration between users unskilled in web technology and text processing, working in a distributed environment. The application comprises:

- the crawling process
- the mapping of the resource in a MySQL database
- user interaction via web interface

RIDIRE-CPI has a modular architecture (see Figure 3), which is made up of:

- a web crawler
- a web interface for crawling management and validation
- conversion tools
- HTML cleaner tools
- anti-duplicate filters
- a language guesser
- a PoS-tagger

The crawling activity, as in the other cited web corpus initiatives, makes use of the Heritrix web crawler (version 3.1.1). However RIDIRE-CPI configures it via the web interface, making it suitable for use in a distributed environment. The crawling activity itself is structured into "jobs" (fully configured crawling sessions) in which the user determines three sets of parameters. First, the user selects the seed URLs from which the crawling activity starts. Then the formats of the resources that should be

downloaded are specified. In addition to HTML, RIDIRE-CPI is able to process TXT, RTF, DOC, and PDF documents. This feature is crucial, since many linguistically relevant resources from the web are not contained in web pages, but in documents of varying formats. The third set of parameters determine the strategy for the selection of content from websites. This step is important in downloading resources which comply with the representativeness requirement, since the reference unit for text on the web (when representing the language of a particular domain) is the web page rather than the website. As a matter of fact, only a subset of the web pages from a given site give information strictly concerning the specific domain to which the site belongs. Within the step, the user selects and/or discards the "resources" specifying

- which found URLs the crawler has to add to the queue ("URL to be navigated");
- which resources the crawler has to download to the file system ("URL to be saved")

Once all the parameters are defined by the user, the crawler starts from the first seed URL, which is put in the processing queue. The crawler accesses the web page relative to the first URL in the queue, extracts all the links that match the "URL to be navigated" rules and saves them in the queue; then, if the page is a "URL to be saved", the crawler downloads the web page content and stores it on the file system. Finally, it goes back to the first step and proceeds recursively until the processing queue is empty.

To maximize the precision of the process, the user can decide to insert a list of complete URLs, to specify website areas with path substrings (any URL containing one of these strings) or to write a customized regular expression that matches desired page URLs. For instance, in Figure 4 the user decided to crawl the website http://musica.atuttonet.it, getting HTML pages only, and further navigating to any link found (this option is set with a regular expression in the *Pattern* field), downloading any pages that do not contain the word *varie* or *artisti* in the URL.

In this stage no technical competence is required, but a pre-analysis of the website(s) is necessary to ensure only relevant information is retrieved.

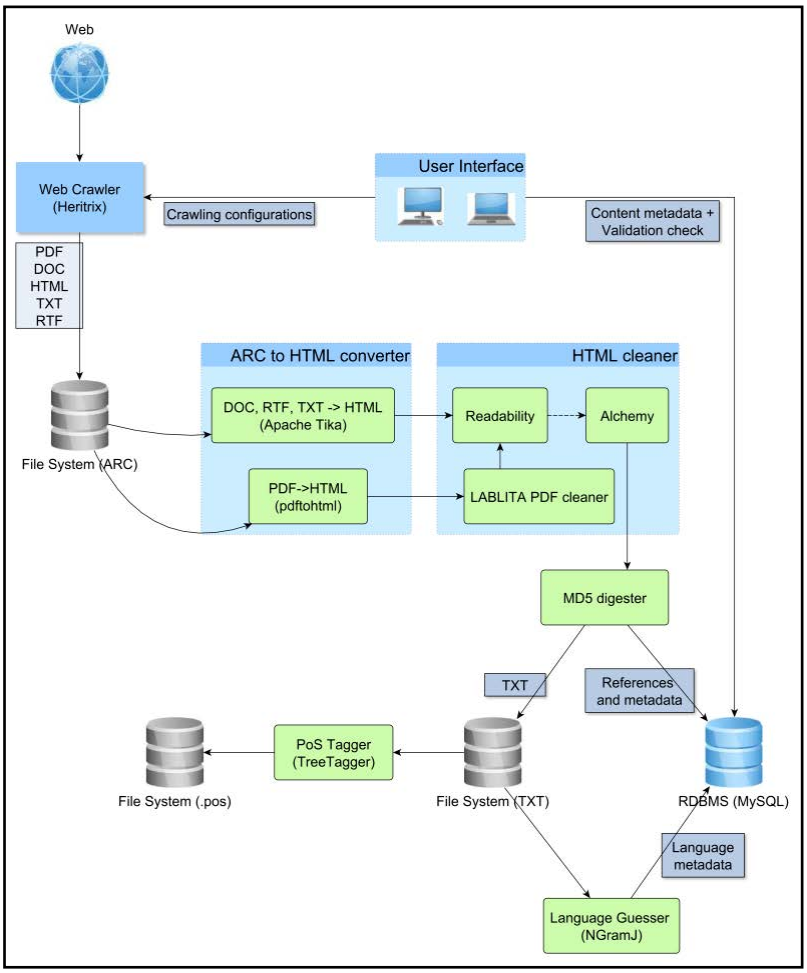


Figure 3: RIDIRE-CPI Architecture.

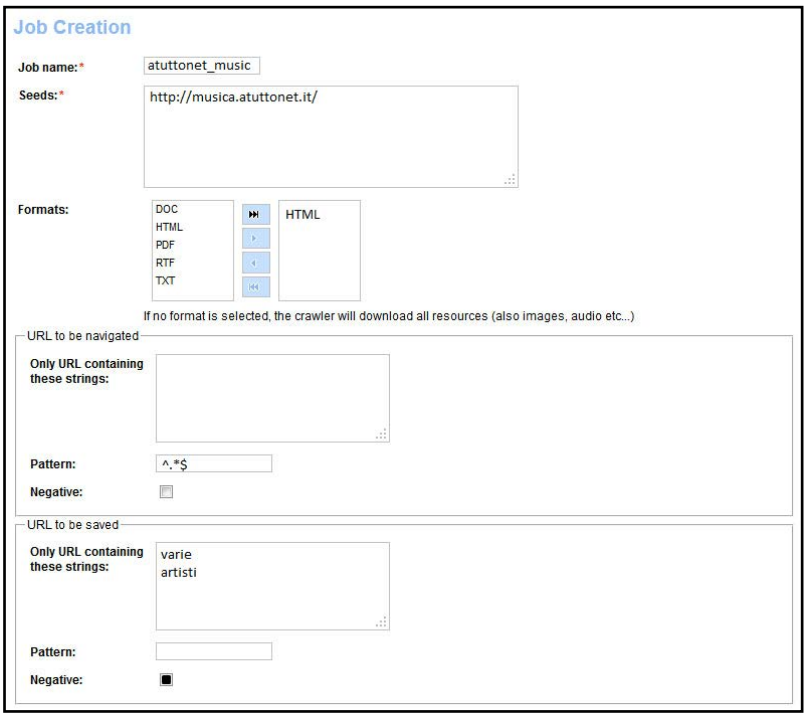


Figure 4: RIDIRE "Job Creation" page.

3.1 The Mapping Process

To be adequate for linguistic research, the crawled data needs to be processed by a procedure that includes text cleaning, duplicate removal, and PoS-tagging (Baroni et al. 2009). To this end, RIDIRE-CPI uses an automatic processing pipeline on the downloaded resources to extract the running text that will constitute the corpus itself. Web pages, as is well known, contain text that is not relevant for the constitution of a corpus e.g. advertising, navigation menus, disclaimers, credits, etc. (the so called "boilerplate"). Each terminated job is first converted into HTML, which involves several tools depending on the input format. After the conversion, the text cleaning is performed. The boilerplate is removed by means of two external tools freely available for research: Readability and Alchemy API. PDF files are more difficult to clean, so they are treated separately with a dedicated tool - PDF-Cleaner - that performs a deep filtering on the content.
Readability is the first option for the HTML cleaner, but if it won't yield results or outputs an error, the Alchemy API provides a second chance. The plain text documents output from the cleaning stage are then processed by a simple MD5 digester to get their signature, which acts as an anti-duplication system allowing the application to discard resources found with the same signature. The last phase of the mapping procedure is the part-of-speech tagging of the plain text resource. The PoS-tagging is performed by TreeTagger, which is run as an external executable by the main application. TreeTagger creates the PoS-tagged file in the correct file location directly.

3.2 Validation and Corpus Creation

RIDIRE-CPI integrates a validation interface dedicated to the evaluation of the crawled resources, which ensures that they belong to the specific domain they should represent. The validation procedure creates a random sample of the resources found and the user can check whether they are adequate with respect to the corpus design or content restrictions. A job can be considered "valid" if it contains non adequate resources under a given percentage (less than 10%, in principle). Since a manual revision is required for a high quality result, but checking the whole corpus is not an option due to its size, the validation process implemented in RIDIRE is a good trade-off between a clean corpus and a fast check. Figure 5 shows how the interface presents a random sampling of one crawled job, allowing direct access to a selection of pages whose adequacy in representing the given domain can be verified.

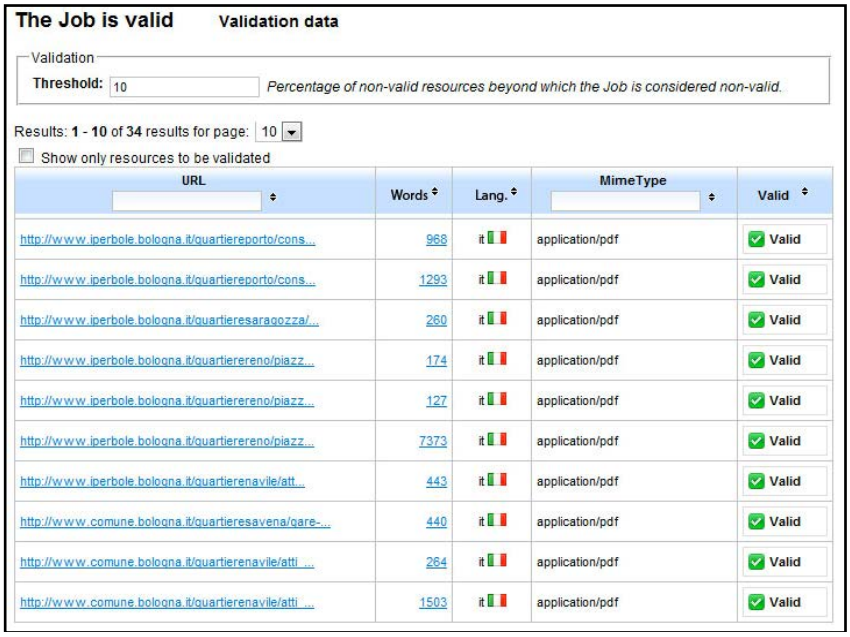


Figure 5: Validation sampling.

Through the content selection, metadata assignment, and validation procedures, the RIDIRE-CPI allows the gathering of linguistic data from the web with a supervised strategy that allows a high level of control. The frequency lists of the various domains provide direct evidence that the crawling performed within expectations. The nouns (i.e. the referred entities) that ranked highly identify each domain (Religion, Fashion and Cookery) quite well, and are shown in Table 2.

4 Methods for the Extraction of Linguistic Information from Corpora in L2 Acquisition and Lexicography

Various experiences in trying to use corpora for second language acquisition purposes clearly show that both learners and teachers are scared by the complexities of techniques involved in corpus linguistics and that the resultant data is difficult to appreciate (Kilgarriff 2009). Concordances provide a large amount of fragmented information that is difficult to read, especially for second language learners. Despite the fact that corpora contain information that is needed and that the tools are pretty powerful (Sinclair 2004; Conrad 2006), the way to use these tools is undefined and the information retrieved is difficult to interpret, with the overall process being felt as time consuming. The challenge for corpus linguistics in the field of second language acquisition is to provide a simple way to link the actual needs of learners to corpus data.

Religion		Fashion		Cooking	
Lemma	**Freq.**	**Lemma**	**Freq**	**Lemma**	**Freq**
vita	210,420	collezione	56,685	ricetta	135,498
uomo	169,995	moda	50,381	iscritto	104,610
amore	110,831	anno	49,369	località	93,692
fede	100,514	colore	32,777	acqua	82,492
mondo	98,913	abito	30,085	farina	81,695
pagina	95,462	mondo	28,816	volta	81,274
parola	92,532	donna	28,657	pasta	75,144
cuore	92,351	stile	26,815	zucchero	67,609
tempo	82,891	linea	26,026	minuto	66,579
giorno	76,190	pelle	20,962	impasto	65,074
figlio	70,231	capo	20,619	forno	61,672
persona	69,251	euro	19,199	olio	59,151
anno	69,054	modello	18,947	cucina	56,065
popolo	66,595	articolo	18,747	gr	55,079
modo	65,716	tempo	18,307	burro	52,101
preghiera	64,907	prodotto	17,365	uovo	49,057
cosa	57,020	marchio	16,968	cosa	48,276
santo	52,341	vita	16,388	tempo	47,712
fratello	51,370	accessorio	16,268	messaggio	47,453
famiglia	51,234	stilista	16,254	parte	46,829

Table 2: The 20 most frequent nouns, taken from 3 different domains.

The types of queries available in RIDIRE are inspired by those from the Sketch Engine and are available for both the general corpus and each sub-corpus:

- frequency lists
- concordances and patterns of words (ranked according to raw frequency)
- collocations (general and restricted to specific PoS)
- Sketches and Sketch Differences (between two words or domains) of collocates for the most relevant patterns of a word

The key strategy adopted in RIDIRE is to give a clear picture of the subset of problems that a learner can solve through corpora access, providing each problem area with a predetermined search path which leads to satisfactory results.

An extension of the concordances search function is the pattern search, where a user can view the concordances of a sequence of words (rather than a single one) specified by a form, lemma or PoS attribute; then, grouping the results together, he can see the more frequent usages of the sequence and

what the allowed syntactic structures are. In Figure 6 we searched the occurrences of the Italian verb *sperare* immediately followed by a preposition and we can see that there are five returned sequences (we excluded the rare occurrences): *sperare di* (68.37%), *sperare in* (13.88%), *sperare per* (4.24%), *sperare nel* (3.7%), *sperare nella* (3.26%). In this way a language learner can understand which prepositions may follow *sperare* and how they may be used by scrolling the occurrences list and looking at the different application contexts.

Figure 6: Pattern search grouped results.

RIDIRE is furthermore characterized by a set of sub-corpora representing Italian usage in different semantic and functional domains. The way in which a concept can be characterized in a given domain is partly a function of idiosyncratic usage conventions and corpus data can show this to the learner. In language this is reflected in particular by adjectives and adverbs, which show preferential meaning and associations and which vary across language usages. For instance, the variety of objects which are modified by the adjective *forte* ("strong") vary when the context of usage is Religion or Cookery. The learner should wonder whether or not this adjective, learned in general, has specific meaning in a domain when applied to its particulars. Here, RIDIRE exploits its corpus variation. Corpus queries based on collocations demonstrate the possible choices, highlighting the adjective's variation across domains.

The collocations in Figure 7 highlight the vastly different meanings conveyed by this adjective in each domain. In Religion, internal state is intensified (*fede*, "faith"; *tentazione*, "temptation"), while in Cookery flavours and smells are augmented. The meaning in one domain cannot automatically be extended to another.

Religion

Risultati: **1** - **20** di **4520**

Parola	Punteggio	Freq. nel corpus	Freq. collocata	Freq. collocata attesa
richiamo	0,0077	47888	245	4.2191
grido	0,0071	20125	129	1.7731
tentazione	0,0069	17810	116	1.5691
fede	0,0065	163796	586	14.431
supplica	0,0053	2467	49	0.2174
legame	0,0051	47612	161	4.1948
lacrima	0,0037	18831	64	1.6591
vento	0,0036	49803	117	4.3878
impulso	0,0036	19318	63	1.702
invito	0,0032	66481	134	5.8572

Cookery

Risultati: **1** - **20** di **2049**

Parola	Punteggio	Freq. nel corpus	Freq. collocata	Freq. collocata attesa
farina	0,169	8023	1254	0.2976
sapore	0,0138	49409	388	1.8329
odore	0,0098	16632	115	0.617
aroma	0,0068	9707	56	0.3601
abbraccio	0,0066	12520	64	0.4644
drago	0,0055	4226	30	0.1568
gr	0,0043	60518	149	2.245
botta	0,0043	7613	31	0.2824
emicrania	0,0043	644	16	0.0239
ricotta	0,0042	15348	46	0.5694

Figure 7: The first 10 collocations (lemmas) of the adjective *forte* in the Religion (left) and Cookery (right) domains.

Despite the versatility of the collocation extraction procedure and its implementation in linguistic applications, a basic knowledge of corpus querying techniques is required for correct usage. RIDIRE collocations across domains can also be extracted with the Sketches tool, which provides a more intuitive way to obtain linguistically relevant information. In other words, Sketches are more suitable for language learners that do not have high competence in corpus linguistics tools, as it provides them with an explicit language acquisition path.

A Sketch is a selection of relevant lemmas that co-occur with the key lemma in a specific syntactic pattern. The relevance of lemmas in each Sketch is determined by a lexical association measure (log-Dice in the RIDIRE implementation). Each Sketch corresponds to a precise grammatical relation[1]; for example, Figure 8 shows the *e_o* Sketch for the adjective *forte* in all domains i.e. the first ten adjectives that co-occur with *forte*, linked to it by a copulative (*e*, "and") or disjunctive (*o*, "or") conjunction:

e_o	39731	
deciso	586	8,15
coraggioso	381	7,73
chiaro	1122	7,4
sano	297	7,09
forte	1148	6,92
debole	313	6,87
potente	266	6,71
incisivo	159	6,61
competitivo	200	6,46
intenso	263	6,42
radicato	118	6,41

Figure 8: Example of a Sketch.

1 RIDIRE Sketches (including both the lexical queries and the visualization layout) are realized with the rules of SketchEngine, that is considered the reference web application for corpus linguistics studies.

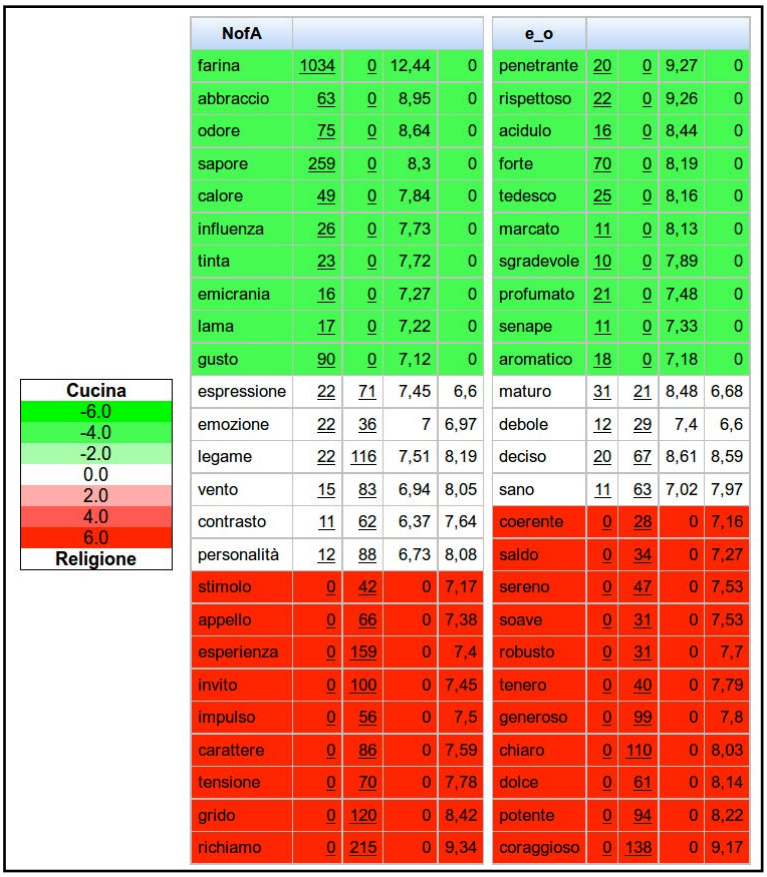
RIDIRE provides two extensions of the Sketch tool: Sketch Difference and Domain Sketch. The Sketch Difference tool shows the difference between the collocational behavior of two lemmas within the same syntactic pattern: we can see the words usable with the first lemma, with the second and with both of them.

In Figure 9 we see the difference between the Italian adjectives *forte* and *resistente* ("resistant") in the Fashion domain; specifically, we select two important Sketches: *e_o*, as in Figure 8, and *NofA*, which selects the nouns related to the adjective. From this example we can see that *forte* has a more varied usage in Fashion and is often related to the characterization of personality traits, while *resistente* is more specific and used for the technical specifications of clothing and accessories.

forte
-6.0
-4.0
-2.0
0.0
2.0
4.0
6.0
resistente

NofA					e_o				
impatto	365	0	10,52	0	deciso	159	0	10,38	0
personalità	291	0	10,02	0	chiaro	59	0	8,31	0
legame	137	0	9,21	0	sicuro	38	0	8,14	0
tinta	169	0	9,18	0	indipendente	23	0	8,05	0
pezzo	254	0	9,15	0	determinato	17	0	7,93	0
carattere	122	0	8,87	0	sensuale	38	0	7,91	0
crescita	174	0	8,85	0	riconoscibile	16	0	7,7	0
identità	108	0	8,85	0	grintoso	19	0	7,67	0
appeal	73	0	8,36	0	sano	18	0	7,65	0
espansione	79	0	8,35	0	coraggioso	12	0	7,46	0
richiamo	71	0	8,26	0	emerso	10	0	7,44	0
presenza	99	0	8,24	0	vivace	19	0	7,37	0
colore	374	0	8,19	0	pratico	0	13	0	7,29
segnale	63	0	8,16	0	aerodinamico	0	3	0	7,31
emozione	70	0	8,12	0	protettivo	0	6	0	7,34
contrasto	94	0	8,09	0	funzionale	0	10	0	7,6
punto	152	0	7,96	0	duttile	0	4	0	7,76
messaggio	51	0	7,67	0	elastico	0	10	0	7,82
connotazione	39	0	7,61	0	leggero	0	56	0	7,94
impronta	40	0	7,57	0	comodo	0	22	0	7,94
contenuto	49	0	7,53	0	robusto	0	9	0	8,54
carica	38	0	7,43	0	capiente	0	13	0	8,63
polimero	0	2	0	7,73	impermeabile	0	24	0	8,96
guaina	0	3	0	8,14	flessibile	0	19	0	9,44
ultra	0	4	0	9,24	ultra	0	11	0	9,51

Figure 9: The Sketch Difference for the adjectives *forte* and *resistente* in the Fashion domain.



Cucina
-6.0
-4.0
-2.0
0.0
2.0
4.0
6.0
Religione

NofA					e_o				
farina	1034	0	12,44	0	penetrante	20	0	9,27	0
abbraccio	63	0	8,95	0	rispettoso	22	0	9,26	0
odore	75	0	8,64	0	acidulo	16	0	8,44	0
sapore	259	0	8,3	0	forte	70	0	8,19	0
calore	49	0	7,84	0	tedesco	25	0	8,16	0
influenza	26	0	7,73	0	marcato	11	0	8,13	0
tinta	23	0	7,72	0	sgradevole	10	0	7,89	0
emicrania	16	0	7,27	0	profumato	21	0	7,48	0
lama	17	0	7,22	0	senape	11	0	7,33	0
gusto	90	0	7,12	0	aromatico	18	0	7,18	0
espressione	22	71	7,45	6,6	maturo	31	21	8,48	6,68
emozione	22	36	7	6,97	debole	12	29	7,4	6,6
legame	22	116	7,51	8,19	deciso	20	67	8,61	8,59
vento	15	83	6,94	8,05	sano	11	63	7,02	7,97
contrasto	11	62	6,37	7,64	coerente	0	28	0	7,16
personalità	12	88	6,73	8,08	saldo	0	34	0	7,27
stimolo	0	42	0	7,17	sereno	0	47	0	7,53
appello	0	66	0	7,38	soave	0	31	0	7,53
esperienza	0	159	0	7,4	robusto	0	31	0	7,7
invito	0	100	0	7,45	tenero	0	40	0	7,79
impulso	0	56	0	7,5	generoso	0	99	0	7,8
carattere	0	86	0	7,59	chiaro	0	110	0	8,03
tensione	0	70	0	7,78	dolce	0	61	0	8,14
grido	0	120	0	8,42	potente	0	94	0	8,22
richiamo	0	215	0	9,34	coraggioso	0	138	0	9,17

Figure 10: Domain Sketches (Cooking *vs.* Religion) for the adjective *forte*.

As the Sketch Difference function displays the contrast between the lexical associations of two lemmas in one corpus domain, the Domain Sketch tool shows the variation of a single lemma between two different domains.

In the Figure 10 we used the Domain Sketch tool to search the differences in usage for the lemma *forte* in two domains: Cooking and Religion. The general difference between these domains (*forte* is applied to flavours and smells in the Cooking domain and to feelings in the Religion domain) has already been demonstrated with the collocation search (Figure 7); however the result here is more fine grained, as it is divided into sketches, giving a more comprehensive overview of the lemma usage.

4.1 Lexicographic applications

Corpora have been widely used as data source in lexicography (Kilgarriff 2013). As a matter of fact, each of the researches presented in the previous section provide very relevant information for the lexical description of a word. Moreover, large corpora can be used as test-beds in order to decide what words and meanings should be inserted in a dictionary.

One of the main application field of corpora in lexicography is the detection of neologism by means of automatic or semi-automatic comparative analysis between an older word lists, taken from a dictionary or from a previous reference corpus, and the newer one, derived from an up-to-date corpus

(O'Donovan & O'Neill 2008). In this respect, web corpora are particularly interesting, since the web can be nowadays considered as the main access to written language, both in comprehension and in production, for a large part of the population.

The dimension and the structure of the RIDIRE corpus make it particularly attractive for lexicographic purposes. For instance, its data have been explored by Carla Marello for the study of Latin loanwords in Italian. The results showed that, in this respect, the corpus is richer than the modern dictionaries: all the Latinisms that are frequent in Italian monolingual dictionaries are frequent also in the corpus, but the corpus contains also various frequent Latinisms that are not reported in the dictionaries (but they probably should be).

The availability of very large corpora gave also a new perspective in the studies of collocations. Starting from these data. for example, it becomes possible to determine the input to which the learners are exposed while reading, and to select the collocations that should be considered during the compilation of monolingual and learner's dictionaries (Marello 2013). The use of sketches, that are a sort of quick synopsis of the grammatical and collocational behavior of a word, makes available a wide range of usage pattern that should be considered during the dictionary creation process.

Moreover, Sketches are useful not only for the detection of collocations, but also to give a quick picture of the distinct meanings of a word, since different meanings often select different collocates (Kilgarriff & Rundell 2002). It has to be noticed that the significance of this "extraction procedure" grows proportionally to the corpus dimension. If detecting meanings and collocations from very large corpora by means of concordance scanning could be very hard and time consuming, for the automatic collocation extraction procedures the bigger is the corpus, the better are the sketches (both in quantitative and in qualitative terms). Finally, the Sketch Differences tool is specifically interesting for comparing a word with its (near) synonyms and antonyms, in a pure lexicographic perspective.

5 Conclusions

Large scale corpora representing a language's domain of usage offer a unique source of data to both learners and lexicographers in accessing information about how the language is actually used. The computational tools now available, including those for web based infrastructures, allow the selection of the relevant information in a simple manner, overcoming significant difficulties encountered by corpus linguistics in meeting second language acquisition needs. Learners, teachers, and lexicographers, however, must be aware of the information required for a proper language acquisition that are up to usage conventions. On the basis of this understanding, corpus querying can be used to solve specific problems and be accepted as a modern method for use in the language acquisition process and in the dictionary creation.

6 References

Alchemy API. Accessed at: http://www.alchemyapi.com/ [06/04/2014].

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

Conrad, S. (2006). Challenges for English Corpus Linguistics in Second Language Acquisition Research. In Y. Kawaguchi, S. Zaima, T. Tackagaki, Y. Tsuruga, M. Usami (eds) *Lingusitcs Informatics and Spoken Language Corpora*. Amsterdam/Philadelphia: John Benjamins.

CQPweb. Accessed at: http://cwb.sourceforge.net/cqpweb.php [06/04/2014].

Heritrix. Accessed at: http://crawler.archive.org/ [06/04/2014].

Kilgarriff, A. (2009). Corpora in the classroom without scaring the students. In *Proceedings of 18th Internat. Symposium on English Teaching, Taipei.* Accessed at: http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc [06/04/2014].

Kilgarriff, A. (2013). Using corpora as data sources for dictionaries. In H. Jackson (ed.), *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 77-96.

Kilgarriff, A., Greffenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. In *Computational Linguistics*, 29(3), pp. 1-15.

Kilgarriff, A., Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications: A Case Study. In A. Braasch, C. Povlsen (eds), *Proceeding of the Tenth Euralex Conference, Copenhagen, 13-17 August 2002*. Copenhagen: University of Copenhagen, pp. 807-818.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds) *Proceeding of the Eleventh Euralex Conference, Lorient (France), 6-10 July 2004*. Lorient: Université de Bretagne-Sud, pp. 105-116.

Marello, C. (2013). Sembra che e subordinate soggettive. Primi sondaggi in italiano L2 scritto. In F. Geymonat (ed.) *Linguistica applicata con stile. In traccia di Bice Mortara Garavelli*. Alessandria: Edizioni dell'Orso, pp. 79-94.

Moneglia, M., Paladini, S. (2010). Le risorse di rete dell'italiano. Presentazione del progetto "RIDIRE.it". In E. Cresti, I. Korzen (eds) *Language, Cognition and Identity*. Firenze: Firenze University Press, pp. 111-128.

O'Donovan, R., O'Neill, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In E. Bernal, J. DeCesaris (eds) *Proceeding of the Thirteenth Euralex Conference, Barcelona, 15-19 July 2008*. Barcelona: Universitat Pompeu Fabra, pp. 571-579.

Panunzi, A., Fabbri, M., Moneglia, M., Gregori, L., Paladini, S. (2012). RIDIRE-CPI: an Open Source Crawling and Processing Infrastructure for Supervised Web-Corpora Building. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds) *Proceedings of Eighth Language Resources and Evaluation Conference (LREC 2012), Istanbul, 23-25 May 2012*. Paris: ELRA, pp. 2274-2279.

Readability. Accessed at: http://www.readability.com/ [06/04/2014].

RIDIRE Corpus Online. Accessed at: http://www.ridire.it [06/04/2014].

RIDIRE-CPI. https://github.com/lablita/ridire-cpi [06/04/2014].

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni, S. Bernardini (eds), *Wacky! Working papers on the Web as Corpus*. Bologna: Gedit, pp. 63-98.

Sinclair, J. (ed.) 2004. *How to use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Sketch Engine. Accessed at: http://www.sketchengine.co.uk/ [06/04/2014].

TreeTagger. Accessed at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [06/04/2014].

WaCky. Accessed at: http://wacky.sslmit.unibo.it/doku.php [06/04/2014].

Acknowledgments

The RIDIRE Project is funded by MIUR – FIRB 2007 and is promoted and maintained by SILFI (Società Internazionale di Linguistica e Filologia Italiana). The web application RIDIRE-CPI was developed by LABLITA and the corpus creation involved six Italian university departments: University of Florence (Dip. Italianistica and Dip. Sistemi e Informatica), University of Turin (Dip. Scienze Letterarie e Filologiche), University of Siena (Dip. Studi Aziendali e Sociali), University of Rome – Roma 3 (Dip. Italianistica), University of Naples – Federico II (Dip. Filologia Moderna).

Empirical Approaches to German Paronyms

Petra Storjohann, Ulrich Schnörch
Institut für Deutsche Sprache Mannheim
storjohann@ids-mannheim.de, schnoerch@ids-mannheim.de

Abstract

German lexical items with similar or related morphological roots and similar meaning potential are easily confused by native speakers and language learners. These include so-called paronyms such as *effektiv/effizient*, *sensitive/sensibel, formell/formal/förmlich*. Although these are generally not regarded as synonyms, empirical studies suggest that in some cases items of a paronym set have undergone meaning change and developed synonymous notions. In other cases, they remain similar in meaning, but show subtle differences in definition and restrictions of usage. Whereas the treatment of synonyms has received attention from corpus-linguists (cf. Partington 1998; Taylor 2003), the subject of paronyms has not been revisited with empirical, data-driven methods neither in terms of semantic theory nor in terms of practical lexicography. As a consequence, we also need to search for suitable corpus methods for detailed semantic investigation. Lexicographically, some German paronyms have been documented in printed dictionaries (e.g. Müller 1973; Pollmann & Wolk 2010). However, there is no corpus-assisted reference guide describing paronyms empirically and enabling readers to find the correct contemporary usage. Therefore, solutions to some lexicographic challenges are required.
Keywords: paronyms; synonyms; easily confused words; collocation profile

1 Introduction

This paper presents a new lexicographic project studying easily confusable words in language use and employing a data-driven approach to the investigation of German paronyms. Although there is a large spectrum of definitions, paronyms are generally referred to as lexical items with both related or similar morphological roots as well as slight morphological difference such as suffixes. But they are not only linked to another by similarity of form and/or sound but also have a similar semantic potential and are hence commonly confused for one another. Alternatively, it is one of the items of a paronymic pair that is commonly misused both by native speakers and learners, respectively. Examples of such items include for example *effektiv/effizient (effective/efficient)*, *sensitive/sensible (sensible/sensitive)*, *Method/Methodik/Methodologie (method/methodology)*, *formell/formal/förmlich (formal)*. They have generally not been regarded as synonyms (cf. Làzàrescu 1995, 1999). However, first empirical studies suggest that in a number of cases items of a paronym set have undergone meaning change and developed synonymous notions (e.g. Storjohann 2013). It is therefore argued here that a sharp distinction between paro-

nymy and synonymy is not always justified. In other cases, they remain similar in meaning, but can show restrictions of usage and subtle differences in definition.

Whereas the treatment of synonyms has received attention from corpus-linguists (cf. Partington 1998, Taylor 2003), the subject of paronymy has not been revisited with empirical, data-driven methods neither in terms of semantic theory nor in terms of practical lexicography. Lexicographically, some German paronyms have been documented in printed dictionaries (Müller 1973; Pollmann & Wolk 2010), although not systematically. However, there is no corpus-guided reference guide describing paronym sets empirically enabling readers to find the correct usage of such lexical items. So overall, paronyms should to be addressed from new perspectives. Firstly, the phenomenon has not been accounted for comprehensively in linguistic theory. Secondly, from a corpus linguistic view, we need to search for suitable corpus methods for detailed semantic investigation. Vachková & Belica (2009) propose the comparison of collocation profiles. They suggest a data-driven method to analyse lexical usage of near-synonyms with self-organising feature maps (SOMs). In this paper, we argue that this might prove a suitable method for the treatment of paronyms too, as it provides insight into both, semantic overlap and differences and it provides instant access to contrastive patterns by examining the concrete collocational behaviour of two items under investigation. Finally, solutions to some lexicographic challenges are required.

2 Linguistic Treatment of Paronyms

As Hausmann (1990) points out the subject of paronymy has mainly been approached linguistically from typological, language contrastive perspectives, particularly in the area of translation studies. These focus on this lexical relation exclusively from a language learner's view. Depending on different parameters, paronyms have been defined differently, covering items such as false friends, homophones, homographs or even cognates (cf. Bußmann 2002). In this paper, paronymy is broadly understood as a lexical relation between two or more items within one language, which are semantically related, have a similar or identical root and which are similar in form and sound but show a slight morphological difference (see morphemic paronyms/paronyms proper in Bolshakov & Gelbukh 2003: 199). Some of the most stringent terminology and a classification model have been proposed by Làzàrescu (1995, 1999) treating paronyms exclusively from a L2-learner's perspective.

Generally, Làzàrescu (1995) distinguishes between phonetically- and orthographically-based lexical confusion (e.g. *Föhn/Fön*), lexicologically-related terms (e.g. *Schifffahrt/Schiffsfahrt, Kindbett/Kinderbett)* and grammatically-based items such as *wohnen/bewohnen* or *dort/dortig*. Another class of paronyms exhibits specific semantic-stylistic features, e.g. as between *Vatermord/Vatermörder*, *Etikett/Etikette*). All these share their potential to be commonly confused in context and then to cause confusion in text reception or production due to similar phonemic representation as well as similarities of form and meaning. Most of these examples are far less likely to be mixed-up by native speakers. The importan-

ce of paronyms is based on the assumption that these items play a vital role for users in the process of second language acquisition and foreign language communication in order to avoid misunderstandings (Làzàrescu 1999). Confusing paronyms is sometimes regarded a violation of semantic correctness. Prescriptive analysts favour semantic correction and the avoidance of such mishaps and argue that paronyms "are important for poorly educated native speakers and for foreigners" (Bolshakov & Gelbukh 2003: 199). Indeed, the alleged misuse of morphologically and semantically similar words also cause linguistic uncertainties for native speakers, as numerous language-related Internet blogs show.

Language learners and native speakers too share their concern of correct language use with the language community as demonstrated by threads and their opening questions such as "Was ist der genaue Unterschied zwischen *effektiv* und *effizient*?" (What is the exact difference between *effektiv* and *effizient*), (see for instance http://www.gutefrage.net/frage/was-ist-der-genaue-unterschied-zwischen-effektiv-und-effizient). The answers of the blog community are impressively diverse.

So far, there is no semantic account encompassing different perspectives on the phenomenon and no satisfactory lexical focus on this relation that goes beyond Làzàrescu's categorisation system and that comprises semantic, diachronic and cognitive aspects. Effectively, there is a large interest in easily confused words from both learners and native users alike, but we lack an empirical treatment and full theoretical account of paronymy in general. So far, the question of what constitutes a relation of paronymy has not been satisfactorily been answered answered as besides lexical features it also involves cognitive aspects. Furthermore, we have no suitable, user-friendly, appropriate dictionaries documenting paronym behavior (see section 4). Hence, there are no widely tested methods that have proved suitable for semantic analysis of such words. To be able to derive conclusion and to develop hypotheses, it is suggested to work with corpus-driven procedures to examine paronyms closer. With the help of corpora and innovative tools it is possible revisit paronymy and also to open up new research questions. Since it is possible to analyse language use synchronically as well as diachronically and observe gradual meaning change even over a short period of time it is possible to detect slight semantic shifts or nuances and also to determine the degree of semantic overlap between similar lexical items both quantitatively and qualitatively. It is argued (see also Storjohann 2013) that in some cases there is no semantic violation when paronyms are being confused. The meanings of typically confused words are more freely exposed to semantic negotiation. Following a descriptive empirical view, the semantics of some paronymic lexical items have adopted new semantic aspects and undergone meaning changes that are observable as regular patterns in a corpus and not as single misused occurrences. Overall, corpus-driven research on paronymy demands a more differentiated look at the phenomenon than has previously been offered. Empirically-based investigations of paronyms can also provide valuable insights into cognitive aspects and the exact circumstances under which two items are being confused as well as possible principles of language change of conceptually associated terms.

3 Corpus-Linguistic Approaches to Paronyms

Currently, researchers face a range of techniques that can be incorporated into the analysis of texts, such as eliciting co-occurrences, extracting keyword lists, investigating concordance and analysing dispersion and frequency of words or patterns. Methodologically, it is advantageous to use corpus tools that are able to provide good access to patterns and structures of lexical use by exploring co-occurrences. Exploring collocational patterns and other syntagmatic patterns has become an established procedure in order to describe the contextual behaviour of a word in empirical lexical semantics and in lexicography. It has also become an established tradition within corpus-linguistics to lead researchers more towards difference-driven analysis (cf. Taylor 2013). However, the examination of paronym sets necessarily incorporates contrastive meaning analyses including the study of similarities too. Therefore, suitable methods should be capable of measuring semantic similarity or distance by contrasting collocation profiles pairwise to systematically uncover overlap and differences in terms of contextual behaviour.

Storjohann (2013) conducted a contrastive analysis of paronymic items on two sets of data. A semantic study of *effizient/effektiv* has been carried out on the basis of a large newspaper corpus, mainly by investigating their collocations. The data used in the analysis consists of 2.7 billion words (cf. Storjohann 2005) and it has been the basis of the lexicographic project *elexiko* (www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html) at the Instiut für Deutsche Sprache. In addition, self-organising feature maps (SOMs) (cf. Kohonen 1990; Keibel & Belica 2007) have been employed which offer visual representation of topographic profiles of the involved lexical items and which complement the collocation analysis. This procedure is implemented into a corpus-linguistic research and development workbench called CCDB (Keibel & Belica 2007). The CCDB is a database containing numerous static lexical co-occurrence profiles. It has been used to extract topographic profiles to break down unstructured collocation patterns and hence complex semantic properties (see Figure 1 and 2). On the basis of collocation profiles, semantic structures are analysed, clustered, visualized in a two-dimensional lattice reflecting different degrees of similarity between various words. As emphasized before, the comparison of paronyms implies the analysis of difference as much as the analysis of overlap. These are systematically being identified between items with overlapping collocation profiles (cf. Vachková & Belica 2009). Self-organising features maps cluster all those items such that proximity of the grid reflects semantic similarity between their semantic profiles. The more similar the colours of two neighbouring groups, the more similar are their collocation profiles although a strict separation is not suggested, as SOMs imply a continuum of semantic shades. The more their colours differ the more semantic differences can be found with regard to their uses (see Figure 1 and 2).

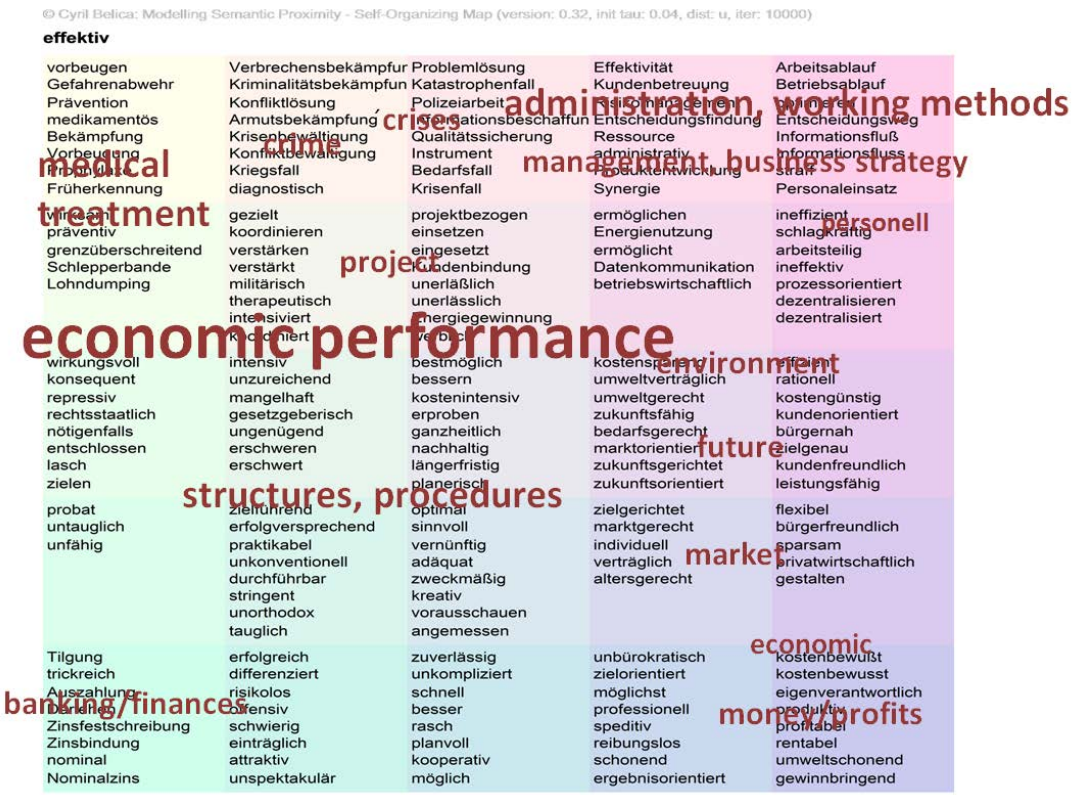


Figure 3: Topographic profiles of German *effektiv*.

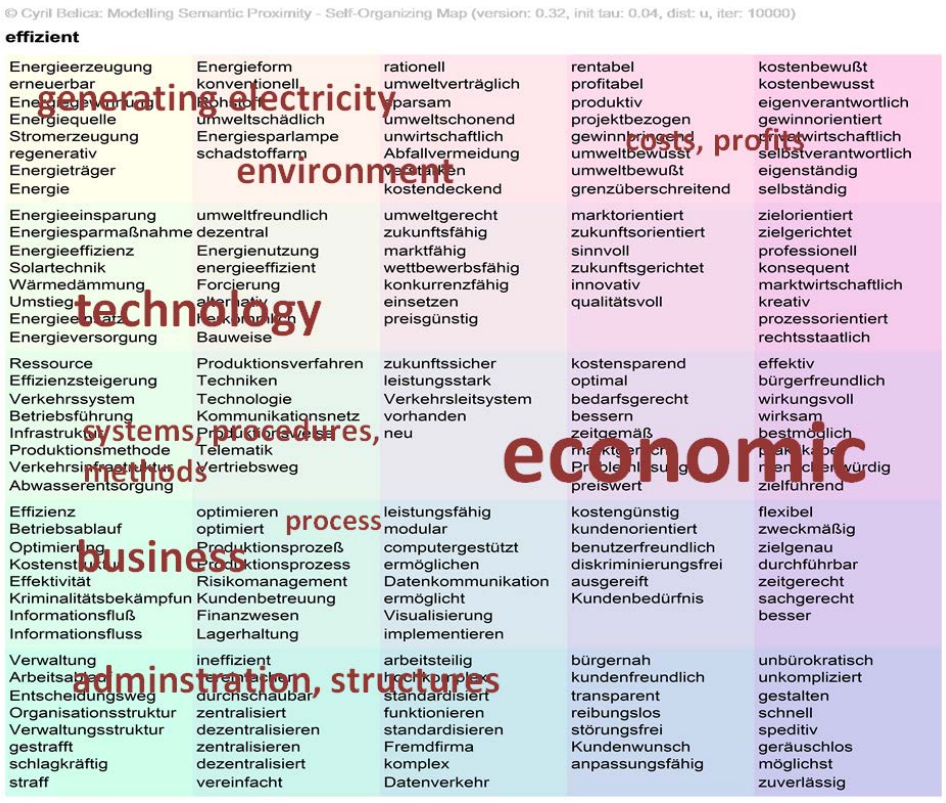


Figure 4: Topographic profiles of German *effizient*.

It is argued here, that the interpretation of such topographic feature maps could be the main entry point into data analysis of paronym behavior. It is a useful device for directing researchers to salient thematic domains associated with the individual terms. Through the process of abstraction a mental associations can be created by looking at individual squares and by moving from one square to the next (see Figure 1 and 2). As Vachková & Belica suggest:

Moving your focus forth and back, try to visualize the boundary where the initial sign eventually faded out, and where a notion of a new supersign entered your mind. Repeat for all corners and all directions. Try to assign each SOM square to at least one SOM supersign. (2009: 228-229)

Hence, a more abstract "supersigns" or superordinate concepts can be derived to categorise key semantic fields, clusters or domains. One major advantage of this procedure is the quick detection of thematic topics or domains in which the lexical items predominantly occur. For both lexical items, these thematic supersigns have been assigned and marked typographically. As a result, the interpretation of the findings as derived from supersigns or general concepts can then be directly compared, as summarised for example in Table 1.

effizient	**effektiv**
systems/procedures/structures administration costs/projects/economy	business/methods/work/management/personnel economic structures/performance
environment/generating electricity	fighting crime crisis management
technology	medical treatment/therapy
	environment/future
	banking & finance

Table 3: Semantic contexts/domains.

The domains in Table 1 are arranged according to their dominance in the underlying corpus. Semantic overlap between *effizient* and *effektiv* can be found in thematic domains where supersigns refer to the notion of business and administration and where both terms characterise methods, structures, procedures and issues of management. These are the discourse areas where ample evidence of synonymous contexts is provided in written German corpora (see examples 1 -3, for more examples see Storjohann 2013).

1. Zwar besteht auch in Hamburg die Vorschrift, dass ein Toter höchstens 36 Stunden in privaten Räumlichkeiten liegen darf und nach 14 Tagen begraben sein muss. „Aber Fakt ist, dass es eine weitere gesetzliche Reglementierung gar nicht gibt", sagt Hillermann. Der sture Ablauf, der fast immer eingehalten wird, sei vielmehr das Ergebnis *effektiver* Arbeitsteilung von Spitälern, Bestattern und den Friedhofsbetreibern. Fast alle größeren Bestattungsunternehmen pflegen in

Deutschland eine *effiziente* Arbeitsteilung. (Die Zeit, 15.04.2004; Wie man in Deutschland begraben wird)

2. Im Jahr 2000 haben sich drei Fünftel der heimischen Gastbetriebe vom Dehoga abgespalten. [...] Der Hoga Rheinland, also der Hotel- und Gaststättenverband im nördlichen Rheinland-Pfalz, ist seinerzeit ausgetreten, weil er die Landes- und Bundesverbandsarbeit für nicht *effektiv* genug gehalten hat und die nicht üppig vorhandenen finanziellen Mittel lieber in regionale Arbeit und Projekte stecken wollte. Die Nordlichter warfen dem Dehoga zu wenig *effiziente* Arbeit vor. Wie sehen Sie das? (Rhein-Zeitung, 05.11.2004; Gastgeber wieder unter einem Dach)

Differences can be found for *effizient* being used as an attribute to characterise equipment, instruments and technological developments as well as being used to specify types and the use of energy resources. *Effektiv*, on the other hand, is more likely associated with the contextual domain of crime and crisis management. It is also attributed to characterise medical treatment and means of saving the climate. Furthermore, it is used within the context of banking and financing. Although, total semantic exclusion is not suggested, within these contexts both terms are far less likely being synonymously. As a result, the self-organising feature maps help us to structure the "unordered" semantics of a lexical item in use. It provides us with necessary details such as semantic dominance and contextual preferences in terms of referential domains and discursive foci.

In a further step, self-organisation maps can be used to contrast patterns of usage between two lexical items such as *effektiv* and *effizient* by comparing them with all those items with which they share parts of their collocation profiles.[1] This means that it is capable of measuring semantic similarity or distance by contrasting typical contextual behavior pairwise. The procedure is referred to as CNS-model (Contrasting Near-Synonyms, cf. Belica 2001 ff) and it is implemented into the workbench CCDB too. This procedure allows linguists to compare and contrast two words visually according to salient collocational contexts. *Effektiv* and *effizient* are being contrasted with each other and with those items with which they share parts of their co-occurrence profiles. The feature maps arrange specific usage aspects that the items in question share and those they do not have in common (see Figure 3).

1 Compare for example corpus software which facilitate the comparison of collocates visually, e.g. SketchEngine's Word Sketch Difference (https://www.sketchengine.co.uk/)

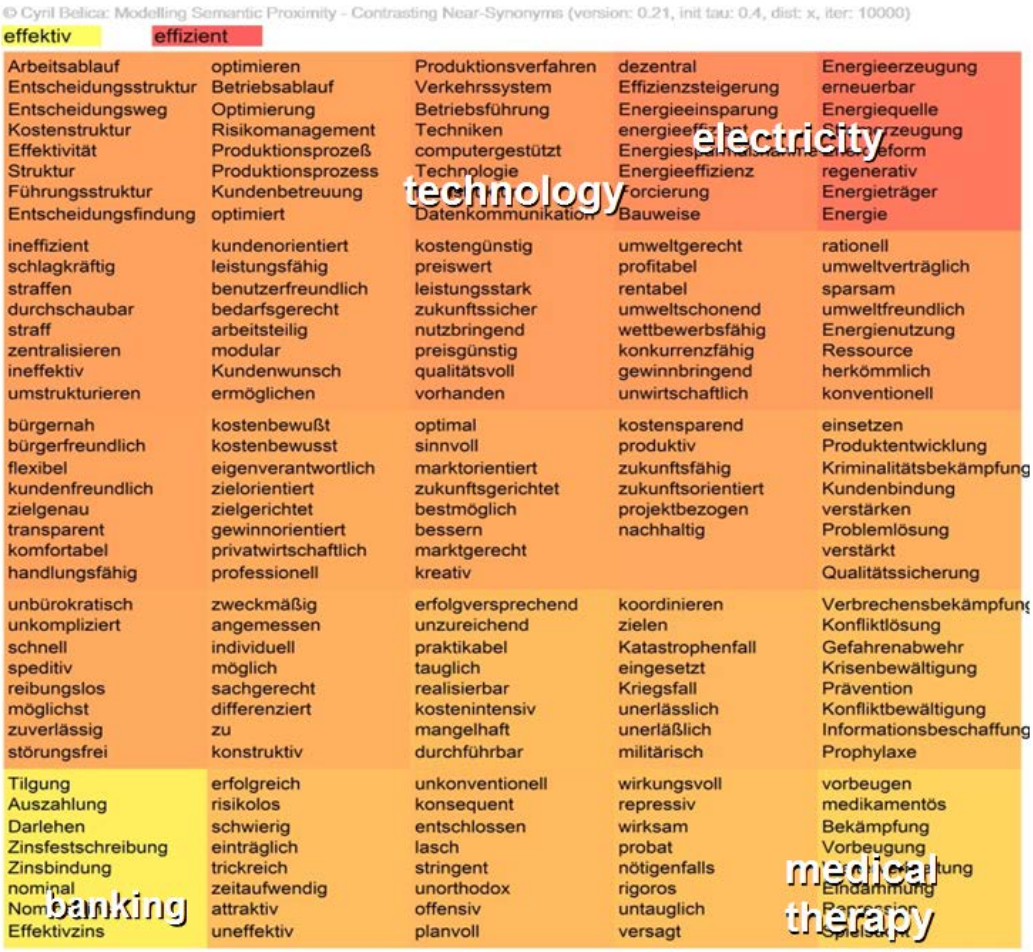


Figure 3: Contrasting German *effektiv* and *effizient* with SOM.

As indicated, the referential domains of 'banking and financing' (bottom left) and 'medical treatment' (bottom right) are preferably assigned to *effektiv*. Ways of generating electricity/energy and defining or specifying technology, however, are more likely a subject where the term *effizient* is preferably attested. In Figure 3, the subjective interpretation of the domains has been added to the map. Large parts of this feature map are shared by both terms. For the purpose being, these have not been further analysed also because a number of referential domains cannot always be clearly determined. Nevertheless, this feature map helps to validate the findings of conceptual similarities and differences as summarised in Table 1.

Generally, feature maps cannot serve as detailed lexicographic documentations to help users to be aware of "appropriate" or "false" usage. With respect to a paronym dictionary, more information as to concrete contextual usage is necessary. In the next stage, these domains or themes could, for example, be exemplified through statistically significant collocates extracted from an underlying corpus and assigned to corresponding discourses in order to illustrate specific preferences and restrictions (see Table 2). This process reveals for example, that one can modify procedures as processes (*Abläufe*, *Arbeitsabläufe*, *Arbeitsweise*, *Betriebsabläufe*), solutions (*Lösung*), structures (*Strukturen*), systems (*System*) and menas/instruments (*Maßnahmen*) both as *effizient* and *effektiv* without implying much of a difference. It is within these contextual domains that similarity between the two terms is most evident

and examples of synonymous usage are being attested in the corpus (compare corpus examples 1 and 2). One does not typically characterise technology or the use of electricity/energy as *effektiv*. Here, the adjective *effizient* is being preferred. There are also efficient combustion engines (*Verbrennungsmotoren*) and efficient power stations (*Kraftwerke*). Other collocates on the topic of health and medical treatment include for example *therapy* (*Behandlungsmethoden*, *Therapie*) and *exercises for your back* (*Rückenübungen*) usually associated with *effective* but not with *efficient*. And terms referring to the banking and finance sector could be *interests*, *rates*, *return on capital* etc. These are modified in German by using *effektiv* (meaning real) exclusively, but never by *effizient*.

effizient		**effektiv**	
collocate	discourse domain	collocate	discourse domain
Abläufe, Betriebsabläufe, Arbeit, Arbeitsabläufe, Einsatz, Lösung, Strukturen, Verwaltungen, Maßnahmen, Arbeitsweise, Bewirtschaftung, Organisation, Wirtschaften, arbeiten, Aufgabenerfüllung, Bewirtschaftung, Methode, Abwicklung ...	systems procedures structures administration costs projects economy	arbeiten, gestalten, Maßnahme, Werbung, Methode, Maßnahme, Lösung, Strukturen, System, Arbeit, Verwaltung, Kommunikation, Kontrolle, Organisation, Arbeitsabläufe, Personaleinsatz, Zeitmanagement, organisieren, Controlling, Techniken, Strukturen ...	business methods work management personnel economic structures performance
Energie, Energieeinsatz, Energienutzung, Energieverwendung, Energieversorgung, Energieverbrauch, Heizen, Heizsysteme, Kohlekraftwerke, Stromnutzung, Stromerzeugung ...	environment generating electricity	bekämpfen, Mittel, Krisenmanagement, Strafverfolgung, schützen, Rechtsschutz, Selbstverteidigungstechniken, Verbrechensbekämpfung, Dopingbekämpfung, Polizeiarbeit, Überwachung ...	fighting crime crisis management
Motoren, Verbrennungsmotoren, Anlagen, Antriebstechnologie, Heizungen, Wärmepumpen, Kraftwerke, Geräte, Technologien, Wärmedämmung, ...	technology	Hochwasserschutz, Ressource, Klimaschutz, Wärmedämmung, Reduzierung	environment future
		Behandlungsmethoden, Ganzkörpertraining, Therapie, Rückenübungen, Prävention, Behandlung, Workout, Schmerztherapie, Beinmuskeltraining, Behandlung ...	medical treatment therapy
		Jahreszins, Rendite, Zins, Zinssatz, Nominalzins	banking & finance

Table 4: Collocates in corresponding discursive domains.

As suggested by Vachková & Belica (2009), this approach to collocational patterning might be applicable to the lexicographic investigation of synonyms. It is argued that

salient SOM features stimulate lexicographers' associative awareness and encourage guided mental imagery leading to valuable insights into both the word semantic structure and the process of discourse-based negotiation of lexical meaning (Vachková & Belica 2009: 239).

The notion of similarity has played a great role in lexicological areas, for instance in the corpus-linguistic investigation of sense relations by using collocational overlap to measure degrees of synonymy (cf. Partington 1998). Marková (2012), for example, puts forward examples of studies of German synonyms where she employed the CNS-model successfully. It is proposed here that consultations and interpretations of self-organisation feature maps might be a suitable approach to the analysis and semantic description of paronyms too where usage aspects that are shared and not shared can be uncovered. Feature maps can guide lexicographers to those contextual patterns where to look for further evidence for example through the analysis of collocations that can be attributed to specific thematic domains. Effectively, the chosen procedures result in a form of methodological triangulation comprising three different analytical stages: first, interpreting SOMs in order to associate domains, secondly, the CNS-Model to validate the previous interpretation and thirdly, collocation analysis to exemplify the given domains/topics contextually.

4 The Lexicographical Treatment

With regard to German, commonly confused words including some paronyms have been described in two printed reference books: Müller (1973) and Pollmann & Wolk (2010). Both are prescriptive documentations aiming at guiding users to the allegedly correct usage and describing a clear distinction between the items in question (see for example Figure 4).

effektiv ··· **effizient**

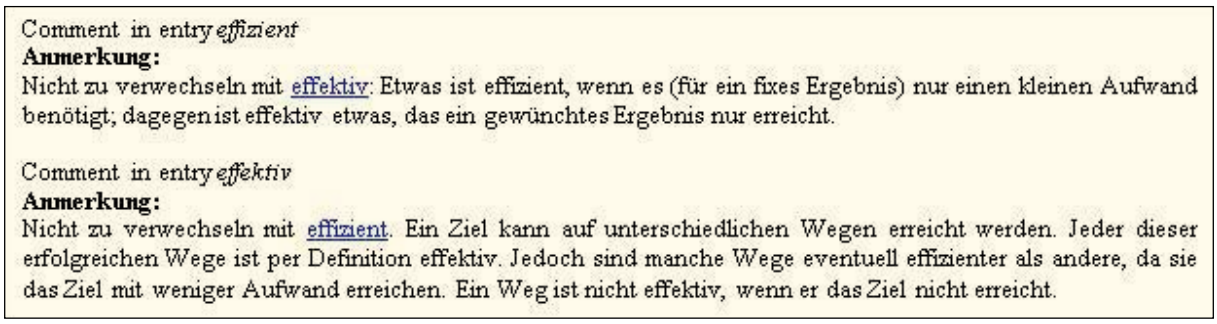
- Ist etwas gemessen an den Mitteln, die zur Erreichung eines bestimmten Ziels ausgerichtet sind verhältnismäßig, dann ist es **effektiv**.
- **Effizient** ist etwas, das besonders wirtschaftlich ist, also die Kosten mit dem Nutzen vergleicht. (Eselsbrücke: -zient / Cent).

Ein Insektenspray vertreibt oder tötet beispielsweise Insekten, somit ist es effektiv. Wesentlich effizienter ist jedoch eine Fliegenklatsche.

Figure 4: Dictionary entry *effektiv/effizient* in Pollmann & Wolk (2010).

The entries contain short meaning descriptions, occasional encyclopaedic comments and citations or examples. Moreover, some normative grammars and lexical studies concerning the didactics of normative language practice contain lists of some paronyms (e.g. Heringer 1989, 1995). Strictly normative language use is also propagated in wiktionary, a popular electronic resource which under an explicit

headline points out that confusion over the two words *effektiv* and *effizient* should be avoided (see Figure 5).


Comment in entry *effizient*
Anmerkung:
Nicht zu verwechseln mit effektiv: Etwas ist effizient, wenn es (für ein fixes Ergebnis) nur einen kleinen Aufwand benötigt; dagegen ist effektiv etwas, das ein gewünchtes Ergebnis nur erreicht.

Comment in entry *effektiv*
Anmerkung:
Nicht zu verwechseln mit effizient. Ein Ziel kann auf unterschiedlichen Wegen erreicht werden. Jeder dieser erfolgreichen Wege ist per Definition effektiv. Jedoch sind manche Wege eventuell effizienter als andere, da sie das Ziel mit weniger Aufwand erreichen. Ein Weg ist nicht effektiv, wenn er das Ziel nicht erreicht.

Figure 5: Comments regarding the misuse of *effektiv* and *effizient* in wiktionary.

From a lexicological point of view the remarks found there are questionable. However, they demonstrate that language users are aware of a potential conflict between *effektiv* and *effizient* and that misuse or confusion of this kind is a rather regular phenomenon that is to be avoided. Overall, in all reference guides findings are neither based on semantic examinations of current natural language in use nor on investigations of large data. Empirical corpus explorations open up the discrepancies to traditional descriptions. The usage restrictions that are documented in these reference books cannot be confirmed through corpus data. Entries lack collocational details referring to recurrent referential domains as for example illustrated in Figures 1 and 2. These provide essential information to users as to in which concrete contexts the corresponding adjectives might be more commonly found. As is the case for *effektiv/effizient*, strict usage lines cannot be sharply drawn which might have been expected intuitively. Conventional reference guides have so far focussed on the differences between commonly confused words. They entirely fail to explain existing similarities. In that respect the methodology is similar to most corpus-linguistic research. Corpus-assisted studies on semantically similar words have so far focussed on the differences between the individual items. As Pearce (2008: 21) points out there is a risk of "the privileging of differences over similarities. [...] the analyst is in danger of exaggerating the differences and overlooking similarities". However, corpus studies also allow for the description of similarities which, on the one hand, might offer a deeper understanding why two words are regularly being confused and, on the other hand, it might indicate ongoing linguistic change worth documenting.

4.1 Challenges

A new project at the Institut für Deutsche Sprache will reopen the chapter on paronymy as a lexical as well we cognitive phenomenon. It will account for paronymy from a corpus-linguistic perspective and it will test methods that will hopefully prove suitable for semantic analysis of such words. It is hoped that data-driven investigations of paronyms can provide valuable insights into principles of

language change in semantically related lexical items. This could enable us to integrate paronymy into a wider theoretical framework. Part of this project is the compilation of the first corpus-assisted paronym dictionary which aims at guiding users descriptively through current usage and contexts. As an electronic resource it will provide adequate pairwise documentations combined with user-friendly navigation and search structures. In the near future, it will be an integral part of the German dictionary portal OWID (see: http://www.owid.de/).

From a lexicographic point of view, a number of challenges are encountered when documenting usage-based findings in a paronyms dictionary where users might demand definite answers for doubtful language situations. One central problem regards the interpretation and documentation of language change and normative restrictions. This is particularly relevant for pairs that are recorded as semantically distinct lexical items in traditional reference works and that have assimilated semantically over time due to common, allegedly "false" use. In some cases, corpus analyses signal tendencies that paronyms might have possibly turned into synonyms. Therefore, one of the major challenges of a corpus-based paronym dictionary is the interpretation of ambiguous data, especially paronym usage with a similar proportion between contexts with clear semantic difference between the terms and contexts exhibiting synonymous use. The lexicographic interpretation of such data requires a certain sensibility, as a specific conflict is expected to be encountered with corpus data. One the one hand, false language use caused by confusing paronyms needs prescriptive correction. On the other hand, gradual language change caused by frequent misuse of a certain lexical item needs descriptive documentation of contemporary language use. Above all, it should be able to explain semantic overlap and to sensitise users for continuous language change. Although this has not been studied, it is assumed that the expectations of users of such a dictionary have rather prescriptive notions due to their habits and their handling with other existing dictionaries. Therefore, the most challenging objective of this descriptive, usage-based dictionary certainly is to offer a reference guide that shows similarities and difference between paronymic items contrastively, including corpus samples, explanations and comments without neglecting to inform users about aspects of semantic overlap, gradual semantic changes, contextual vagueness, possible substitutability and at the same time still answering their look-up questions satisfactorily.

4.2 Presentation

As an e-dictionary the new German paronym guide can go far beyond the depth of information found in the two existing printed dictionaries. It will also have to consider different options with respect to navigation, visualisation, cross-referencing, linking and searching in order to exploit the possibilities of the electronic medium profitably and in order to create a user-friendly instrument. Traditional dictionary entries contain explanations of the formula "*effizient* is something that ..." or alternatively "if something is *effizient* than ..." (see Figure 4). The answer as to what exactly this something is can be found in contexts and collocates. Although at this point, no finite solution of the details of presentati-

on can be given, it is inarguably information on co-occurring patterns that most attention will be drawn to. Users can expect direct access to collocational patterns of two or more easily confused words together at the same time with their interpreted thematic domains in which both are likely to occur together with their preferences or their restrictions. Corpus samples will illustrate the information given. The depth of information could be realised as optional user-customised views. Overall, two aspects need further exploring. Firstly, the possibilities of the electronic medium still need to be examined and exploited to create a reliable and usable environment enabling users to make correct choices. Secondly, research on the users' needs and behaviour provides us with valuable insights (cf. Tarp 2008; Müller-Spitzer 2014) and these should be profitably incorporated into modes of presentation.

5 Conclusion

In this paper, a multifaceted approach to the study of regularly confused words in German has been discussed, suggesting a method of investigation that implies both differences and similarities between paronymic items. The phenomenon of paronymy has not been accounted for empirically. It has neither been reconsidered in recent linguistic theories and models nor in lexicographic practice. This gap will be closed by a new project which accounts for paronymy from a corpus-linguistic perspective where methods of investigation will be tested to find suitable tools for semantic analysis. With the help of the example *effective/effizient*, we have set out one possible way to implement different software-driven resources facilitating the search for similarity and difference. Corpus-assisted investigations of easily confused words and their usage over recent decades can provide valuable insight into principles of semantic shift. It is argued here, that such analyses might enable semanticists to integrate the phenomenon into a wider theoretical framework on the one hand and into appropriate lexicographic descriptions on the other hand.

6 References

Belica, C. (1995). Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache.

Belica, C. (2001ff). *Kookkurrenzdatenbank CCDB – V3.3. Eine korpuslinguistische Denk- und Experimentierplattform*. Mannheim: Institut für Deutsche Sprache. Accessed at http://corpora.ids-mannheim.de/ccdb/ [27/3/2014].

Belica, C. (2006). Modellierung semantischer Nähe: Kontrastierung von nahen Synonymen. Korpusanalytische Methode. Mannheim: Institut für Deutsche Sprache.

Belica, C. (2011). Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In A. Abel, R. Zanin (eds.) *Korpora in Lehre und Forschung*. Freie Universität Bozen: Bozen University Press, pp. 155-178.

Bolshakov, I.A., Gelbukh, A. (2003). Paronyms for accellerated correction of semantic errors. In *International Journal Information Theories & Applications*, 10, pp. 198-204.
Bußman, H. (2002). *Lexikon der Sprachwissenschaft*. 3rd edition, Stuttgart: Kröner Verlag.
elexiko. Accessed at http://www.owid.de/wb/elexiko/start.html [27/3/2014].
elexiko-corpus. Accessed at http://www.owid.de/wb/elexiko/glossar/elexiko-Korpus.html [27/3/2014].
Hausmann, F. J. (1990). Das Wörterbuch der Homonyme, Homophone und Paronyme. In F. J. Hausmann, O. Reichmann, H.E. Wiegand (eds.) *Wörterbücher. Dictionaries. Dictionnaires. Vol. 2.* Berlin/New York: de Gruyter, pp. 1120-1125.
Heringer, H.J. (1989). Lesen – lehren –lernen. Eine rezeptive Grammatik des Deutschen. Studienausgabe. Tübingen: Niemeyer.
Heringer, H.J. (1995). Grammatik und Stil. Praktische Grammatik des Deutschen. Berlin: Cornelsen.
Keibel, H., Cyril, B. (2007). CCDB. A Corpus-Linguistic Research and Development Workbench. In *Proceedings of the 4th Corpus Linguistics Conference, CL 2007, 27-30 July 2007*. University of Birmingham, UK.
Kohonen, T. (1990). The Self-Organizing Map. New Concepts in Computer Science. In *Proceedings of the IEEE*, Vol. 78 (9), pp. 1464-1480.
Làzàrescu, I. (1995). Deutsche Paronyme. In *Grazer Linguistische Studien* 43, pp. 85-93.
Marková, V. (2012). Synonyme unter dem Mikroskop. Eine korpuslinguistische Studie. Tübingen: Gunter Narr.
Müller, W. (1973). *Leicht verwechselbare Wörter. Duden Taschenwörterbücher*, Vol. 17. Mannheim: Bibliographisches Institut.
Müller-Spitzer, C. (ed.) (2014). *Using Online Dictionaries*. Berlin/New York: de Gruyter. (Lexicographica: Series Maior 145).
OWID Online-Wortschatz-Informationssystem Deutsch. Accessed at: http://www.owid.de/ [27/3/2014].
Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
Pearce, M. (2008). Investigating the Collocational Behaviour of Man and Woman in the BNC using Sketch Engine. In *Corpora* 3(1), pp. 1-29.
Pollmann, Ch., Wolk, U. (2010). *Wörterbuch der verwechselten Wörter*. Stuttgart: Pons.
SketchEngine. Accessed at: https://www.sketchengine.co.uk/ [27/3/2014].
Storjohann, P. (2005). Das elexiko-Korpus. Aufbau und Zusammensetzung. In U. Haß (ed.) *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. (Schriften des Instituts für Deutsche Sprache 12). Berlin/New York: de Gruyter, pp. 55-70
Storjohann, P. (2013). Korpuslinguistische und lexikografische Ansätze zur Beschreibung deutscher Paronyme. In H. Scheuringer, D. Sava (eds.) *Im Dienste des Wortes. Lexikologische und lexikografische Streifzüge*. Festschrift für Ioan Làzàrescu. (Reihe: Forschung zur deutschen Sprache in Mittel-, Ost- und Südosteuropa Vol. 3). Passau: Stutz, pp. 401-418.
Vachková, M., Belica, C. (2009). Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography. In *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, 13(2), pp. 223-260.
Wiktionary. Das freie Wörterbuch. Accessed at: http://de.wiktionary.org [27/3/2014].
Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Niemeyer.
Taylor, Ch. (2013). Searching for Similarities Using Corpus-assisted Discourse Studies. In *Corpora*, 8, pp. 81-113.
Taylor, J. (2003). Near synonyms as coextensive categories: 'Tall' and 'high' revisited. In *Languages Science*, 25, pp. 263-284.

Pragmatic Meaning in Lexicographical Description: Semantic Prosody on the Go

Mojca Šorli
Trojina - Institute for Applied Slovene Studies
mojca.sorli@trojina.si

Abstract

The present paper focuses on ways in which the pragmatic (functional) meaning, known in corpus linguistics as semantic prosody, is treated in monolingual and also bilingual dictionaries. We have analysed a selection of lexicographical descriptions, as they are represented in the *Slovene Lexical Database* (SLD), comparing them to a number of English and Slovene lexical sources, and demonstrated how corpus-derived pragmatic meaning can become an integral part of dictionary definitions. This is particularly important for the treatment of phraseology and idiomatics, where greater involvement of pragmatics is demonstrated. A tentative typology was compiled for the purpose of this analysis in order to categorise lexical units according to their inner semantic-pragmatic relations, with a view to examining the implications for the monolingual dictionary treatment of individual items, as well as any potential strategies that could be applied, on the basis of the posited categories, to their translation. We have also pointed out the treatment of individual lexical units in the selected bilingual dictionaries.

Keywords: lexicographical description; lexical database; monolingual/bilingual dictionary; pragmatics; semantic prosody

1 Introduction

The present paper is based on some results obtained in the course of doctoral research into the ways in which the pragmatic (functional) meaning that arises from various contextual features, known in corpus linguistics as semantic prosody (Sinclair 1991, 1996; Louw 1993; Stubbs 1995, 2001; Partington 1998; Tognini-Bonelli 2001; Whitsitt 2005; Hunston 2007, etc.) can become an integral part of (monolingual and bilingual) dictionaries. We will attempt to demonstrate the value of the explicit description of pragmatic meaning, i.e., semantic prosody, as implemented in the Slovene Lexical Database (SLD),[1] while also presenting some conclusions based on the exploration of the possibilities of recording semantic prosody in a bilingual perspective. We posit various types of meaning that are codified in specific types of linguistic form or patterns of use. Of central concern to us is the meaning de-

1 Slovene Lexical Database (2008-2012): The project was co-financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia.

scribed as "peripheral" or "underspecified" (see Philip 2009). This can only be studied in context, as it is completely dependent on collocation and syntagmatic relations, and therefore cannot be attributed solely to a concrete word form. The basic pattern of language use is represented by collocation based on the distributional features of words, while a more abstract type of pattern is derived from "inter-collocational" generalisations, which include semantic prosody (Ellis et al. 2009: 89-90). In other words, we not only speak of lexicogrammatical patterns, but also of pragmatic patterns of language use. In the present paper, we adopt the view that the function of a dictionary should not be limited to presenting the "referential", "denotative", "cognitive", "semantic", "dictionary", etc., meaning, but should contain a comprehensive description of inherent semantic features of words, as well as the pragmatic circumstances of their use. A number of successful (English) language learner's dictionaries have been designed to take into account functional aspects of meaning. Although still lagging behind, bilingual dictionaries have also moved on from being mere "glossaries" expected to provide no more than "prototypical", "systemic" or "cognitive" equivalents to not only corpus-based but "corpus-like" language resources, in which the user can explore words in real use. With space restrictions no longer in place, electronic lexicography now has the means and the opportunity to devote more attention to the textual and pragmatic dimensions of meaning, such as the complexities of semantic prosodies, which, as research shows (Hunston 2002; Zethsen 2006; Zhang 2009, etc.), contemporary monolingual lexical databases and dictionaries still fail to convey, typically implying them in examples of (typical) usage. Moreover, in bilingual dictionaries semantic prosody is typically ignored altogether.

2 Forms of Encoded Pragmatic Meaning: Semantic Prosody

Semantic prosody is an integral part of an (extended) unit of meaning, identifiable only by examining its repeated occurrences in a large amount of (corpus) data. For example, at first glance "situation" seems perfectly neutral, but examining a large number of contexts of situation shows that it typically occurs as the node of units of meaning in contexts conveying negative events, facts or features that evoke negative associations and carry negative semantic prosody. In the 112-million BNC reference corpus there are 19,569 hits (174,4 per million). Collocations are very dispersed, with some of them being seemingly neutral, e.g., *present, given, similar*, but the wider context reveals negative circumstances of meaning. The most frequent collocations are: *the situation is [complicated, worse, hopeless, conducive, analogous, desperate, unsatisfactory, unstable, confusing, confused, vacant, tense, grave, favourable, different]* and *[current, present, given, dangerous, similar, particular, intolerable, economic, stressful, difficult, ideal, financial, complex, deteriorating, worsening, etc.] situation*. In the first collocate set, we can identify 11, and in the second 5 out of 15 collocates that could be marked as "negative", i.e., carrying negative implications and associations arising from the extended units of meaning. The concordance shows that prosody is neutral mainly in (semi)terminological contexts. Even though semantic preference and semantic prosody overlap to an extent in general

contexts, semantic prosody is usually about a particular "scenario" rather than merely a "preference" related to a semantic field. Amongst the verb collocates *[react, respond, adapt, adjust, correspond, apply, refer, relate, etc.]* stand out. These are again rather neutral at first glance, but are confirmed as predominantly negative upon the examination of the wider co-text and context:

> *(1) Clientelism is a strategy used by capitalists and workers to adapt to a* ***situation*** *where there is limited mobility.*
> *(2) In parliament it was difficult to adjust to the new* ***situation****, whereby the party was supposed to abstain from all criticism of the government but had no say in its decisions.*

Of the verbs with "situation" as the prepositional object, the most typical are *[cope, deal, face, confront, compare, etc.] with the situation; the first 6 collocates in the genitive relation, which indicate difficulty, are [seriousness, gravity, urgency, reality, complexity, absurdity] of the situation, while an even more explicit reference to concrete sociopolitical conditions is made by the first* 6 telling collocates in the prepositional phrases *the situation in [Somalia, Yugoslavia, Gulf, Russia, Africa, Iraq]*. The research presented in the continuation is based on an examination of the possibilities of including pragmatic information in lexical-lexicographical descriptions from two perspectives: 1) as above, monolingually, identifying and recording pragmatic components in the context, and 2) exploring options for conveying pragmatic information, especially semantic prosody, in bilingual dictionaries. For this purpose, we have analysed a selection of meaning descriptions from the SLD and studied *a prima facie* translation equivalents of the relevant lexical units, drawn from the DANTE lexical database, the selected EFLs[2] and Collins English Dictionary (CED), which were then checked against a selection of bilingual (corpus-based) dictionaries.[3] The bilingual focus remains throughout on the Slovene-English rather than English-Slovene perspective. We did, however, consult two English-Slovene sources to check the degree of bidirectionality of the translations. A tentative typology was compiled in order to categorise lexical units according to their inner semantic-pragmatic relations, with a view to examining the implications for the monolingual dictionary treatment of individual items, as well as any potential strategies that could be applied, on the basis of the posited categories, to their translation. The schematic representation is not based on the structural relations between the components of the lexical units, but is driven by the semantic-pragmatic relations established by each category.

2 Monolingual learner's dictionaries: COBUILD, MED, LDOCE and MWLD. See bibliography.
3 Bilingual Slovene-English dictionaries: Concise Slovenian-English Dictionary DZS (1st Ed.) (PSA), Slovenian-English Pocket Dictionary DZS (MSA), PONS Slovenian-English (& English-Slovenian) Dictionary, DZS-Oxford Comprehensive English-Slovenian Dictionary (VASS) and the automatically reversed VASS database with 120,000 entries, nicknamed OXZILLA DZS.

3 Pragmatic Analysis Based on the Different Situations of Meaning

3.1 Meaning as an Inherent Markedness (at the Morphosyntactic Level): Connotation

Some words can be identified pragmatically as morphosyntactically non-neutral. In the case below, the SLD fails to convey the pragmatic meaning either with a label or in the description:

Example 1: *debeluh* – fatso; fatty

SLD: debel človek (Eng.: a fat person)
(3) ***Fatties*** who lose a lot of weight talk about the need for mental adjustment. /.../[4]
The English sources label both lexical items, "fatty" and "fatso", as informal or slang (DANTE and MWLD), also indicating their connotation with the label "insulting" or with the inclusion in the definition of "an insulting word for" (CED, LDOCE and MED).

DANTE: fatty: 1 n **[offens] [inf]** nickname or appellation or a fat person; fatso /

(4) He reviewed a gallery of the great ***fatties*** of all time, from Nero through Falstaff to Arbuckle.
(5) Go on ***fatty*** !

The appropriate semantic-pragmatic profile of this lexical unit can thus be created with the use of labels or, as in most sources, with a combination of the label and the definition. A key component of meaning is the speaker's intent to insult, in the case of direct address, and, in the case of third person use, to establish a certain distance and/or to express disdain in relation to such a person. Below is a summary of monolingual and bilingual treatment of this lexical unit in the relevant dictionaries and lexical databases.

4 The examples taken from the SLD are translated into English as literally as possible to preserve the original meaning.

COBUILD **fatty**	LDOCE	MED	MWLD	CED
/	fatty [countable] informal. **an insulting word for** someone who is fat	noun [countable] informal. **an insulting word for** someone who is fat	[count] informal + **offensive :** a fat person	(informal) a fat person
fatso				
/	[countable] informal. **an insulting word for** someone who is fat	**an insulting word for** someone who is fat	[count] informal + **offensive :** a fat person	fatso: **(slang)** a fat person: used as an insulting or disparaging term of address
bilingual				
PSA ***debeluh***	**MSA** ***debeluh***	**OXZILLA** ***debeluh***	**PONS** ***debeluh***	**VASS** ***fatty***
pejor. fatso, fatty	pejor. fatso, fatty	inf. Am. **fat-ass**; inf. Br. **lard-ass**; inf. pejor. **fatso, fatty;** inf. Am. **blimp**; [...]	fatty; fatso Am.	pog. žalj. baj(i), bajsa; **debeluh(ar)**, debeluhinja

Table 1: The treatment of *debeluh* in the selected monolingual and bilingual sources.

3.2 Meaning as a Matter of Attitude Towards a Pragmatic Situation

Another tentative, but lexicographically significant category has been posited that will not be treated in detail here due to a lack of space. In some situations of meaning, such as in *plezati* 'make one's way through/over/out of an uncomfortable position with effort', a set of circumstances has been identified that differs considerably in terms of prosodies from the main sense, thus calling for a separate (sub) sense:

Example 2: plezati (čez/skozi kaj; iz česa) – climb (over/through/from sth); clamber (over/across/into/out of)

SLD: če ČLOVEK pleza preko OVIRE, skozi ODPRTINO ali iz neudobnega POLOŽAJA, se skuša s pomočjo celega telesa premakniti v želeno smer, navadno s trudom ali težavo
(Eng.: if a HUMAN climbs over an OBSTACLE, through an OPENING, or from an uncomfortable POSITION s/he, using all limbs and her/his whole body, attempts to move in the desired direction, usually with great difficulty or with some effort)
(6) The studio personnel must sometimes **climb** over heaps of presents, nappies and various toys.

(7) The track is in parts unsurpassable, we have to **climb** up on the dug up deviations.
(8) At that point, the angry supporters at the west stand began to **climb** over the fence, while objects were being thrown to the area next to the play field where the referees were.
(9) Thousands tried to come into the stadium **climbing** over the fences and closed gates.
(10) Often they would have to **climb** through the windows, the side exits and run away from the girls through the kitchen.
(11) The airbag had already emptied itself and released the pressure on my body. I slowly unfastened the safety belt and started **climbing** out of the vehicle – through the co-driver's window!

In this case, with the subject typically human, the act of climbing is largely unplanned or undesired, bringing with it the semantic prosody of anguish or despair arising from the frustration at not being able to move faster or with greater ease. All of the examples of use above contain some element that indicates unfavourable circumstances surrounding the central event, accessible through the speaker's attitude, which is not at all typical of climbing a ladder, a tree or a mountain; therefore, a separate sense is in place to capture adequately the identified pragmatic components.

COBUILD **climb**	**LDOCE**	**MED**	**MWLD**	**CED**
/	with difficulty [intransitive always + adverb/preposition] to move into, out of, or through something slowly and awkwardly	*[intransitive/transitive] to use your hands and feet to move up, over, down, or across something*	always followed by an adverb or preposition [no obj] **:** to move yourself in a way that usually involves going up or down	/
bilingual				
PSA **plezati**	MSA plezati	OXZILLA plezati	PONS plezati	VASS climb
(*with difficulty*) to clamber; to scramble	(*with difficulty*) to clamber; to scramble	clamber; scramble; climb	climb	splezati, (po) vzpeti se na; plezati po; vzpenjati se po

Table 2: The treatment of *climb* in the selected monolingual and bilingual sources.

3.3 Meaning as a Matter of Emphasis

The way meaning and use are in fact two distinct, yet inseparable, facets of language is illustrated by the analysis of the following example, which shows how meaning can arise from an emphasis on a

particular aspect of the (pragmatic) situation rather than from the word's inherent (semantic) features:

Example 2: ***bobnati (pri kom)*** – to drum with/in (=to be the drummer with a group)

The prototypical meaning defined by most (Slovene and English) sources as "to beat or play a drum, or a set of drums" (MWLD) or "to play a drum" (MED, LDOCE) (but not listed in COBUILD) is not the most frequent meaning of *bobnati*. On closer inspection and driven by a pragmatic function of meaning a subsense can be separated out, i.e., to play the drums in a particular band or group on a regular basis:

SLD*: če ČLOVEK **bobna**, se posveča igranju na bobne kot stalni član glasbene zasedbe, navadno v določenem obdobju (Eng: if a HUMAN **drums** s/he is engaged in playing the drums as a regular member of a music group, usually for a period of time)
(12) The new drummer is Nenad Kostadinovski, who used to **drum** with groups such as Scuffy Dogs and Traffic Religion.
(13) Do you then have a drummer for your concerts? At the concerts we use Moreno Buttinar, who is Lara Baruca's drummer. We have also practiced with Janez, who **drums** with Miladojka Youneed.
(14) Meanwhile, Eva and Nataša sing and **drum** on Laibach's concert tour, while Darja has used the short period of lesser working intensity really well and has freshly fallen in love.
(15) Cecil Durkin was a druggy, a knife cutter and a prison rapist, but he also **drummed** in a few good jazz bands.
This use of the verb is distinct, as its colligational behaviour in particular indicates: it typically requires explicit temporal complementation (recently, at night, later, on tour, in the election time, sometimes, etc.) or the time of the action is implied in the context by the use of, typically, a past tense. It is commonly used with prepositional complementation (drum with, at, in, etc.) denoting individual people or groups with whom one drums; proper names, therefore, appear regularly in the co-text as part of listings and/or coordinate structures with "and". Of course, the prototypical meaning "to play a drum" is still present, but it is now a secondary rather than the key component of the conveyed sense. The emphasis is on the fact of being engaged as member of a performing musical group, often in the context of other players who make up the group, i.e., on the role ensuing from the ability to play the drums. A colligational feature is that the verb cannot be pre- or postmodified, e.g., by an adverb. The semantic preference is for musical groups, players and settings, from which the association of an opportunity for success and fame emerges. In some cases semantic prosody, which seems to lie here first and foremost in "renewing the connection of this semantic information with the reality of language in use /.../," (Philip 2009) relies more heavily on the colligates than the collocates alone, "if anything tending to favour the patterns and participants in verbal processes over lexical-semantic features per se" (ibid.).

COBUILD	LDOCE	MED	MWLD	CED
/	to play a drum	Music. to play a drum	to beat or play a drum or set of drums	to play (music) on or as if on a drum
bilingual				
PSA **bobnati**	MSA bobnati	OXZILLA bobnati	PONS bobnati	VASS to drum
(as a profession) to play the drums	(as a profession) to play the drums	to drum	to drum	bobnati

Table 3: The treatment of *bobnati* in the selected monolingual and bilingual sources.

3.4 Restricted Meaning in Semantically Analysable Units: Collocation

This is probably the most widespread, yet lexicographically somewhat neglected lexical category that lies at the very core of semantic prosody. Semantic analysability can be observed along a continuum stretching from collocations – commonly thought to be transparent, but so only in their restricted meanings, i.e., in only one of the possible meanings resulting from the various meanings of their components, as well as of their various combinations – to the most opaque idioms that lie at the other extreme (Philip 2009). The collocation below can be observed in its restricted meaning "dull weather":

Example 4: *kislo vreme* – sour/grey/dull weather

SLD: kislo VREME je takrat, ko ni sonca ali dežuje (Eng: sour WEATHER is when it rains or the sky is overcast)

The only two collocates that stand out are *zagosti* and *pokvariti* ("to spoil"), paralleled in scarcity by colligational patterns. The semantic preference is for cultural and sports events, and, within a limited spectrum, for agricultural products, especially grapes which are expected to ripen and develop sugar in the sun. The semantic association of physical and mental discomfort caused by the weather conditions helps build up the semantic prosody based on imminent danger of poor turnouts at public events or people not going ahead with their outdoor plans, such as tourists cancelling their bookings:
(16) A general characteristic is that the camping sites in Gorenjska are pretty full, and even the **sour weather** of recent days has not chased away the tourists.
(17) That is why the construction workers are working at a good pace, but, on the other hand, due to **sour weather** the owners of Bioterme are in no hurry to open the swimming pool.
(18) The expected **sour weather** will cause malaise or indisposition in many people.

(19) When the summer is sunny and September alike, the grapes will be sweet, but **sour weather** will give us grapes that will be hard to sell

An important fact about *kislo vreme* is that it is often used with its verbal collocates to express the opposite, i.e., to convey that unfavourable conditions did not, in fact, have the expected effect and did not put people off from coming and/or having a good time. Colligationally, therefore the use of negation is noticeable:

(20) The good spirits were not destroyed even by somewhat **sour weather**, which towards noon cleared up, so they set off for short or longer walks in the surrounding areas [...].

COBUILD **dull**	**LDOCE** **grey/dull weather**	**MED** **dull**	**MWLD** **dull**	**CED** **dull**
You say the weather is **dull** when it is very cloudy.	cloudy and not bright	if the weather is dull, there are a lot of clouds and it is rather dark	not sunny **:** having a lot of clouds ▪ *a dull winter sky*	(of weather) not bright or clear; cloudy
bilingual				
PSA **kislo vreme**	MSA **kislo vreme**	OXZILLA **kislo vreme**	PONS **kislo vreme**	VASS dull sky
bad, nasty, foul	bad, nasty, foul	/	/	oblačno nebo

Table 4: The treatment of *kislo vreme* in the selected monolingual and bilingual sources.

3.5 Meaning as Encyclopedic Knowledge: (Terminological) Compounds

There are some words and phrases that cannot be understood without knowledge of the real world or so-called encyclopedic knowledge, such as compounds (*kislo zelje* - sauerkraut) conveying different degrees of (semi)terminological meaning. Semantic prosody is the least prominent here, as the meaning has already been fixed by word-semantics and extra-linguistic knowledge. Due to this fact, and lack of space, we will not treat this category in more detail.

Meaning as (Pragmatic) Knowledge about Language Use: Idiomatic Expressions

In some expressions, where the salient (or metaphorical) meaning of either of words is insufficient for the reader to know their overall meaning, "word-semantics are redundant and yield entirely to the pragmatic reality of use, both textual and contextual, as the meaning of the phrase relies heavily on knowledge of semantic associations and semantic prosody" (Philip 2009):

Example 5: ***deklica za vse*** – dogsbody; girl/gal Friday

SLD: nekdo, ki kje opravlja najrazličnejša dela, od najnižjih do najzahtevnejših

(21) Having finished school, she, soon after World War II, got a job with the Slovenian Railways where she persisted for almost 34. She was a ***general dogsbody***: worker, booking clerk and paymistress at the Head Office of the Slovenian Railways in Ljubljana.
(22) Jože Klemenčič is a ***general dogsbody*** in Slovenian langlauf: vodja, koordinator in pomočnik trenerja.
(23) He is now involved fixing computers, in the shop and on the terrain, but he calls himself a ***dogsbody*** as he does all sorts of jobs.

The concordance to *deklica za vse* shows that the expression is used both for men and women, and that, interestingly enough, it is not possible to identify a clearly negative attitude towards this enforced role. The emphasis is on the variety and unpredictability of the tasks that one is expected to perform, which in some contexts even conveys positive prosodies ensuing from the fact that such a person typically displays positive qualities such as dedication, resourcefulness and efficiency in the assigned tasks.

DANTE: dogsbody: n [inf] [non_AmE] sb who has to do any unpleasant jobs that nobody else wants to do

(24) We pull over to the side of the road and 10 minutes later meet George Wolter who is to be our guide, host, translator, organiser and ***general dogsbody*** for the next 10 days. George is a freelance translator in English and Spanish and has been working for the Political Song Festival for the last 10 years.
(25) Then I suppose, when the festival had opened, I'd already gone to the site with the school, before I left, in the Easter of 1951 and then I started an apprenticeship with Vickers, but I had to wait some time before I could actually start, and I was given the job as a ***dogsbody***, you know a fetcher and carrier.
(26) STRUCTURE N_mod➪ He started off as '***kitchen dogsbody***' at 13 and went on to study at Westminster Catering College. /.../

According to the English database, "dogsbody" can be premodified by an adjective, as in "general dogsbody". In the English-Slovene PONS, we do indeed find "general dogsbody" with the translation *deklica za vse*, while for the entry *deklica za vse* it lists "jack-of-all-trades", a unit less frequent and perhaps semantically more comparable with *mojster za vse*, which evokes associations of "to be a jack-of-all-trades (and master of none)" and thus conveys negative attitudes. The suggested translations are relatively consistent in all of the bilingual sources (dogsbody, Br. Eng./girl Friday, informal old-fashioned);

however, the interlingual differences as well as the monolingual differences (synonymy) should be explicitly highlighted. As LDOCE, MED and MWLD report, "girl Friday" is old-fashioned and seems to refer exclusively to women who do several different jobs in an office, or are, in fact, secretaries (e.g., MED, also MWLD). "Dogsbody", on the other hand, is used for both sexes and has, in comparison with the Slovene *deklica za vse*, a much more evidently negative connotation (a person is forced to do the jobs that other people refuse to do). In addition, the functional (pragmatic) meaning of "dogsbody" is different to an extent, conveying feelings of bitterness due to being systematically exploited, which leads us to posit negative prosody.

COBUILD dogsbody	**LDOCE**	**MED**	**MWLD**	**CED**
British, informal A dogsbody is a person who has to do all the boring jobs that nobody else wants to do.	someone who has to do all the small boring jobs that no one else wants to do	**British informal** *someone who is forced to do all the jobs that no one else wants to do*	**informal + old-fashioned :** a woman who does many different jobs in an office	**informal** a person who carries out menial tasks for others; drudge
girl Friday				
	British English old-fashioned a girl or woman worker who does several different jobs in an office	**informal old-fashioned** *a female secretary*	**informal + old-fashioned** a female office assistant	a female employee who has a wide range of duties, usually including secretarial and clerical work
bilingual				
PSA deklica za vse	MSA deklica za vse	OXZILLA deklica za vse	PONS deklica za vse	VASS dogsbody
dogsbody BR; girl/gal Friday esp. AM	dogsbody BR; girl/gal Friday esp. AM	dogsbody; girl Friday; pejor. Am. ward-wheeler	jack-of-all-trades	deklica za vse; dekla, hlapec

Table 5: The treatment of *deklica za vse* in the selected monolingual and bilingual sources.

4 Discussion

In the SLD, an attempt was made to incorporate, to a maximum degree, corpus-derived pragmatic components, i.e., semantic prosody, into meaning descriptions by means of various definition strategies. One way of highlighting a particular (shade of) meaning is also to introduce a new (sub)sense. Thus the concordance for *plezati čez/skozi kaj, plezati iz česa* shows consistently unfavourable condi-

tions in which *plezati* occurs, expressing a negative attitude of the speaker towards the central act of "climbing", which in the other listed senses appears to be neutral, if not positive. However, with the exception of LDOCE, the English monolingual sources do not identify this use, nor does the Dictionary of Standard Slovene, while both corpus-based Slovene-English dictionaries do, listing it as a new subsense introduced by the semantic indicator (*s težavo se premikati*) (to move with difficulty), which in most cases will yield the translation "clamber" rather than "climb". In the selected subsense of *bobnati* treated in the SLD, the emphasis is on the fact of being engaged as member of a performing musical group, rather than on the actual physical act of playing a drum, which is indicated by a separate subsense. The bilingual sources do not recognise a separate sense, with the exception of PSA and MSA, which list the relevant subsense introduced by (*kot poklic*) (as a profession) and provide an alternative translation "to play the drums". Collocational meaning, such as in *kislo vreme*, is not regarded as semantically transparent but restricted in semantically analysable units. The most extensive collocational range for "X weather" is provided by LDOCE and MED, with the most typical collocates being "dull" and "grey". The semantic equivalent "sour weather" is practically non-existent, but there are some corpus occurrences, e.g., "the weather turned sour with thunderstorms and heavy rain," confirming the translation, albeit hypothetically. The semantic prosody in the case of "dull/grey weather" cannot be paralleled to that of *kislo vreme* as described in 3.4, and there are not enough corpus examples of "sour weather" to establish its prosody, which could be interlingual. The type of meaning referred to as encyclopedic, such as in *kislo zelje (sauerkraut)*, is very much dependent on real world knowledge, but much less sensitive to axiological aspects of meaning, which is generally the case in (semi)terminological lexical units. Determining interlingual equivalence is often rather straightforward and, in a monolingual perspective, is the least prone to language change. Finally, with reference to idiomatic units of meaning, which are normally treated as phraseological units, it can be said about *deklica za vse* that its functional equivalent "dogsbody" conveys the subject's feelings of bitterness due to being systematically exploited, which leads us to posit negative prosody; "girl Friday", on the other hand, can only refer to a woman and is, as COBUILD and MWLD demonstrate, old-fashioned. The choice of the translation – in our case "dogsbody", "girl Friday" or "jack-of-all-trades" – is dependent on the circumstances of meaning, and on whether (or not) semantic prosody, which often plays a part in phraseology, is detected in the source language. The analysis of lexical units has shown that we are not dealing with full equivalents, but, at best, with comparable idiomatic units, where meaning requires greater involvement of pragmatic knowledge.

Based on a tentative typology of meaning we have discussed a number of strategies for tackling pragmatics monolingually: first we dealt with an example of connotation where in some sources an adequate label (pejorative/disparaging) was applied, while other sources combined labelling and defining (informal/slang + an insulting word for). The focus of the analysis, however, was on those instances that encode pragmatic meaning as part of their meaning, rather than that which is traditionally subsumed under connotation. In the SLD, some meanings that would traditionally be labelled as "humorous" or "showing disapproval" were, where possible, made part of the definition. Usually the first part

of the definition provides the general semantic-syntactic pattern and the second describes pragmatic circumstances, including the semantic prosody where appropriate. In other words, the first part of the definition is a straightforward explanation, while typically the second part tells us about the speaker's attitude to the meaning situation and, expressing the pragmatic function, clarifies why a particular lexical choice has been made by the speaker. We have also pointed out the treatment of individual lexical units in the selected bilingual dictionaries. Any indication of prosodic meaning is consistently implicit, even when the prosody in the two languages is different, or when only a particular (sub)sense of L1 carries prosodies and the translation differs accordingly. The identified ways of conveying pragmatics in the examined bilingual dictionaries are: functionally adequate translations, i.e., a new (sub)sense (with an appropriate semantic indicator), such as in *plezati: (s težavo se premikati)*, examples of use and labels. The bilingual focus remains throughout on the Slovene-English rather than English-Slovene perspective. Considerable progress is noted in the treatment of language use in contemporary bilingual corpus-based dictionaries and databases, such as PSA and VASS.

5 Conclusion

Corpus-derived facts that concern axiological aspects of meaning, such as those related to "situation" and the other lexical units examined, can be explicitly described in monolingual lexical databases and dictionaries. Lexicography, by definition, is concerned with the investigation and recording of all aspects of lexical meaning. Semantic prosody can be viewed as a link between the lexical and the textual or discoursal levels. By including information not only on collocational but also lexical-textual co-selection, we are bound to improve the dictionary, this practical tool, equally serving language learners, translators and interpreters, as well as communicologists, copywriters, etc. It is well known (cf. Partington 1998: 72) that signals on semantic prosody are particularly important for second or foreign language learners, as they do not have the subconscious understanding of pragmatic meaning that, presumably, native speakers do. However, assuming that semantic prosody is part of a dictionary "sense" and, in many cases, the key to its actual identification, explicating prosodic meaning in native-speaker dictionaries cannot be considered insignificant. Establishing a difference between the lexis that requires labelling due to its connotational meaning that is morphosyntactically coded (*fatty, fatso, bunny,* etc.) and the lexis that displays other, contextual types of pragmatic meaning naturally leads to three basic solutions: a label, a definition or both. Collocational meaning, particularly semantic prosody, which is by nature delexical, functional, phraseological, textual and abstracted from various contextual features – which is why some authors have described the phenomenon as "collocational", "discourse" or "pragmatic" prosody (e.g., Stubbs 1995, 2001), or "semantic harmony" (Lewandowska-Tomaszczyk 1996) – should be presented as part of the definition or in an additional gloss. The question remains as to how exactly semantic prosodies should be presented in bilingual dictionaries, and this is currently under investigation in the doctoral research.

6 References

6.1 Corpora, Dictionaries and Databases

British National Corpus (BNC). Accessed at: https://the.sketchengine.co.uk/.
Collins COBUILD English Dictionary for Advanced Learners, HaarperCollins Publishers (COBUILD). Accessed at: http://www.collinsdictionary.com/dictionary/english-cobuild-learners.
Collins English Dictionary (CED). Accessed at: http://www.collinsdictionary.com/dictionary/english.
DANTE - A Lexical Database for English. Accessed at: http://www.webdante.com/.
Dictionary of Standard Slovene (SSKJ). Accessed at: http://bos.zrc-sazu.si/sskj.html.
Longman Dictionary of Contemporary English Online (LDOCE). Accessed at: http://www.ldoceonline.com/.
Macmillan English Dictionary Online (MED). Oxford: Macmillan Education. Accessed at: http://www.macmillandictionary.com/.
Mali angleško-slovenski in slovensko-angleški slovar (MSA) = English-Slovenian & Slovenian-English Pocket Dictionary. 1st edn. Zaranšek, P. ed. 2006. Ljubljana: DZS.
Merriam Webster's Learner's Dictionary (MWLD). Accessed at: http://www.learnersdictionary.com/.
PONS Splošni slovensko-angleški slovar in angleško-slovenski slovar (PONS)
Accessed at: http://www.pons.si/pons-splosni-slovensko-angleski-slovar-p-518.php.
Priročni angleško-slovenski in slovensko-angleški slovar (PSA) = Concise English-Slovenian & Slovenian-English dictionary. 1st edn. Drinovec Sever, N., Pogačnik, A., Žerak A. eds. 2010: Ljubljana: DZS.
Slovene Lexical Database (SLD). Accessed at: http://www.slovenscina.eu/spletni-slovar.
Veliki angleško-slovenski slovar Oxford-DZS (VASS) = Oxford-DZS Comprehensive English-Slovenian Dictionary. Krek, S. (ed.) [Vol. 1, 2005; Vol. 2, 2006]. Ljubljana: DZS.

All links accessed [05/04/2014].

6.2 Other Literature

Ellis, Nick. C., Frey, E. in Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1): Lexical Access. In *Exploring the lexis-grammar interface*, U. Romer, R. Schultze (eds.). Amsterdam/ Philadelphia: John Benjamins.
Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: CUP.
Lewandowska-Tomaszczyk, B. (1996). Cross-linguistic and language-specific aspects of semantic prosody. *Language Sciences*, 18, pp. 153-178.
Louw, W. E. (1993). Irony in the Text or Insincerity in the Writer?: The Diagnostic Potential of Semantic Prosodies. In *Text and technology: in honour of John Sinclair*, M. Baker, G. Francis and E. Tognini Bonelli (eds). Amsterdam: John Benjamins, pp. 157-176.
Partington, A. S. (1998). Patterns and Meanings. Using Corpora for English Language Research and Teaching. SCL 2. Amsterdam and Philadelphia: John Benjamins.
Philip, G. (2009). Why prosodies aren't always there: Insights into the idiom principle. Corpus Linguistics Conference. Liverpool. Great Britain. http://ucrel.lancs.ac.uk/publications/cl2009/ . [Accessed 06/09/2012].
Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.
Sinclair, J. M. (1996). *The Search for Units of Meaning*. TEXTUS IX.

Stubbs, M. (1995). Corpus evidence for norms of lexical collocation. In *Principle and Practice in Applied Linguistics*, Cook, G. and Seidlhofer, B. (eds.), Oxford: Oxford University Press, pp. 245-6.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Šorli, M. (2013). Forms of encoded pragmatic meaning: semantic prosody: a lexicographic perspective. In *Lingue e linguaggi* [Online ed.], 2013, Vol. 10, pp. 95-111.

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

Whitsitt, S. (2005). A critique of the concept of semantic prosody. In *International Journal of Corpus Linguistics* 10, pp. 283-305.

Zethsen, K. K. (2006). Semantic Prosody: Creating Awareness about a Versatile Tool. *Tidsskrift for Sprogforskning*, 4(1), pp. 275-294.

Zhang, W. (2009). Semantic Prosody and ESL/EFL Vocabulary Pedagogy. *TESL Canada Journal/Revue TESL DU CANADA*. VOL. 26, NO 2, SPRING 2009

Bi-and Multilingual Lexicography

Linking a Dictionary to Other Open Data – Better Access to More Specific Information for the Users

Ulrich Apel
Eberhard Karls University Tübingen
ulrich.apel@uni-tuebingen.de

Abstract

The project *WaDokuJT* (2013) has developed into the most comprehensive Japanese-German dictionary in regard to covered lemmata and translation equivalents. The dictionary is used by many users from German speaking countries and from Japan. The project was designed concentrating on German speaking users who want to read and translate Japanese texts. So, priority is given to a rather complete coverage of orthographic variations of Japanese headwords on the one side, and German translation equivalents on the other side. Definitions are given only when they are really necessary and only in German language; further, they are as short as possible.

This means that the dictionary project may have certain shortcomings for text production, for Japanese users or for users who need deeper encyclopedic explanations connected to a certain dictionary headword.

This paper presents an approach how to alleviate such deficiencies by providing hyperlinks or other references to data of other dictionaries and encyclopedia projects which may contain the information, users are looking for. For this aim, *WaDokuJT* data refers especially to open source projects, dictionaries, which aren't protected by copyright anymore, and collaborating projects, the data of which can be accessed easily at least in parts.

Keywords: References; Japanese; German; hyperlinks

1 *WaDokuJT* – Project overview

The Japanese-German dictionary project *WaDokuJT* (2013) started in 1998 as individual initiative at Osaka University (Apel, 2001). In the meantime it covers approximately 115,000 lemmata and around 275,000 database records. Records like derivations, compounds, examples of usage and example sentences are not main headwords.

Records contain the Japanese lemmata, their pronunciation written in Japanese syllable alphabet – *hiragana* – and translation equivalents. Orthographical variety in Japanese is represented in more than half a million different written forms of the entries. The *kana* transliteration is extended by a mark-up to calculate Rōmaji transcription following the new standard *DIN 32708 – Transliteration of*

Japanese (*NA 009 Normenausschuss Bibliotheks- und Dokumentationswesen*, 2013). The number of translation equivalents is more than half a million.

2 Weakness investigation of the project and possible improvements

The project was designed mainly for text reception of German users. Sometimes additional information like domain, definitions, explanations etc. are given in German, too. Some information for text production is added, like the Japanese pitch accent for correct pronunciation or conjugation types for Japanese verbs. A certain concession for Japanese users is the addition of a mark-up for the gender of German headwords within the German translation equivalents.

The dictionary's limitations seem to be not too serious, as the dictionary is used by Japanese users and also for text production.[1] Nevertheless, improvements are possible. Further information could be added directly to the dictionary with a specific mark-up and would be displayed only for certain user profiles or use situations. Unfortunately this would complicate the data management not to mention the lack of resources to provide such detailed information. Another way to make further information available without too much trouble for the users and lexicographers is to include links and references to other projects or dictionaries with additional information.

Giving the example of this dictionary project, the paper presents how such links can be added by hand or with automated suggestions of link candidates.

3 Linking to data in public domain or with open source license

3.1 Großes Japanisch-Deutsches Wörterbuch from 1937

The most comprehensive printed Japanese-German dictionary was edited by Kinji Kimura and was published first in 1937. It is still reprinted since then (*Großes Japanisch-Deutsches Wörterbuch*, 1952). The

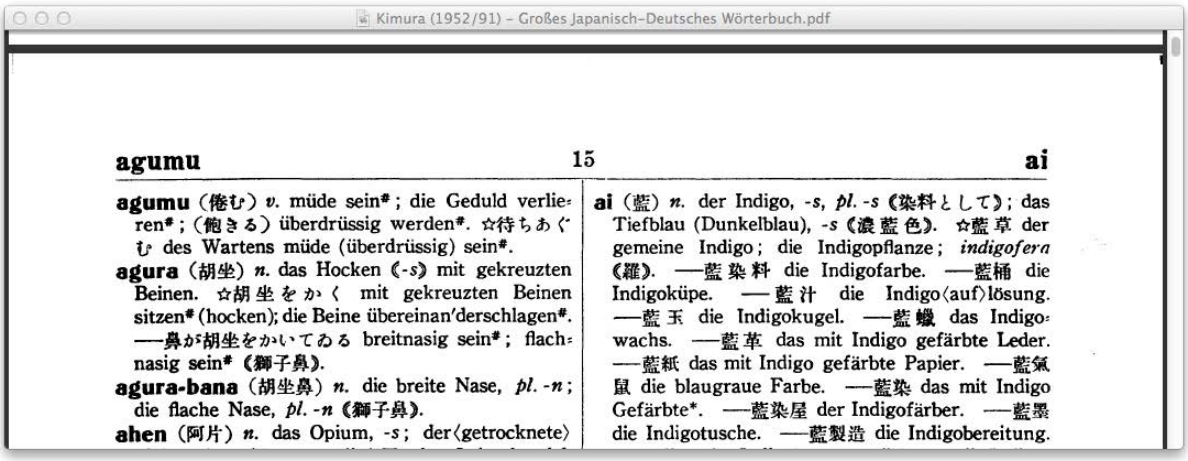
1 Actually, the dictionary is often used as German-Japanese dictionary, since the German translation side can be easily searched, too. This of course is not recommended because users may get overwhelmed by possible Japanese translations, some of which are outdated or are used only in rather specific situations. Since there is quite a number of rather good German-Japanese dictionaries, there would be enough alternatives, although most of these dictionaries concentrate on the needs of text reception by Japanese users (e.g. *Shogakukan Großes Deutsch-Japanisches Wörterbuch*, 1998, or *Wörterbuch der deutschen und japanischen Gegenwartssprache. Deutsch-Japanisch*, 1989).

author deceased in 1948, and the copyright expired in 1998 – fifty years later according to Japanese copyright laws.

This dictionary seems to be written with the purpose to enable Japanese diplomats and writers to explain Japanese intentions and ways of thinking to Westerners and especially to speakers of German. Different meanings of polysemous words are defined in Japanese and the corresponding translation equivalents are given in German. The German equivalents are completed by articles to show gender, by verb conjugation, by noun and adjective declension or by adjective comparison etc.

This information is still rather valuable, especially for Japanese users and for German text production, even though the Kimura dictionary is outdated in certain aspects. For example, it doesn't reflect orthographic reforms in Japan and Germany or it contains many Japanese names for cities in Manchuria, which was *de facto* a Japanese colony at the time of the compilation. Nowadays, these names are of very limited use.

For German users the Kimura dictionary is also lacking for example the pronunciation of Japanese subentries, the conjugation of Japanese verbs and further German explanations.


Kimura (1952/91) – Großes Japanisch-Deutsches Wörterbuch.pdf

agumu 15 ai

agumu (倦む) *v.* müde sein#; die Geduld verlieren#; (飽きる) überdrüssig werden#. ☆待ちあぐむ des Wartens müde (überdrüssig) sein#.
agura (胡坐) *n.* das Hocken (*-s*) mit gekreuzten Beinen. ☆胡坐をかく mit gekreuzten Beinen sitzen# (hocken); die Beine übereinan'derschlagen#. ——鼻が胡坐をかいてゐる breitnasig sein#; flachnasig sein# (獅子鼻).
agura-bana (胡坐鼻) *n.* die breite Nase, *pl.* *-n*; die flache Nase, *pl.* *-n* (獅子鼻).
ahen (阿片) *n.* das Opium, *-s*; der⟨getrocknete⟩

ai (藍) *n.* der Indigo, *-s*, *pl.* *-s* (染料として); das Tiefblau (Dunkelblau), *-s* (濃藍色). ☆藍草 der gemeine Indigo; die Indigopflanze; *indigofera* (羅). ——藍染料 die Indigofarbe. ——藍桶 die Indigoküpe. ——藍汁 die Indigo⟨auf⟩lösung. ——藍玉 die Indigokugel. ——藍蠟 das Indigowachs. ——藍革 das mit Indigo gefärbte Leder. ——藍紙 das mit Indigo gefärbte Papier. ——藍氣鼠 die blaugraue Farbe. ——藍染 das mit Indigo Gefärbte*. ——藍染屋 der Indigofärber. ——藍墨 die Indigotusche. ——藍製造 die Indigobereitung.

Figure 1: Screenshot of a sample page of the *Großes Japanisch-Deutsches Wörterbuch* (1952) in a digitalized version as PDF with the example entry of *agura*.

The *WaDokuJT* project is now adding gradually tags that refer from one *WaDokuJT* entry to the corresponding page, column and entry count of the Kimura dictionary. This process involves quite a lot of manual operations, but it can be supported by electronic means. A scan of the Kimura dictionary had enough quality to run an optical character recognition for Latin characters.[2] Since both dictionaries contain headwords in Rōmaji transcription, one can set up a database relation using the pronunciation transcription as a common key.

2 Japanese character recognition doesn't work well with pre-war orthography, since modern OCR technique concentrates on the most probable characters and prefers contemporary short forms. Further, Japanese OCR has problems with e.g. German umlauts.

Unfortunately, the Japanese language has a lot of homonyms, especially as a result of the intensive borrowing of loan words from Chinese which were adapted to Japanese pronunciation while losing some distinguishing features in the process. For example, the Kimura dictionary gives 38 entries with the pronunciation of *ko* or *kō*.

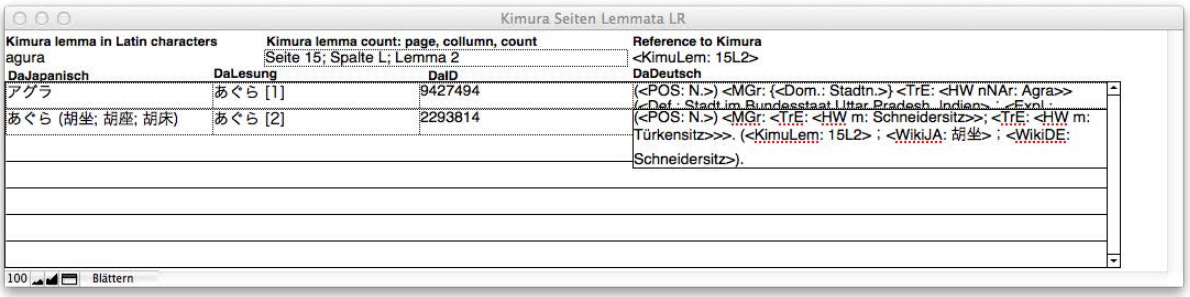


Figure 2: Screenshot of the entry *agura* in a database layout relating to two *WaDokuJT* entries with the same pronunciation.

A relational database being able to deal with homophony – or to be more precise with identical transcriptions of the pronunciation in Latin characters – is shown in Figure 2. The entry with the pronunciation *agura* relates potentially to two entries of the *WaDokuJT* dictionary. The screenshot shows them displayed in a portal of a database layout using a desktop application. A human editor can then enter the reference to the corresponding Kimura entry to the *WaDokuJT* entry via script and a defined hotkey.

A similar approach is used by a students' project at the Institute of Asian Studies – Department of Japanese Studies of Tübingen university in an online version. The project makes serious and steady progress, and, since the first few thousand entries are already covered, it is only a matter of time until the whole dictionary is adapted in this way.

The next step is to give the users access from the *WaDokuJT* webpage to the Kimura information. We will ad a hyperlink to the interface of the online dictionary which will open a scan of the page with the corresponding entry from the Kimura dictionary. This process will be very similar to the one that is explained in Paragraph 4.2, where a hyperlink from the online *WaDokuJT* dictionary opens a page in *Google Books*.

Linking the *WaDokuJT* dictionary with the Kimura dictionary will hopefully also lead to a new correction iteration of the *WaDokuJT* data. As for instance, missing translation equivalents or example sentences from the Kimura dictionary can be added more easily.

3.2 German and Japanese *Wikipedia*

The Japanese *Wikipedia* (2013) is the largest edition of *Wikipedia* in a non-European language. Although the worth of *Wikipedia* as a primary source or reliable references is disputed in academia, it is an easy accessible and important resource. In most cases, it gives a better overview on a certain topic

than would be possible in any ordinary bilingual dictionary. As the *WaDokuJT* dictionary offers mainly translation equivalents and only short hints on meaning or very short definitions, more extensive explanations might be of advantage to many users. *Wikipedia* often refers to corresponding entries in other languages, and Japanese-German pairs can be found, too.

In the source data of the *WaDokuJT* project, the *Wikipedia* article's title is used as unique reference with a mark-up for *Wikipedia* and its language version. As far as possible, *WaDokuJT* refers to both, the Japanese and the German *Wikipedia*. Unfortunately, in many cases a Japanese *Wikipedia* entry has no German equivalent, and often entries, that are marked as corresponding, differ relatively far in meaning and can definitely not be considered as translation equivalents.

The process of finding correspondent *Wikipedia* articles can be automated to a certain extend, although orthographic variety of Japanese makes this process more difficult. Sometimes, several trials with different orthographical forms are necessary. Names of plants and animals for example are written in *katakana* characters within the Japanese *Wikipedia* – as is customary with scientists in the domain of biology – while many traditional dictionaries use *kanji* writings. The *WaDokuJT* project tries to give the most frequent Japanese writing first. Doing so, it is possible to generate automatically links for example to the *Wikipedia* with high probability to work correctly.

Figure 3: Screenshot of the database interface for working on links to the Japanese and German *Wikipedia*.

The screenshot in Figure 3 shows a window from the database application that is used to add links in the *WaDokuJT* dictionary to *Wikipedia* from existing entries. The example entry has four Japanese writings and a generated link to the Japanese *Wikipedia*. A "Web Viewer" field displays the content of this link and enables the editor to check easily the correct correspondence.

A similar link to the German *Wikipedia* can be generated using the first German translation equivalent. If these links between *WaDokuJT* and different *Wikipedia* versions work correctly in spite of possible polysemy or homography, a certain mark-up shorthand as reference to the Japanese or the German *Wikipedia* can be added via script and a hotkey.

The advantage of such a reference for users is, besides getting the translation equivalents which are to be expected from a bilingual dictionary, an easy access to up-to-date encyclopaedic information as well and in many cases even directly both in Japanese and in German.

4 Cooperations with partially open source projects

Currently, the *WaDokuJT* project is cooperating with two domain specific dictionaries, which are the topic of this paragraph. Further cooperation with specialists in other fields is discussed, also. The domains are for example: disaster prevention, mechanical engineering, economy, medicine and life sciences, traditional craft professions and martial arts. Thanks to the great variety and most diverse possibilities, one can be rather optimistic about the results.
The main characteristic of the presented approach is that every project can keep its individuality and its own strengths. The projects shouldn't compete for users or financing, but complement each other in very productive way.

4.1 *Sōgō bukkyō daijiten* – a Buddhist encyclopedia

One cooperating dictionary project is the translation of a Buddhist encyclopaedia (*Sōgō bukkyō daijiten*, 1987; Aoyama et al. 2006; Aoyama et al. 2013). Since a number of years, the articles are translated and extended by information that is relevant for German speaking users. Recently, these data were converted into an online version stored on the same server of Tübingen University as the *WaDokuJT* dictionary.
The entries of the Japanese version of the lexicon are in the Japanese *a-i-u-e-o* order. The online version identifies them by page and count of the entry on the very page. *WaDokuJT* uses this information as unique ID to refer to the Japanese version of the Lexicon as well as to the German translation of the online dictionary. The data of the encyclopaedia will be easier accessible and *WaDokuJT* can concentrate on the translation equivalents and leave buddhological explanations to that lexicon.

4.2 Japanese-German archaeological dictionary and its new Japanese-English-German version

Another collaborating project is an archaeological dictionary project, of which a Japanese-German version already exists as a print-on-demand book: *Kleines Wörterbuch zur Japanischen Archäologie* (Steinhaus, 2010). This project addresses scientific archaeologists and tries to give the latest and often normative translation equivalents, reflecting the current state of the art in the field.

The *WaDokuJT* project has obtained an electronic version of the book from the author and now points from one *WaDokuJT* entry to the corresponding page of the archaeological dictionary. With this information, one can also generate a hyperlink which opens the dictionary page in *Google Books*.

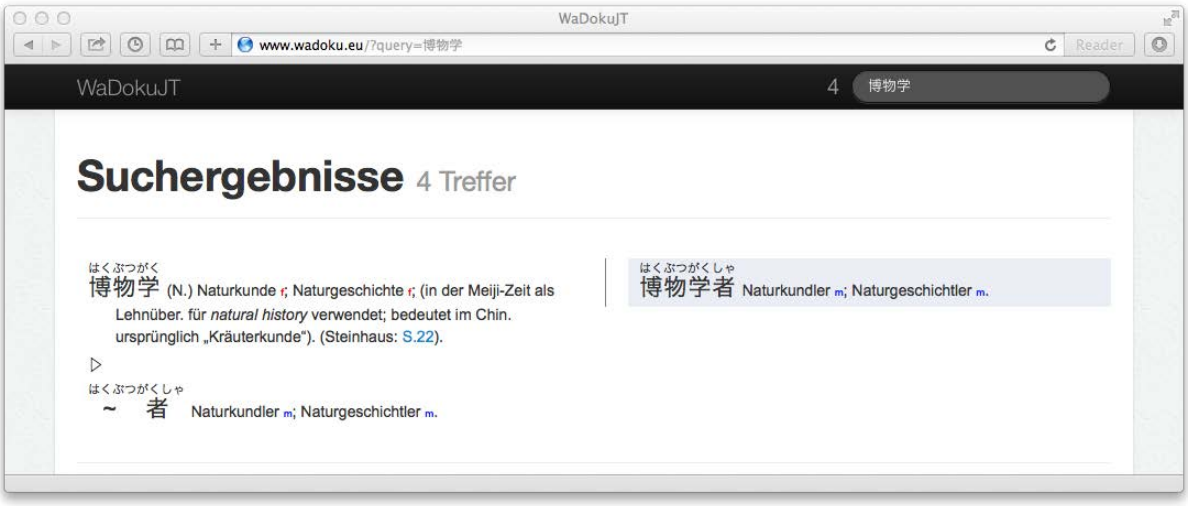


Figure 4: Web interface of WaDokuJT, displaying an entry with hyperlink to Steinhaus *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch* (2010) at Google Books.

Figure 4 is a screenshot form the online version of the *WaDokuJT* project *wadoku.eu* which shows an entry with reference to Steinhaus *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch* (2010). The page information is also a hyperlink which by clicking it opens the mentioned page of the book in its online version at *Google Books*.

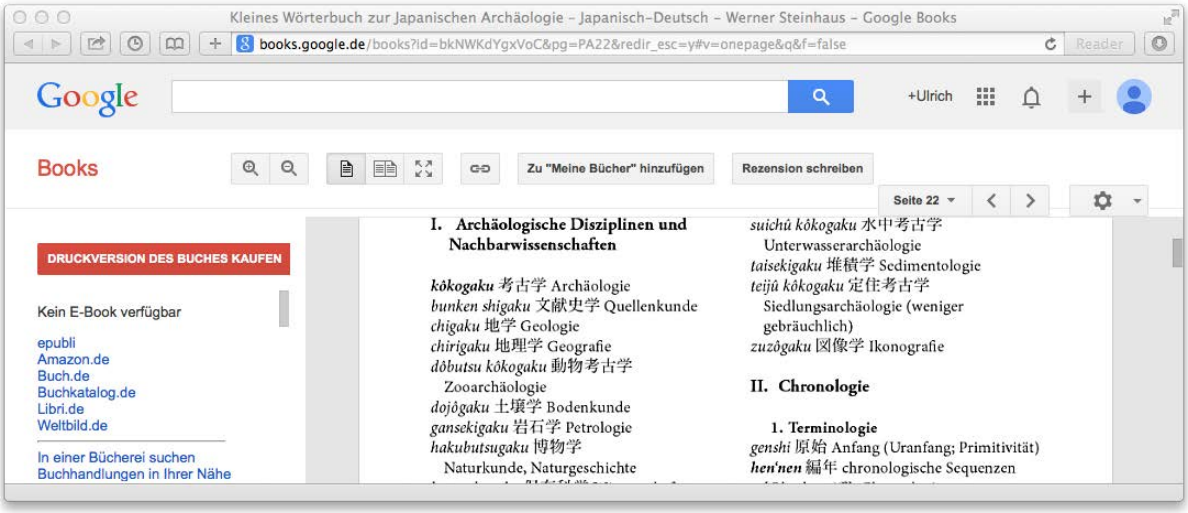


Figure 5: Target page at *Google Books* with the entry *hakubutsugaku* "Naturkunde, Naturgeschichte".

Figure 5 shows a screenshot of the corresponding page of Steinhaus (2010) which is opened at *Google Books*. Here, the translation equivalents are more streamlined for the needs of archaeologists, who

aren't interested in e. g. linguistic explanations about when the word came into usage in Japan and what it meant in its Chinese version.
One main feature of the original archaeological dictionary is the arrangement of entries around certain topics. This means that the archaeological dictionary and the *WaDokuJT* project are no competitors. Further, *WaDokuJT* has no claims of being normative for certain scientific fields but gives possible translations. Through the reference to the archaeological dictionary, users get easy access to the normative translations, too.
The development of the archaeological dictionary goes on in a rather fast pace. In collaboration with the University of East Anglia, Norwich, Great Britain, it is extended to a Japanese-English-German version. In addition, it is planned, that the new version will be hosted at Tübingen as well.

5 Projects linked to *WaDokuJT*

Linking dictionaries to other projects doesn't need to be a one-way-street. Other projects link to the Japanese-German data. For example, *JMdict* (2013), a multilingual lexical database with Japanese as the pivot language uses *WaDokuJT* data for German translations.
Further, a new Korean-German dictionary *Handok.eu* (2013) refers to *WaDokuJT* entries via their ID. Korean and Japanese use a lot of common loan words from Chinese. The project founded by Benjamin Rusch, a former student of the Department of Japanese Studies, may not have the same amount of translation equivalents in the beginning and users may be pleased to get more choice from the *WaDokuJT* project.

6 Managing project derivates

Linking data plays an important role in managing two forks of the *WaDokuJT* project, too. An online interface of the project developed a life of its own, as format changes have rendered it incompatible with the original data. Users can add new entries or suggest corrections of existing entries online. But also on the original data side, new entries and corrections have been added.
Both sides use unique IDs – seven figure IDs for common entries and eight figure numbers for new entries for the online version at *wadoku.de*. Via these IDs entries and changes can be at least monitored, what makes common improvements possible and has most benefit for the users. Data import from *wadoku.de* also includes another edition cycle and should lead to better overall quality of the project.

7 Conclusion

In our opinion, one dictionary alone cannot satisfy all possible users' needs. An attempt to do so may turn the data rather hard to manage and very difficult to use. For additional information, we suggest the approach to link one dictionary to other dictionaries and sources. Our examples come from the daily lexicographical praxis and are added one by one, entry by entry. An automatic system that suggests links, will be helpful for the lexicographer in many cases, but we mistrust a full automatic systems at the moment.

The ultimate aim is to provide users with more information that is better tailored to their specific needs without corrupting usability and manageability of the data.

8 References

Aoyama, T., Paul, G., Schmidt-Glintzer, H., Schmithausen, L. and Wittern, C. (ed.) (2006). *Das Große Lexikon des Buddhismus – Erste Lieferung: A–Bai*. Munich: Iudicium.

Aoyama, T., Paul, G., Rotermund, H. O, Schmithausen, L., Steineck, R. C. and Wittern, C. (ed.) (2013). *Das Große Lexikon des Buddhismus – Zweite Lieferung: Bait–D*. Munich: Iudicium.

Apel, U. (2001): Ein elektronisches japanisch-deutsches Wörterbuch auf Datenbankbasis – Über das Finden von Wörterbucheinträgen im Computer-Zeitalter. In Gössmann, H. and Mrugalla, A. (eds.). *11. Deutschsprachiger Japanologentag in Trier 1999*. Bd. II. Hamburg: Lit. 627–644.

Google Books. http://books.google.com/ [10/11/2013].

Großes Japanisch-Deutsches Wörterbuch (1952). [First edition 1937, Kimura, K. (ed.)]. Tokyo: Hakuyūsha.

Handok.eu. http://handok.eu/ [founded by Benjamin Rusch; 10/11/2013]

JMdict. EDRDG. http://www.csse.monash.edu.au/~jwb/jmdict.html [10/11/2013].

NA 009 Normenausschuss Bibliotheks- und Dokumentationswesen (*NABD*) [draft for DIN 32708, Information und Dokumentation – Umschrift des Japanischen]: http://www.nabd.din.de/projekte/DIN+32708/de/147167717.html [10/11/2013]

Shogakukan Großes Deutsch-Japanisches Wörterbuch (1998): Kunimatsu K. (ed.) [revised edition]. Tokyo: Shōgakukan.

Steinhaus, W. (2010): *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch*. Berlin: Epubli. [Also: on line]. http://books.google.de/books?id=bkNWKdYgxVoC [10/11/2013].

Sōgō bukkyō daijiten (1987): Sōgō Bukkyō Daijiten Henshū Iinkai (ed.). Tokyo: Hōzōkan.

WaDokuJT: http://wadoku.eu/ or http://wadoku.de/ [10/11/2013].

Wikipedia – Die freie Enzyklopädie: http://de.wikipedia.org/wiki/Wikipedia:Hauptseite [10/11/2013]

Wikipedia [Japanese version]: http://ja.wikipedia.org/wiki/メインページ [10/11/2013]

Wörterbuch der deutschen und japanischen Gegenwartssprache. Deutsch-Japanisch (1989). Schinzinger, R., Yamamoto, A. and Nanbara, M. (ed.): Tokyo: Sanshusha.

Bilingual Word Sketches: the *translate* Button

Vít Baisa, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý
Lexical Computing Ltd, UK; Faculty of Informatics, Masaryk University, Czech Republic
{xbaisa,jak,xkovar3,pary}@fi.muni.cz, adam@lexmasterclass.com

Abstract

We present bilingual word sketches: automatic, corpus based summaries of the grammatical and collocational behaviour of a word in one language and its translation equivalent in another. We explore, with examples, various ways that this can be done, using parallel corpora, comparable corpora and bilingual dictionaries. We present the formalism for specifying equivalences between gramrels in the two languages. We show how bilingual word sketches can be useful for dictionary-making and we present additional functionality to make them more useful. We state the language pairs for which bilingual word sketches are currently available, and our plans for adding more pairs.
Keywords: word sketch; corpus lexicography; lexical computing

1 Introduction

Word sketches are one-page, automatic corpus-based accounts of a word's grammatical and collocational behaviour (Kilgarriff et al 2004). Since their introduction in 1998 they have come to be widely used in lexicography, often serving as the first port of call for a lexicographer analysing a word. Until recently, they have been monolingual. Bilingual lexicographers would like to see the word and its grammar and collocations, matched up with its translation and its grammar and collocations. In this paper we discuss how this might be done, and present the solution we have adopted and now make available within the Sketch Engine.

Two open questions are:

- how should the source word's translation be identified?
- how should the grammar and collocations be matched up?

We first describe three responses, based on different answers to the first question: *bics, bips* and *bims.* We then describe how we amalgamate all three to give a single, easy-to-use ***translate*** function, with a report as illustrated in Figure 1.

Figure 1: Bilingual word sketch for fire/Feuer. The user can click on alternative translations to see alternative sketches.

In all the approaches discussed, we start from the word sketch for one language (hereafter L1) and augment it with information from the other language (L2). Note that, here, L1 and L2 are neither 'source' and 'target' as understood by translators, nor 'mother tongue' and 'language being learnt' as in the language learning literature. They simply reflect the fact that the user (and algorithm) starts from one language and adds information from another. At some point we may develop symmetrical, direction-independent bilingual word sketches but we have not done so yet.

2 BIPs

Bips are bilingual word sketches based on **p**arallel corpora. A dictionary is not needed because the connections between the languages can be inferred, by looking to see which <L1, L2> pairs of words are frequently found in aligned chunks <L1, L2> chunks (where the chunks are usually sentences). We first count occurrences in aligned chunks for all <L1, L2> word pairs, and then use the Dice coefficient to identify candidate translations.

Where a lemmatiser is available for the language, the corpus is first lemmatised and the dictionary-induction process is applied to lemmas rather than word forms.

This provides a bilingual dictionary, with each lemma in each language having a list of candidate translations, with confidence scores. The default setting is that the ten top candidates for each lemma are retained.

Similar methods are used in GIZA++ (Och and Ney, 2000) and other tools for statistical machine translation, though usually applying to word forms rather than lemmas, and with different statistics and objectives.

Once the blingual dictionary is in place, the algorithm for creating the bip sketch is as follows:

(1) the user inputs an L1 headword.

(2) take the set of L1 collocates[1] from the L1 word sketch and translate them using the bilingual dictionary.

(3) take the top translation candidate for the L1 headword: call it the L2 headword

(4) take the set of collocates from the word sketch of the L2 headword

(5) perform an intersection between the translations-of-collocates from step 2 and the collocates-of-translation from step 4

(6) for each item in the intersection,
- is there at least one pair of aligned chunks in the parallel corpus where
  - the L1 collocation occurs in the L1 chunk, and
  - the L2 collocation occurs in the L2 chunk?
- if yes
  - present as a translation candidate for the L1 collocation,
  - illustrate with the aligned chunk in the two languages

(7) For L1 collocates with no translation candidates in the intersection, or where there were items in the intersection but there were no instances of corresponding collocations in aligned chunks, present the L1 collocate monolingually, without any candidate translations.

A bilingual sketch produced using this method, also showing the aligned chunks with both L1 and L2 collocations, is shown in Figure 2.

Figure 2: Bip word sketch for *declaration*/*déclaration*.

1 Our terminology here is that a *collocation* comprises a *headword* and a *collocate* (in a specific grammatical relation).

The 'intersection' method follows Grefenstette (1999): to find translations for compositional collocations like English *work group* (into, e. g., French) he looked up *work* and *group* in an English-French dictionary, where he found three translations for *work,* five for *group.* That gives 3 x 5 = 15 possible combinations. He then checked to see which was commonest in a French corpus, and presented that as the candidate translation. This core method is explored in much comparable corpus work (see Sharoff et al 2013 for the state of the art).

Linguee[2] provides users with some similar functionality, with aligned bilingual concordance data together with the dictionary translations of the search and some matched pairs of collocations.

2.1 In Praise of EUROPARL

A central constraint on methods using parallel corpora is the quality, genre, size and availability of parallel data for each language pair. For the 22 official EU languages[3] and corresponding 253 language pairs, we are fortunate: the EUROPARL corpora are large, contain professional quality translations, and are of a text type - European parliamentary speeches - which, while far from perfect for general-language lexicography, is far more general than the language of, for example, software manuals or patient information leaflets, two other domains where parallel data is available in bulk. Moreover the EUROPARL corpora have been prepared and made ready for language technology use (Koehn 2005). We are currently only exploring parallel-corpus methods for EU language pairs, for this reason.

3 BICs

Bics are bilingual word sketches based on **c**omparable corpora. They require a bilingual dictionary, as well as two comparable corpora, as input. Our first attempts at bilingual word sketches took a comparable-corpus approach, with dictionaries from publishers (Kilgarriff et al 2011). Our conclusion was that this left us too dependent on dictionaries from publishers, which were highly variable in availability and licence terms, not to mention format, size, quality and lexicographic approach. The approach was unviable for extending to multiple language pairs.

Now that we can build bilingual dictionaries for all EU-language pairs, these dictionaries can be used as free-standing resources for building bic sketches**.** The algorithm again uses the intersection method, and is as presented above for bips, except that step 6 is not available. Wherever an L2 headword's collocate is a translation of an L1 collocate, it is presented as a candidate translation, as shown in Figure 3.

2 http://www.linguee.com

3 Excluding Irish: irish ahs a distinct status to the other 22 languages and there is far less data available.

Figure 3: Bic word sketch for *declaration*/*déclaration*.

4 BIMs

Bims are bilingual word sketches based on **m**anual selection of headwords. In this approach the user chooses the two words, usually translation equivalents from two different languages, whose word sketches they want to compare, and the corpora to be used. The two word sketches are spliced together. This takes forward something that bilingual lexicographers have been doing since word sketches were first developed: opening two browser windows side by side, with one word sketch in each.

A bilingual sketch for English *brown* and Portuguese *marrom* is shown in Figure 4.

Figure 4: Bim word sketch for *marrom*/*brown*.

4.1 Alignment of grammatical relations

The lexicographer would like to see collocations and their translation equivalents aligned. This is possible to some extent, and is attempted in bip and bic sketches, but is difficult and error-prone. Also the lexicographer would often like more control, and finds it straightforward to match, eg, brown *leather* and *couro* marrom where they are in columns next to each other. So in bim sketches we set ourselves the more limited ambition of matching up columns, so that collocates for corresponding grammatical relations are shown next to each other. (In word sketches, 'grammatical relations' or 'gramrels' are the relations such as *object*, *object_of*, *modifier* that are specified in the sketch grammar and are showed at the heads of the columns in a word sketch.)

Where the gramrels have the same names for two languages, this is trivial: we simply show same-name gramrels next to each other. However sketch grammars for different languages are usually prepared independently of each other; the grammar of different languages is different, underlying part-of-speech taggers may use different conceptualisations and word classes; and gramrels will often be given names in the language that the word sketch is for: the French equivalent of *object_of* is called *objet_de*. So matching gramrels with identical names is not the standard case. We need a mapping between the gramrels of the two languages.

A mapping for each set of names to some 'master' set is preferable to a different mapping for each pair. In our current setting, we use the English names as the master. Both 1:1 and m:n mappings are possible.

The mappings are defined in the sketch grammar using a newly introduced processing directive *UNIMAP. The following definition from a French sketch grammar

```
*DUAL
=objet/objet_de
*UNIMAP object/object_of
...
```

says that *objet* should be joined with *object* and *objet_de* should be joined with *object_of* (or the gramrels paired with English *object* and *object_of* in other languages). The algorithm for finding a target language (TL) gramrel to display next to a source language gramrel X is:

- if there is one or more TL gramrel with a UNIMAP value matching the UNIMAP value of X, select that one/them
- else if there is a TL gramrel of the same name, select that one
- else, nothing is aligned with X.

Left-over, unaligned TL gramrels are shown after SL and aligned gramrels.

4.2 An inline form for finding missing items

For a user, some matching pairs are immediately evident from the bim sketch (*brown leather, couro marrom*) but others are not. We do not find anything equivalent to *brown rice* on the Portuguese side of Figure 5. The user then wants to know "how do you say brown rice in Portuguese?" To meet this need, a new function has been added: the user may click on *rice* to reveal a text-input box where they can input the missing target-language equivalent (here, *arroz*). A new bim word sketch appears, as illustrated in the same Figure 5. This feature helps the user find the missing translation: here, *arroz integral* for *brown rice*.

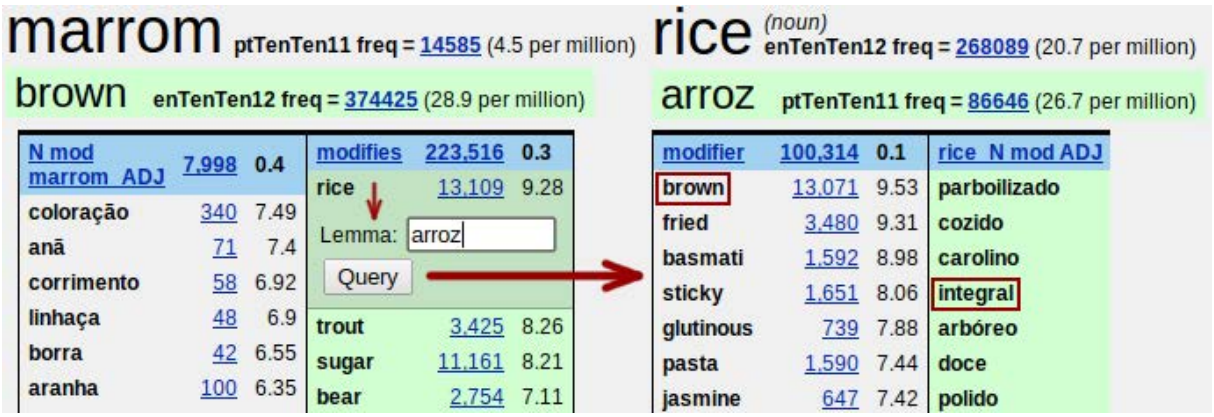


Figure 5: "Finding the missing translation" functionality.

5 Observations

Bics, bips and bims were developed in 2013.[4] In the course of presenting and beginning to use them, we made several observations:

- The bilingual dictionaries created from EUROPARL were of good quality. Most of the time, the top candidate translation was valid.
- For most collocations on the L1 word sketch, we did not find any strong candidates for L2 translation equivalents. The *declaration* examples above were selected because, there, the algorithm did find a translation candidate for many of the L1 collocations. More often, there were very few translations offered. This affected both bics and bips.
- Bims often worked well, but the access method did not: users were required to select, first, the L2; then, the L2 corpus from which the L2 word sketch should be drawn, and then, the L2 lemma. Unless they made all the right choices, they were going to be disappointed.

4 They were presented at the e-Lexicography conference in Tallinn in October 2013.

6 The *translate* button

Most users just want to select the target language, and do not want to think about 'which corpus'. The system developers are the people who know which corpus has the best word sketches for a language, so they should make that choice.

Many users would also rather not have to think of, and input, the L2 lemma. For all EU language pairs, the EUROPARL dictionary offers candidate translations, so here again, more can be done by the system leaving less work for the user. The user can be offered a BIM sketch in which the L2 word is the top candidate from the EUROPARL bilingual dictionary (with second and other candidates also offered.)

With these choices made, we can add a 'translate' button to the word sketch. The 'translate' button gives the user a choice of languages. The languages that the user can choose between are all of those where there is:

- a bilingual dictionary between L1 and the language
- high-quality word sketches for the language
- as at April 2014, there are twenty such languages
- a UNIMAP mapping

Then, when the user has selected the L2, they see the bilingual sketch directly, as in Figures 1 and 6:

amore *(noun)* itTenTen10 freq = 634049 (206.1 per million)

Liebe *(noun)* deTenTen10 freq = 294973 (103.7 per million)

n_modifier	3,656	0.6	modifier	281,129	0.5	preN_V	120,341	2.0	VerbY+Subst...	24,736	0.5
sponsale	129	9.84	wahr	2,537	6.76	giurare	430	6.71	gestehen	750	7.54
lesbico	116	9.39	bedingungslos	583	5.93	dichiarare	1,080	6.7	erwidern	139	6.1
oblativo	41	8.46	ewig	870	5.86	donare	573	6.66	besingen	48	5.38
karmico	34	8.0	dein	3,017	5.71	chiamare	1,754	6.62	entdecken	836	5.03
leopardiano	20	7.15	unerfüllt	465	5.7	Un	1,160	6.49	beteuern	43	4.87
lesbo	16	6.7	unglücklich	539	5.69	cantare	601	6.49	schwören	59	4.45
lesbici	10	6.29	innig	468	5.65	nascere	1,388	6.29	schenken	194	4.27
fato	31	6.21	verboten	490	5.58	provare	901	6.27	entgegenbringen	24	4.22
anoressico	9	5.87	romantisch	552	5.45	vivere	1,643	6.21	wünschen	634	4.22
cannibale	13	5.85	heimlich	476	5.4	confessare	311	6.18	symbolisieren	45	4.14
conta	23	5.7	groß	12,073	5.38	ricambiare	277	6.13	empfinden	200	4.12
cinefilo	9	5.18	leidenschaftlich	423	5.35	sbocciare	260	6.08	predigen	38	4.09
court	12	5.05	viel	6,631	5.33	trasmettere	584	6.05	grüßen	85	4.08

Figure 6: Bilingual word sketch for *amore*/*Liebe*.

The functionality is currently available for all combinations of English, French, German, Italian and Spanish.

7 Current and future work

- **Aligning collocates within BIMs**

  A feature of bics and bips that bims were lacking, was the alignment of translation-equivalent collocations. Although, often, no alignments were found, where an alignment was found, it was usually valid and helpful. So, we can further enrich bim-style sketches, by re-ordering the collocates in the L2 word sketch table so that matched L1, L2 pairs are next to each other.

This is currently in progress.

- **More languages**

  For monolingual or bilingual word sketches to work well, there are a number of prerequisites; first a very large corpus, then processing tools including a tokeniser, lemmatiser, part-of-speech tagger and sketch grammar. All components are currently in place for all major world languages, all EU languages, and a number of others; languages where we intend to get all components working well in the near future include Icelandic, Malay, Bahasa Indonesia and Burmese.[5]

- **More bilingual dictionaries and language pairs**

  We are looking into parallel resources for other language pairs including English and the non-EU major world languages, in particular Arabic-English, Chinese-English, Japanese-English, Russian-English.

- **Evaluating EUROPARL-based bilingual dictionaries**

  For Czech-English we are currently assessing our induced dictionary by comparison with a publisher's dictionary. We shall also compare the Dice algorithm with the Giza++ algorithm for dictionary induction. We shall then extend the evaluation to other language pairs.

8 References

Grefenstette, G. 1(999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *Proceedings of Aslib Conference on Translating and the Computer 21.* London.

Kilgarriff, A., Avinesh, P. V. S., & Pomikálek, J. (2011). Comparable Corpora BootCaT. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex*, 122-128.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proc. Euralex.* Lorient, France.

Koehn P. (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. *Proc. Machine Translation Summit*.

Och, F. & Ney, H. (2000).. Improved Statistical Alignment Models. *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hong Kong, China.

Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P., editors (2013). *Building and Using Comparable Corpora.* Springer.

5 In most cases the work has been a collaboration with linguists of the language in question.

Creating a Bilingual Italian-English Dictionary of Collocations

Barbara Berti, Laura Pinnavaia
Università degli Studi dell'Insubria, Università Statale di Milano
barbara.berti@uninsubria.it, laura.pinnavaia@unimi.it

Abstract

Collocations are among the most challenging issues that learners of a foreign language have to face. The main obstacle for Italian learners of English is encountered in active tasks, and especially, when translating or writing essays in English. The encoding of collocations in fact often results in the selection of wrong terms. There is no doubt that the production and translation of Italian collocations would be facilitated were students to have at their disposal a bilingual Italian-English Dictionary of Collocations. The aim of this work is to endorse the importance of the existence of such a tool and to show the beginnings of its compilation.

Keywords: collocations; bilingual lexicography; corpus linguistics.

> It is not so much the words of English nor the grammar of English that makes English difficult ... The vague and undefined obstacle to progress ... consists for the most part in the existence of so many odd comings-together-of-words. (Cowie 1999: 52-53)

1 Introduction

At the intersection between syntax and semantics, collocations are among the most challenging issues that learners of a foreign language have to face. Not only are they massively present in languages (Hoey 2005), but they are also (or at least appear to be) arbitrary and highly specific for a given language. This poses problems in second language acquisition. In fact, at some stage of the learning process, learners will inevitably have to deal with the way in which words combine with each other and the restrictions that they are subject to. Moreover, their first language will be likely to interfere to some extent with the production of natural-sounding word combinations in the second language and will cause them to produce incorrect *verbum pro verbo* translations (e.g. *make a photo** as a calque of *fare una foto*).

The main obstacle for learners of English is, in fact, encountered in production tasks, and, especially, when translating from Italian, or writing essays in English. In reception tasks, the meaning of English collocations can often be derived quite easily, especially if the context in which they occur is suf-

ficiently clear. For example, while in isolation the meaning of the collocation *to explode a myth* might be opaque to a learner of English, in an appropriate context, such as *scientists wanted to explode the myth of a monster living in the lake*, it becomes apparent. However, to produce the collocation from scratch would undoubtedly be a problem for the learner. It is highly likely that he/she would not select the correct terms, especially if, as in the case of Italian, the collocate of the noun *myth* is semantically unrelated to the verb *explode* (*sfatare un mito*). It thus becomes apparent that learners need to have at their disposal a resource that will help them to overcome such problems.

Nowadays, English monolingual lexicography does tackle the issue of collocations quite extensively, and especially dictionaries intended for learners of English; yet, there is evidence that Italian learners of English feel more at ease using bilingual tools rather than monolingual ones (Baxter 1980; Bensoussan, Sim and Weis 1984; Atkins 1985; Atkins and Knowles 1990; MacFarquhar and Richards 1983; Piotrowski 1989; Rundell 1999; Scholfield 1999). Unfortunately, however, Italian-English bilingual lexicographical resources poorly document this phenomenon from both a quantitative and a qualitative point of view (Berti 2010, 2012). That is why, in addition to implementing bilingual dictionaries with more and more carefully selected word combinations, a bilingual Italian-English dictionary of collocations would be very useful for Italian learners of English.

Successive to a theoretically-oriented study regarding the creation of a corpus-based Italian-English dictionary of collocations (Berti & Pinnavaia 2012), this work comes as a practical presentation of what selecting collocates entails, followed by two sample entries. The dictionary has as its principal target Italian learners of English, as well as professional translators. It will be unidirectional, presenting Italian collocations with their equivalents in the English language. The aim is not just to provide viable translations of Italian collocations, but also to offer users the most suitable English equivalents based upon frequency of use and situational appropriateness. This can only be achieved if the equivalents are drawn from linguistic corpora of English and carefully evaluated by the compilers, whose linguistic expertise must be bilingual.

2 Methodology

Nowadays, the use of corpora is of fundamental importance in lexicography (Atkins 1994), offering a more "objective" perspective on language. Over the years, in fact, it has become progressively clearer that resources based on the lexicographers' *Sprachgefühl* are subjected to a number of limitations and that they cannot be regarded as fully representative of the current state of a language. Diastratic and diatopic variations, together with a speaker's idiolect, exert an influence on the type of material that is included or excluded in and from a lexicographic tool, and even the work of a composite team of experts cannot guarantee an objective treatment of the language, or the selection of appropriate material.

In particular, as regards collocations, research shows that acceptability is extremely subjective and that the individuals' knowledge of word combinations in their language might vary to a great extent (Berti, *forthcoming*; Nesselhauf 2003). Acceptability depends, firstly, on semantics and, secondly, and most importantly, on usage. Indeed, it is usage that ultimately provides us with the intuition we need in order to discern between acceptable and unacceptable combinations in our mother tongue. Yet, given that the linguistic experience of each individual is absolutely unique and that the factors that have an influence on it are numerous, there will be a large grey area of disagreement in the judgement of word combinations. For this reason, it is nowadays essential to query corpora for the selection of the lexical material to be used for lexicographical purposes.

Since they are the parts of speech that occur more often in collocational patterns and are the most needed by students, it is nouns and verbs that we decided to select. This choice also seemed reasonable in virtue of the manner in which our thoughts are shaped. A speaker, unaware of the adjective + noun combination broad daylight, will more naturally think of the noun daylight before its qualifier. Similarly, the retrieval of any verb and adverb combination will normally imply the consultation of the verb before that of the adverb. This seems to be the rationale behind the organization of most dictionaries of collocations, among others, Rundell (2010) and the Oxford Dictionary of Collocations (see ODC 2002). The idea is thus to create an onomasiological dictionary that relies on the well-established distinction between base and collocate.

For the present study we took the Italian noun *odore* (a smell) and verb *pagare* (to pay) as our working examples retrieving the Italian collocations from the *Dizionario delle collocazioni* (Tiberii 2012). The decision to collect our sources from an already existing dictionary of collocations as opposed to extracting collocations from an Italian corpus was taken on recognizing the efficiency of this tool and after having ascertained the overall paucity of general Italian corpora compared to English.

We proceeded to find the collocates of the English equivalents of *odore* and *pagare*, *smell* and *pay* in the British National Corpus (BNC). Each Italian collocate was then matched with its English semantic equivalent. The explanation of the procedure for and the problems encountered with the pairing off of items ensues.

2.1 The Noun *Odore*

In Tiberii (2012), the following collocates for the noun *odore* were found:

> **odore** *nm* accattivante, acre, acuto, aspro, buono, caratteristico, delicate, disgustoso, distintivo, fetido, forte, fresco, gradevole, inconfondibile, inebriante, intense, invitante, leggero, nauseabondo, nauseante, opprimente, particolare, penetrante pungente, ripugnante, rivoltante, sgradevole, sottile, stomachevole, strano, tenue, terribile, vago.

Considering that the Italian collocates for *odore* are all adjectives, we searched for adjective-noun collocational patterns in the BNC, obtaining the following one hundred collocates for the noun *smell*, listed in order of decreasing frequency:

- sweet
- strong
- pungent
- strange
- musty
- rich
- faint
- sour
- bad
- unpleasant
- acrid
- stale
- warm
- delicious
- fresh
- heavy
- horrible
- lovely
- different
- familiar
- good
- nice
- able
- clean
- damp
- distinctive
- sickly
- dank
- funny
- other
- overpowering
- wonderful
- awful
- hot
- new
- old
- pleasant
- aware
- earthy
- lingering
- peculiar
- thick
- close
- distinct
- metallic
- oily
- foul
- full
- masculine
- musky
- open
- particular
- sharp
- spicy
- chemical
- cold
- cloying
- comfortable
- dirty
- fishy
- fragrant
- free
- great
- green
- milky
- nasty
- putrid
- real
- slight
- smoky
- soft
- tangy
- usual
- very
- whole
- curious
- dangerous
- comforting
- clinical
- characteristic
- burning
- blue
- better
- bitter-sweet
- acute
- dry
- evil
- fine
- homely
- keen
- little
- noxious
- powerful
- rancid
- rotten
- savoury
- sick
- stronger
- supposed
- sure

The next step was to match each Italian collocate from Tiberii (2012) with an equivalent English one from the above list. For example, we paired off *odore accattivante* with *captivating smell*. On finding Italian collocations that correspond to two or more English ones, as in the case of *odore gradevole* which is equivalent to both *sweet* and *pleasant smell*, none were discarded and all were kept. This was done even when one collocation was found to be more frequent than another (i.e. *sweet smell*), out of scientific rigour, on the one hand, and in order to make up for the scant number of collocations provided in bilingual dictionaries, which were naturally also consulted, on the other. When, moreover, it was noticed that some of the collocates in Italian are synonymous, these were grouped together; for instance, *acre* and *aspro*, can both be rendered as *acrid, sour* or *sharp* in English, just as *leggero, tenue, sottile, vago* can be translated as *slight* and *faint*. When the close examination of the list of English collocations revealed the absence of a number of Italian collocations in the source text, it was deemed opportune to introduce them, especially if the English equivalents are frequently occurring collocations. For example, it seemed important to include the collocations *cattivo odore* and *odore stucchevole*, missing from Tiberii (2012), aware of the fact that the collocations *unpleasant/musty smell* and *heavy/cloying smell* are commonly used in English. Lastly, when equivalents for Italian collocates were not found in the BNC, we consulted WebCorp, which owing to its larger size, often managed to provide suitable equivalent adjectives, such as *captivating, delicate, intoxicating* and *penetrating smell* for *accattivante, delicato, inebriante* and *penetrante*.

2.2 The Verb *Pagare*

Under the lemma *pagare*, Tiberii (2012) lists the following adverbial collocates:

> **pagare** v. abbondantemente, anticipatamente, caro, comodamente, completamente, generosamente, immeditamente, integralmente, interamente, obbligatoriamente, pesantemente, profumatamente, prontamente, provvisoriamente, puntualmente, regolarmente, tempestivamente, volontariamente.

While the collocates we retrieved in the BNC for *pay* amount to forty-three, only eight have been reported here. This is because, unlike the list of collocates for *smell*, the list for *pay* included many that were not pertinent. It was necessary, in fact, to shortlist the relevant ones by examining the concordances. During this process, we also decided to exclude collocates with a single occurrence, following Sinclair's tenet that a single occurrence does not represent a "settled pattern in the language" (2003: 15).

- promptly
- dearly
- hereby
- handsomely
- proportionately
- compulsorily

- upfront
- punctually

As earlier, the Italian and English items were matched. For example, pagare obbligatoriamente was matched with pay compulsorily. Similarly, synonymous collocations were grouped together, as in pagare abbondantemente/generosamente/profumatamente, which were all associated to pay handsomely. However, the pairing off of collocates proved to be more problematic than it was for odore/smell. Firstly, because there are very few correspondences between the Italian and English collocates; secondly, because it is not always easy to understand the real meaning of certain Italian collocations on account of the absence of contextual data in Tiberii (2012). For example, pagare anticipatamente/comodamente/integralmente/interamente/puntualmente/regolarmente/volontariamente do not have a corresponding collocate in the BNC. For this reason, in order to provide an equivalent, not only did we have to consult various bilingual dictionaries along with WebCorp, but we also had to rely on our own knowledge of the language. At times this resulted in equivalent collocations (e.g. pagare comodamente = pay easily), at other times in lexical sequences that do not fit the verb + adverb pattern (e.g. pagare anticipatamente = to pay in advance), at other times in single words (e.g. pagare anticipatamente = to prepay). The results we obtained can be observed in the sample provided in the following section. Not having examples of usage to refer to made the translation of the items pagare completamente and pagare provvisoriamente particularly problematic: while the former sounds like a free combination, the latter is clearly semantically ambiguous. The way in which the collocates of pagare/pay have been rendered in English can be observed in the sample provided in the following section.

3 The Sample Entries

Here follow the sample entries for the noun *odore* and verb *pagare*.

odore - smell n.

ADJ. + NOUN
accattivante – captivating
acre/aspro – acrid/sour/sharp
acuto/pungente – pungent
buono – good/nice/delicious
caratteristico – characteristic
cattivo/sgradevole – unpleasant/musty
delicato – delicate
disgustoso/ripugnante/rivoltante/terribile – bad/ horrible/nasty/awful
distintivo – distinctive
fetido – foul/putrid
forte – strong/overpowering
fresco – fresh
gradevole – sweet/pleasant

inconfondibile – distinct
inebriante – intoxicating
intenso – powerful/keen
invitante – warm
leggero/tenue/sottile/vago – slight/faint
nauseabondo/nauseante/stomachevole – sickly
opprimente – overpowering
particolare – curious/peculiar/particular
penetrante – penetrating
persistente – lingering
strano – strange
stucchevole – heavy/cloying

pagare - pay v.

ADV. + VERB
Abbondantemente/generosamente/profumatamente – handsomely
anticipatamente – in advance/prepay (v.)
caro/pesantemente – dearly
comodamente – easily
immediatamente/prontamente/tempestivamente – promptly
integralmente/interamente – in full
obbligatoriamente – compulsorily
puntualmente – punctually
regolarmente – regularly
volontariamente – voluntarily

As the examples show, the entries provide a list of the Italian collocates, documented for the noun *odore* and for the verb *pagare*, along with their respective adjectival and adverbial equivalents for the English noun *smell* and verb *pay*. The Italian collocates appear listed in alphabetical order, and sometimes include more than one exemplar when they have similar meanings (e.g. *odore disgustoso/ripugnante/rivoltante/terribile; pagare immediatamente/prontamente/tempestivamente*). Next to them appear the equivalent English collocates, retrieved from the data in the BNC and WebCorp. As can be seen, the correspondences can be of various types. It is possible to have collocations in Italian and English that have the same number of collocates (e.g. *strano odore = a strange smell*; *cattivo/sgradevole odore = unpleasant/musty smell; pagare puntualmente = to pay punctually*); sometimes the Italian language has more (*nauseante, nauseabondo, stomachevole odore = sickly smell; pagare immediatamente/prontamente/tempestivamente = to pay promptly*); at other times, fewer (e.g. *odore buono = good/nice/delicious smell; pagare anticipatamente = to pay in advance/to prepay*). In all cases, it is our intention that dictionary-users have all the most frequent and most appropriate equivalent English collocations for the Italian ones.

4 Conclusions

While this work is still in progress and has still not fully solved all the problems involved in retrieving and selecting equivalent English collocates for Italian ones, the objective of compiling a bilingual dictionary of Italian and English collocations is, in our view, invaluable for the quantitative and

qualitative advantages it can offer next to general bilingual and monolingual dictionaries. Compared to a normal Italian-English bilingual dictionary, such as *Il Ragazzini2011*, a bilingual collocations dictionary can, first of all, provide much more information. While certainly being an efficient tool, *Il Ragazzini2011*, like any other bilingual dictionary, has to include all aspects of the Italian language. It cannot, evidently, devote as much space to collocations. In fact, it only provides the following English collocations for *smell* under the noun *odore*: a *good smell, a pleasant smell, a nice smell, a bad smell, a nasty smell* and an *offensive smell*. Despite offering a wide range of noun + verb collocations under the lemma *pagare*, with regards to the lexical pattern adverb + *pagare*, the dictionary only includes *pagare salato* and *pagare profumatamente,* both translated as *to pay dearly*. Clearly, bilingual dictionaries need to make a selection among the various collocations to include, whereas a dictionary of collocations can be much more exhaustive.

Of course, it could be argued that learners or translators, needing to find English collocations, can directly consult the monolingual tools already available. There is no doubt that monolingual English dictionaries of collocations are excellently compiled: once the user is aware of the English base, then he/she can consult such dictionaries to find collocates. It is also true, however, that this task in a monolingual dictionary takes longer and requires greater linguistic skills than it would do in a bilingual collocations dictionary. As seen above, the Italian-English bilingual collocations dictionary could more carefully account for meaning nuances in English, based on the Italian collocates. In finding the English collocations, *an acrid smell, a sour smell, a sharp smell,* and *a pungent smell*, an Italian learner or translator might be tempted to think that they are all equivalents. We have seen, however, that while the first three can be considered synonyms, the last, *a pungent smell*, means something a little different as its Italian equivalents endorse. The use of a bilingual collocations dictionary should, thus secondly, save the dictionary-user time in trying to individuate what collocation to choose, resulting in a more qualitative and effortless task. The combination of a fine lexicographic and linguistic analysis of Italian and English words will hopefully result in a bilingual reference work that will be quantitatively and qualitatively valid.

5 References

Atkins, B. T. S. (1985). Monolingual and bilingual learners' dictionaries: A comparison. In R. Ilson (ed.), Dictionaries, Lexicography and Language Learning. Oxford: Pergamon Press and British Council, pp. 15-24.

Atkins, B.T.S. (1994) A corpus-based dictionary. In *Oxford-Hachette English- French Dictionary* (Introductory section). Oxford: Oxford University Press: xix – xxxii.

Atkins, B.T., Knowles, F.E. (1990). Interim report on the EURALEX/AILA research project into dictionary use. In I. Magay, J. Zigany (eds.), *Budalex 88 Proceedings.* Budapest, Akademiai Kiado, pp. 381-392.

Baxter, J. (1980). The dictionary and vocabulary behavior: a single word or a handful? In *TESOL Quarterly* XIV, 3, pp. 325-336.

Bensoussan, M., Sim, D. & Weiss, R. (1984). The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. In *Reading in a Foreign Language,* 2, pp. 262-276.

Berti, B. (2010). The treatment of lexical collocations of six adjectives related to feelings in a sample of bilingual dictionaries English-Italian. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of XIV Euralex International Congress, Leeuwarden, 6-10 July 2010,* Fryske Akademy, Leeuwarden. [contribution on CD-rom]

Berti, B. (2012), Le collocazioni lessicali nel dizionario bilingue. Uno studio su tre dizionari italiano-inglese. In G. Gobber *Usi dei vocabolari nell'apprendimento delle lingue,* Milano: Vita e Pensiero, pp. 5-25.

Berti, B. (forthcoming). 'This sounds odd to me'. Issues of collocational acceptability in a bilingual English-Italian dictionary". In *Proceedings of the conference Norma e uso nella lessicografia bilingue*, *18-20 October 2012*, Ragusa.

Berti, B., Pinnavaia, L. (2012). Towards a Corpus-driven Italian-English Dictionary of Collocations. In R. Facchinetti (ed.) *English Dictionaries as Cultural Mines*, Newcastle-Upon-Tyne: Cambridge Scholars Publishing, pp. 201-222.

Cowie, A. (1999). *English Dictionaries for Foreign Learners: A History*. Oxford: Oxford University Press.

Facchinetti (ed.) English Dictionaries as Cultural Mines, Cambridge Scholars Publishing, Newcastle-Upon-Tyne: 201-222.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

MacFarquhar, P. D., Jack C. R. (1983). On dictionaries and definitions. *RELC Journal*, 14, pp. 111-124.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. In *Applied Linguistics*. 24(2): pp. 223-242.

Oxford Collocations Dictionary for Students of English (2012). Oxford: Oxford University Press.

Piotrowski, T. (1989). Monolingual and bilingual dictionaries: Fundamental differences. In Makhan L. Tickoo (ed.), *Learner's Dictionaries: State of the Art*. Singapore: Seamo Relc, pp. 72-83.

Il Ragazzini 2011. Dizionario inglese-italiano, italiano-inglese, Bologna: Zanichelli.

Rundell, M. (1999). Dictionary use in production. In *International Journal of Lexicography*, 12, (1): pp. 35-54.

Rundell, M. (2010). *Macmillan Collocations Dictionary for Learners of English*, Oxford: Macmillan Publishers Limited.

Scholfield, P. J. (1999). Dictionary use in reception. In *International Journal of Lexicography* 12, (1): pp.13-35.

Sinclair, J. (2003). *Reading concordances. An introduction.* Harlow: Pearson Education Limited.

Tiberii, P. (2012). *Dizionario delle Collocazioni. La Combinazione delle Parole in Italiano*, Bologna: Zanichelli.

Websites

BNC: *British National Corpus*, accessible at http://www.natcorp.ox.ac.uk.

WebCorp, accessible at http://www.webcorp.org.uk/live/.

La valencia del adjetivo en diccionarios bilingües alemán-español-alemán

Andreu Castell, Natàlia Català, María Bargalló
Universitat Rovira i Virgili – Tarragona (España)
andreu.castell.@urv.cat, natalia.catala@urv.cat, maria@bargallo@urv.cat

Resumen

El verbo constituye el elemento fundamental en la teoría de valencias, por lo que es normal que se le haya prestado siempre una especial atención. Menos atención se le ha dedicado hasta ahora a la valencia del nombre y mucho menos aún a la del adjetivo. El estudio que se presenta forma parte de la fase inicial de un proyecto que se está desarrollando con el fin de establecer las bases teóricas y metodológicas para la elaboración de un diccionario sintáctico del adjetivo en alemán y en español. Se analiza la información sintáctica que ofrecen diversos diccionarios bilingües alemán-español-alemán sobre la complementación de ocho adjetivos que, en principio, poseen los mismos complementos argumentales en ambas lenguas. Se estudia, en concreto, qué información ofrecen respecto de una serie de criterios que desde el punto de vista de la gramática de valencias se deberían considerar fundamentales, así como hasta qué punto la información ofrecida, independientemente de los presupuestos de la teoría de valencias, resulta realmente útil al usuario. También se analiza el grado de coherencia interna que presentan los diccionarios estudiados en las diversas entradas.
Palabras clave: lexicografía; adjetivos; valencia del adjetivo

1 Introducción

La valencia del adjetivo se trata de forma más o menos detallada en casi todas las gramáticas de la lengua alemana aparecidas en la propia Alemania a partir de los años setenta del siglo pasado: desde Helbig/Buscha (1972: 283-285) y Duden (1973: 229-230) hasta Duden (2009: 361, 927-933), pasando por Eisenberg (1986: 93-95) y Engel (1988: 592-600 y 2004: 353-361). Este último llega a adjudicarle al elemento que nos ocupa ni más ni menos que 16 complementos argumentales distintos. Es más, para la lengua alemana incluso existe, aunque sea el único hasta ahora, un diccionario de valencias del adjetivo (Sommerfeldt/Schreiber 1974).

Respecto del español, a pesar del artículo publicado por Bosque (1983), la atención que se ha prestado a la valencia del adjetivo ha sido, como constatan Garriga/Bargalló (2000: 602), mucho menor: "En muchas ocasiones las gramáticas se limitan a comentar esta cuestión y ofrecen unos pocos ejemplos que no cubren todas las posibilidades, entre ellas, especialmente, la que aquí nos ocupa: el régimen del adjetivo". No es hasta la aparición de las dos últimas grandes gramáticas de la lengua española (Bosque/

Demonte 1999; Real Academia Española 2009) que ha variado el panorama. Que un adjetivo posee complementos ya se insinúa en gramáticas anteriores: "La cualidad que expresa un adjetivo con respecto al sustantivo a que se refiere puede limitarse o concretarse por medio de un sustantivo precedido de preposición; p.e. *dócil a la advertencia* [...]" (Real Academia Española 1973: 414). Incluso se encuentran, aunque muy excepcionalmente, referencias al respecto en libros de texto para la enseñanza secundaria: "Al igual que el sustantivo, el adjetivo puede recibir complementación mediante un sintagma preposicional. Este sintagma se denomina complemento del adjetivo" (Llácer et al. 1996: 135). Sin embargo, que un adjetivo pueda poseer un complemento indirecto tan solo llega a afirmarse o a insinuarse en las dos gramáticas más recientes: "Concluimos, pues, que tanto sustantivos como adjetivos pueden aparecer con un complemento indirecto" (Bosque/Demonte 1999: 1563); [...] se plantea inmediatamente la cuestión de si los adjetivos y los sustantivos pueden tener complemento indirecto" (Real Academia Española 2009: 2659).

Se constata, en general, como señala Castell (2009: 183 y ss.), una recepción mucho mayor de la teoría de valencias en las gramáticas de la lengua alemana, por lo que podría suponerse que ello debería reflejarse también en los correspondientes diccionarios monolingües. Sin embargo, dejando de lado alguna excepción en el caso del alemán (como Langenscheidt 2003), lo cierto es que dicha suposición no resulta del todo correcta. En un trabajo todavía inédito, realizado como parte del proyecto en el que se enmarca el presente estudio, se constata que por ejemplo el diccionario de Moliner (2008) y el *Diccionario Salamanca* para el español y el *Deutsches Universalwörterbuch* (Duden 2008) y el diccionario *Deutsch als Fremdsprache on line* de Pons para el alemán presentan carencias muy similares en cuanto a la información que ofrecen sobre la complementación del adjetivo.

En el presente estudio se analiza qué información sintáctica ofrecen respecto del adjetivo diversos diccionarios bilingües alemán-español-alemán. En principio cabe suponer que los usuarios de dichos diccionarios son fundamentalmente hablantes de la lengua de partida que, aún poseyendo un nivel muy alto de la misma, desconocen del todo o en gran medida cómo funciona la lengua de destino. Por lo tanto, este tipo de diccionarios debería ofrecer una información aún más detallada y específica que la que ofrecen los diccionarios monolingües.

Se analizan a este respecto diccionarios bilingües, tanto en forma impresa como electrónica o accesibles online: *Diccionario de las lenguas española y alemana* (Slaby/Grossmann 1993 y Slaby/Grossmann/Illig 1991) (SLG), *Hueber Wörterbuch. Diccionario Alemán-Español/Español-Alemán* (2007) (HUE), *Langenscheidt e-Taschenwörterbuch Spanisch-Deusch-Spanisch* (Langenscheidt 2002) (LNG), *Pons Diccionario español-alemán-español online* (PON), *Larousse Diccionario español-alemán-español online* (LAR), *LEO Spanisch-Deutsch-Spanisch online* (LEO), *Reverso Diccionario Collins español-alemán-español online* (REV), *Beolingus online* (BEO) y *DIX online* (DIX).

Los diccionarios citados se han analizado en cuanto a la información sintáctica que ofrecen respecto de ocho adjetivos que, en principio, poseen los mismos complementos argumentales en ambas lenguas:

1) Adjetivos con un complemento argumental preposicional que puede ser realizado mediante una oración: *fähig/capaz*, con el significado de "estar dispuesto o atreverse a algo" e *interessiert/interesado*, con el significado de "mostrar interés por alguna cosa o alguien":

(1) Er ist *zu allem fähig*.

(2) Es *capaz de todo*.

(3) Ich bin *an dem Projekt interessiert*.

(4) Estoy *interesado en el proyecto*.

El complemento argumental preposicional de *fähig/capaz* puede ser realizado también por una oración de infinitivo, el de *interessiert/interesado* admite su realización tanto mediante una oración de infinitivo como mediante una oración subordinada finita con *dass/que*:

(5) Er war nicht *fähig, seine Meinung zu verteidigen*.

(6) No fue *capaz de defender su opinión*.

(7) Ich bin *daran interessiert, an dem Projekt teilzunehmen*.

(8) Estoy *interesado en participar en el proyecto*.

(9) Ich bin sehr *daran interessiert, dass Sie das Projekt befürworten*.

(10) Estoy muy *interesado en que recomiende usted el proyecto*.

2) Adjetivos con un complemento argumental preposicional que no puede ser realizado mediante una oración: *reich/rico*, con el significado de "tener mucho de lo que se expresa":

(11) Das Land ist *reich an Energiequellen*.

(12) El país es *rico en recursos energéticos*.

3) Adjetivos con un complemento argumental direccional o local que no puede ser realizado mediante una oración: *gebürtig/natural*, con el significado de "nacido en un lugar" y *wohnhaft/residente*, con el significado de "habitar en un lugar":

(13) Meine Nachbarin ist *aus Ungarn gebürtig*.

(14) Mi vecina es *natural de Hungría*.

(15) Beide waren *in Madrid wohnhaft*.

(16) Ambos eran *residentes en Madrid*.

4) Adjetivos con un complemento argumental dativo y otro preposicional, con los que el último puede ser realizado mediante una oración: *dankbar/agradecido*, con el significado de "dar muestras de gratitud por algún favor recibido":

(17) Wir sind *Ihnen für Ihre Unterstützung dankbar*.

(18) *Le* estamos *agradecidos por su apoyo*.

En alemán, el complemento argumental prepositivo puede ser realizado mediante una oración subordinada finita con *dass*, mientras que en español resulta más habitual una oración de infinitivo:

(19) Wir sind Ihnen *dankbar dafür, dass Sie uns unterstützt haben*.

(20) Le estamos muy *agradecidos por habernos apoyado*.

No está claro que el complemento prepositivo pueda ser realizado en español mediante una oración subordinada finita con *por que*. En el *Corpus de referencia del español actual* (Real Academia Española) se

encuentran numerosos ejemplos de realización mediante oraciones subordinadas de infinitivo, pero también se encuentran ejemplos como los siguientes:
(21) Estuve sonriendo, asombrado y *agradecido porque fuera tan fácil* [...].
(22) Le estaba muy *agradecido porque le hacía las cuentas* [...].
En ellos se tendería a analizar la oración subordinada como un complemento adjunto de causa. Sin embargo, debe tenerse en cuenta que ambas expresiones equivalen a:
(23) Estuve sonriendo, asombrado y *agradecido por el hecho de que fuera tan fácil*.
(24) Le estaba muy agradecido *por el hecho de que hiciera las cuentas*.
Aquí nos hallamos, sin duda alguna, ante un complemento argumental prepositivo, por lo que, en principio, no deberían descartarse como agramaticales expresiones como las siguientes (cfr. Brucart/Gallego (2009: 166-167) y Pavón (2012: §2.1)):
(25) ?Estuve sonriendo, asombrado y *agradecido por que fuera tan fácil*.
(26) ?Le estaba muy *agradecido por que le hiciera las cuentas*.

5) Adjetivos con un complemento argumental dativo que no puede ser realizado mediante una oración: *treu/fiel*, con el significado de "no faltar a sus ideas o a su compromiso con su pareja" y *untreu/infiel* como antónimos de los anteriores:
(27) Er ist *seinen Idealen* immer *treu* gewesen.
(28) Siempre ha sido *fiel a sus ideales*.
(29) Ich bin *ihr* immer *treu* gewesen.
(30) Siempre *le* he sido *fiel*.

2 ¿Qué información deberían ofrecer los diccionarios?

En los diez diccionarios analizados se ha estudiado lo siguiente:

1) En el caso de los complementos argumentales prepositivos, direccionales y locales se observa si se nombra de forma explícita la preposición correspondiente y, en cuanto al alemán, el caso que ésta rige, de modo que no haya que deducirlo exclusivamente de los ejemplos aportados. Se considera adecuada e imprescindible una indicación del tipo *interessiert an + dat = interesado en*, puesto que *an* puede regir más de un caso. Sin embargo también se consideraría adecuado que únicamente se ofreciera información sobre el caso que rige la preposición cuando ésta puede regir más de un caso, como en el ejemplo anterior, siempre que este proceder se diese de forma sistemática. Sería por tanto justificable que en el caso de *fähig zu* (*capaz de*) o *dankbar für* (*agradecido por*) no se indicara el caso, puesto que las preposiciones *zu* y *für* siempre rigen, respectivamente, dativo y acusativo.

2) En el caso de los complementos argumentales realizables también mediante una oración, se observa si se hace referencia explícita al tipo de subordinada, por ejemplo mediante indicaciones como *NSdass* o *NSInf* para el alemán u *OSque* u *OSinf* para el español. También se considera aportada esta información si por ejemplo en el caso de *interessiert/interesado* se ofrecen paráfrasis del tipo *daran interessiert sein, etwas zu tun/estar interesado en hacer algo* y *daran interessiert sein, dass jemand etwas tut oder dass etwas stattfindet/estar interesado que alguien haga algo o que suceda alguna cosa*. En estrecha relación con ello, los diccionarios también deberían ofrecer información sobre si la oración principal alemana debe o no contener el elemento catafórico formado mediante *da(r)*+preposición. Esta información es importante, puesto que hay adjetivos que lo exigen (31) y otros en los que es facultativo (33):

(31) Wir sind sehr *daran* interessiert, an dem Projekt teilzunehmen.
(32) Estamos muy interesados en participar en el proyecto.
(33) Er war nicht *(dazu)* fähig, seine Mitarbeiter zu überzeugen.
(34) No fue capaz de convencer a sus colaboradores.

3) En el caso del complemento argumental dativo, se observa si se ofrece alguna paráfrasis que evidencie inequívocamente que se trata de este tipo de complemento. En el caso del alemán se consideraría adecuada una indicación del tipo *jemandem dankbar*, puesto que la forma declinada de *jemand* señala claramente que se trata de un dativo. En cambio, debería considerarse insuficiente, por lo que respecta al español, la indicación equivalente *agradecido a alguien*, puesto que el sintagma preposicional no garantiza que el usuario lo interprete inequívocamente como un complemento indirecto. Una indicación del tipo *agradecido a alguien/le* o *estar(le) agradecido a alguien* sería sin duda más adecuada.

4) En todos los casos se analiza si se indica de alguna manera el hecho de que el complemento argumental en cuestión es obligatorio o facultativo. Al usuario le puede interesar, por ejemplo, si los adjetivos *dankbar/agradecido* pueden aparecer con solo uno de los dos complementos argumentales que rigen o incluso sin ninguno de ellos, sin que por ello cambie el significado del enunciado:

(35) Ich bin Ihnen sehr dankbar für Ihre Hilfe.
(36) Le estoy muy agradecido por su ayuda.
(37) Ich bin Ihnen sehr dankbar.
(38) Le estoy muy agradecido.
(39) Ich bin sehr dankbar für Ihre Hilfe.
(40) Estoy muy agradecido por su ayuda.
(41) ?Ich bin sehr dankbar.
(42) ?Estoy muy agradecido.

O si los adjetivos *interessiert/interesado* pueden aparecer sin su complemento argumental prepositivo:

(43) ?Wir sind sehr interessiert.
(44) ?Estamos muy interesados.

Se considera información adecuada al respecto que se enuncie de forma explícita o que se indique, ya sea en la paráfrasis introductoria o en los ejemplos aportados, poniendo p. e. los complementos argu-

mentales facultativos entre paréntesis (previa explicación en las instrucciones de uso del diccionario). Así, en cuanto a los adjetivos *dankbar/agradecido*, si se considera que ambos complementos argumentales son realmente facultativos, podría indicarse mediante una paráfrasis como *(jemandem) (für etwas) dankbar/agradecido (a alguien/le) (por algo)* o poniendo entre paréntesis los complementos argumentales en los ejemplos (35) y (36). Del mismo modo también debería señalarse de alguna manera si el adjetivo debe aparecer como mínimo con uno de ellos, si no fuese posible la elisión de ambos a la vez.

5) También se analiza en todos los casos si se ofrece información sobre restricciones semánticas que afectan al complemento argumental en cuestión. Se considera información explícita al respecto que, independientemente de los ejemplos aportados, pueda deducirse de forma inequívoca si el complemento puede referirse a personas o a cosas o a ambas indistintamente. Se trataría de ofrecer p. e. una paráfrasis como la mencionada en el punto anterior en cuanto a los adjetivos *dankbar/agradecido* u otra del tipo *an etwas/jemandem interessiert /interesado en algo/alguien* en el caso de los adjetivos *interessiert/interesado*.

6) Finalmente se analiza en todos los casos si los diccionarios ofrecen ejemplos unívocos de todo lo anterior. Como tales se consideran únicamente oraciones enteras como *Wir sind an diesem Projekt interessiert/Estamos interesados en este proyecto o Wir sind daran interessiert, dass alle an dem Projekt teilnehmen/ Estamos interesados en que todos participen en el proyecto* y, en ningún caso, posibles paráfrasis como *an etwas interessiert sein/estar interesado en algo*.

3 Los adjetivos y sus complementos en los diccionarios analizados

1) De los nueve diccionarios analizados, solo uno (BEO) prescinde por completo de informar sobre las preposiciones que pueden o deben intervenir en los complementos argumentales según lo definido en el punto (2) del apartado anterior. Eso no significa, sin embargo, que los demás sí lo hagan. En realidad, ninguno de ellos lo hace en cada una de las 12 ocasiones en que deberían informar al respecto (*fähig/capaz, interessiert/interesado, reich/rico, wohnhaft/residente, gebürtig/natural* y *dankbar/agradecido)*. Si ordenamos los diccionarios de menor a mayor frecuencia, obtenemos el siguiente resultado: DIX (1), REV (2), LEO (6), LAR (6), SLG (6), PON (7), HUE (7) y LNG (9). Por otra parte, en los seis últimos se observa que la información no se ofrece de forma sistemática en las dos partes del diccionario. Valgan los casos de LNG y SLG como ejemplo: En el primero se ofrece la información pertinente de forma bidireccional respecto de *fähig/capaz*, *interessiert/interesado* y *reich/rico* tanto en la parte alemana como en la española; en cambio, respecto de *wohnhaft/residente*, *gebürtig/natural* y *dankbar/agradecido* únicamente se ofrece en la parte alemana. En el caso de SLG solo se ofrece información bidireccional en ambas partes con respecto a *gebürtig/natural*, mientras que la información correspondiente respecto de *fähig/capaz*,

reich/rico y *dankbar/agradecido* sólo se ofrece en la parte alemana y la de *wohnhaft/residente* e *interessiert/interesado* únicamente en la parte española.

En cuanto al caso que rige la preposición alemana, exceptuando LEO, que cuando ofrece información sobre la preposición lo hace indicando siempre el caso que rige, incluso cuando ésta solo puede regir un único caso, los demás diccionarios optan mayormente por ofrecer dicha información solo cuando la preposición puede regir dos casos diferentes. Así, en HUE encontramos junto a la indicación "zu etw. fähig sein ser capaz de algo", en la que no se señala el caso, la indicación "reich an etw. dat rico en algo", mediante la que se evidencia que la preposición *an* rige dativo. Mientras que algunos diccionarios proceden de forma sistemática al respecto, en otros, como en PON, el hecho de informar sobre el caso parece producirse de forma aleatoria: Los complementos argumentales prepositivos de *interessiert* y *reich* se introducen mediante la preposición *an*, que en principio podría regir tanto acusativo como dativo. Sin embargo sólo se indica el caso concreto en las entradas *interessiert* ("an etw dat/jdm interessiert sein estar interesado en [oder por] algo/alguien") e *interesado* ("estar interesado en algo an etw dat interessiert sein"), mientras que en *reich* solo se dice: "reich an". El adjetivo *gebürtig* posee un complemento direccional introducido por *aus*, preposición que siempre rige dativo. Mientras que en la entrada española correspondiente natural se indica "natural de gebürtig aus + dat", en la entrada del adjetivo alemán solo se lee "gebürtig aus natural de".

2) La información que ofrecen los diccionarios sobre las distintas posibilidades de realización del complemento argumental solo puede calificarse de muy deficiente. Son tres las parejas de adjetivos que también pueden ser realizados mediante oraciones subordinadas, ya sean finitas o de infinitivo: *fähig/capaz*, *interessiert/interesado* y *dankbar/agradecido*. La información debería ser aportada, pues, en un total de seis casos en todos los diccionarios, por lo tanto en un total de 54 ocasiones. Sin embargo solo se encuentra indicada o al menos insinuada la posibilidad de realización oracional en dos ocasiones y lo cierto es que en solo una de ellas aparece formulada de la manera que se ha descrito más arriba como adecuada. Se trata de la información aportada sobre *fähig* en HUE: "fähig sein, etwas zu tun ser capaz de hacer algo", paráfrasis que aparece junto a la que haría referencia a la realización nominal o pronominal "zu etw. fähig sein ser capaz de algo". La segunda se encuentra en PON, donde respecto de *interessiert* se ofrece la siguiente información, que en realidad no es la típica paráfrasis, pero que tampoco llega a la categoría de ejemplo: "ich bin nicht daran interessiert, dass … no me interesa que… +subj". Resulta curioso que en la correspondencia en español se opte por la expresión verbal "no me interesa que", cuando a fin de hacer más comprensible la construcción alemana se habría poder optado perfectamente, al menos complementariamente, por la expresión adjetival paralela "no estoy interesado en que". Lo cierto es que al consultar en el mismo diccionario la entrada *interesado* tampoco se encuentra referencia alguna a una posible realización oracional.

Ninguno de los diccionarios analizados incide de algún modo en la obligatoriedad o facultatividad del elemento catafórico que debe o puede aparecer en el caso de realización oracional en el caso del ale-

mán. De ahí que el usuario de PON no pueda deducir de ninguna manera si el adverbio pronominal *daran* en la indicación que se acaba de mencionar es obligatorio o no.

3) Son tres las parejas de adjetivos que poseen un complemento indirecto: *dankbar/agradecido, treu/fiel* y *untreu/infiel*. En SLG, LEO y DIX no se encuentra referencia alguna al respecto. En los demás, la información se ofrece de forma dispersa y poco sistemática. Ofrecen información parcial HUE (para *agradecido* y *fiel*), REV (para *treu*) y BEO (para *treu/fiel*). Se trata de indicaciones como "jemandem treu sein ser fiel a alguien" (BEO), en las que solo resulta evidente que se trata de un complemento indirecto en alemán. Solo en LNG (para *dankbar* y *untreu*) y en PON y LAR (para *treu* y *untreu*) se encuentran formulaciones que evidencian que también en español se trata de un complemento indirecto: "estar(le) a alg agradecido por a/c" (LNG) o "serle infiel a alguien" (LAR). La falta de coherencia interna de los diccionarios también resulta evidente aquí, al comparar la última información en LAR con la que el mismo diccionario ofrece respecto de *dankbar*: "jm (für etw) dankbar sein estar agradecido (f agradecida) a alguien (por algo)" (LAR). Nótese que aquí no se incluye, como en el caso de *infiel*, el pronombre *le*.

4) Una posible indicación en cuanto a la obligatoriedad o facultatividad del complemento argumental, tal y como se define como adecuada en el apartado 2, solo se encuentra en HUE y LAR, concretamente respecto de *dankbar/agradecido*. Así, en la paráfrasis definitoria de *dankbar* (no en la de *agradecido*) en LAR se dice: "jm (für etw) dankbar sein estar agradecido (f agradecida) a alguien (por algo)". En HUE se encuentra el mismo tipo de indicación, pero en este caso en la entrada de *agradecido* (no en la de *dankbar*): "estar agradecido con alguien (por algo) jdm (für etw.) dankbar sein". Podría deducirse de ello que los dos diccionarios consideran, al colocarlo entre paréntesis, que el complemento argumental prepositivo de *dankbar/agradecido* es facultativo en ambas lenguas. Sin embargo, no existe en los citados diccionarios ninguna instrucción de uso que explique el significado de estos paréntesis. Si se consultan en ellos otros adjetivos con los que el complemento argumental prepositivo debería considerarse facultativo, como por ejemplo *stolz/orgulloso* o *verliebt/enamorado*, no se halla ninguna indicación similar. Y lo mismo sucede si se consulta en ellos la descripción de verbos como por ejemplo *sich ärgern (enfadarse)*. En ninguno de los dos diccionarios aparece el complemento argumental prepositivo entre paréntesis con el fin de indicar que es facultativo: "sich über etw./jdn ärgern enfadarse por algo/con alguien" (HUE); "sich über etw/jn ärgern enfadarse por algo/por culpa de alguien" (LAR). Podría deducirse de ello que los paréntesis arriba mencionados son debidos a la mera casualidad.

5) En todos los diccionarios, excepto en REV y DIX, se encuentra alguna información sobre restricciones semánticas que afectan a los complementos argumentales de los adjetivos *fähig/capaz, interessiert/interesado, reich/rico, dankbar/agradecido, treu/fiel* y *untreu/infiel*. Valga como ejemplo lo que se dice en LNG respecto de *dankbar*: "jemandem für etwas dankbar sein estar(le) a alg agradecido por a/c". Pero también aquí se observa que los diccionarios no proceden de forma sistemática, como lo demuestra la simple comparación cuantitativa. Del total de 12 ocasiones en que podría ofrecerse este tipo de infor-

mación, ninguno de los diccionarios lo hace en todas ellas. Si se ordenan de menor a mayor frecuencia, el resultado es el siguiente: SLG (1), BEO (1), LEO (2), LNG (2), PON (4), LAR (7) y HUE (8). Incluso en el caso de este último resulta difícil de entender por qué en cuanto a *fähig/capaz, interessiert/interesado* y *reich/rico* se ofrece dicha información de manera bidireccional tanto en la parte alemana como en la española, mientras que en el caso de *dankbar/agradecido* y *treu/fiel* solo se hace en la parte española, prescindiendo además por completo de dicha información por lo que respecta a *untreu/infiel*. Lo cierto es que, incomprensiblemente, el adjetivo alemán *untreu* ni tan solo aparece en el diccionario.

6) Contrariamente a lo que sería deseable en diccionarios bilingües, la mayoría de diccionarios analizados prescinden totalmente (REV) o al menos en gran medida de ofrecer ejemplos de uso de los adjetivos estudiados. Entre los últimos se encuentran SLG y LAR, que solo ofrecen un único ejemplo de los 16 posibles, y LNG y DIX, que solo los ofrecen en dos ocasiones. Únicamente HUE y PON se acercan a lo que sería deseable, al ofrecerlos en ocho ocasiones. Se trata de ejemplos como: "Er ist zu allem fähig", Das Essen ist reich an Kohlehydraten", "Ich wäre Ihnen für eine schnelle Antwort sehr dankbar" e "Ihr Mann war (ihr) untreu" (HUE) o "estos muchachos son capaces de todo diese Burschen sind zu allem fähig", "er ist aus Valencia gebürtig es natural de Valencia" e "ich weiß, dass mein Mann mir treu ist sé que mi marido me es fiel" (PON). Cabe señalar que HUE solo ofrece ejemplos en alemán, proceder que podría considerarse justificado por el hecho de ir dirigido a usuarios que aprenden alemán y que por lo tanto ya saben español, pero que en última instancia puede resultar contraproducente, si se tiene en cuenta que dichos usuarios consultan el diccionario precisamente por desconocer dicho idioma, lo que puede comportar que no acaben de entender los ejemplos ofrecidos. LEO y BEO, que no se caracterizan precisamente por aportar una información detallada sobre la complementación, ofrecen un sinfín de ejemplos reales, que no pueden considerarse realmente útiles al ofrecerse de forma totalmente asistemática.

4 Conclusión

Todos los diccionarios analizados ofrecen una información claramente deficiente respecto a la complementación del adjetivo. Entre las deficiencias cabe destacar sobre todo que no ofrecen siempre información detallada sobre la preposición que introduce el complemento y que la información sobre la posible realización oracional de los complementos y sobre su obligatoriedad es prácticamente inexistente. En todos ellos se observa además que son poco sistemáticos, puesto que en determinados casos ofrecen cierta información al respecto, mientras que en otros prescinden totalmente de ello, sin que dicho proceder pueda atribuirse a unos parámetros determinados y lógicos. Por otra parte se constata que algunos se limitan a actuar como simples traductores de palabras sueltas, sin ofrecer al usuario indicaciones respecto a su uso. Del mismo modo se constata que la falta de ejemplos es una característica común a la gran mayoría de ellos. En los dos diccionarios en que se ofrece una gran variedad de

ejemplos reales (LEO y BEO), esta característica no puede considerarse positiva, puesto que al usuario le puede resultar muy difícil encontrar lo que busca.

5 Referencias

5.1 Diccionarios analizados

Beolingus online. Acceso: http://dict.tu-chemnitz.de [10/03/2014].
DIX online. Acceso: http://dix.osola.com/index.es.php [10/03/2014].
Hueber Wörterbuch. Diccionario Alemán-Español / Español-Alemán (2007). Ismaning: Hueber.
Langenscheidt (2002). *e-Taschenwörterbuch Spanisch-Deusch-Spanisch* (CD Version 4.0). Berlin/München: Langenscheidt.
Larousse Diccionario español-alemán-español online. Acceso: http://www.larousse.com/es/diccionarios/allemand-espanol [10/03/2014].
LEO Spanisch-Deutsch-Spanisch online. Acceso: http://dict.leo.org/esde/index_de.html [10/03/2014].
Pons Diccionario español-alemán-español online. Acceso: http://es.pons.eu/ [10/03/2014].
Reverso Diccionario Collins español-alemán-español online. Acceso: http://diccionario.reverso.net/aleman-espanol [10/03/2014].
Slaby, R., Grossmann, R. (1993) Diccionario de las lenguas española y alemana I. Español-Alemán. Barcelona: Herder.
Slaby, R., Grossmann, R., Illig, C. (1991) Diccionario de las lenguas española y alemana II. Alemán-Español. Barcelona: Herder.

5.2 Otras

Bosque, I. (1983). El complemento del adjetivo. En *Lingüística Española Actual* V, pp. 1-14.
Bosque, I., Demonte, V. (eds.) (1999). *Gramática descriptiva de la lengua española*. Barcelona: Ariel.
Brucart, J.M., Gallego, Á.J. (2009). "L'estudi formal de la subordinació i l'estatus de les subordinades adverbials". En: *Llengua & Literatura*, 20, pp. 139-191.
Castell, A. La recepción de la teoría de valencias en España. Reflexiones en torno a la adaptación al español de los términos 'Ergänzungen' y 'Angaben'. En *Revista de Filología Alemana* 17, pp. 183-204.
Diccionario Salamanca de la lengua española. Acceso: http://fenix/cnice.mec.es/diccionario/ [10/03/2014].
Duden (1973). *Grammatik der deutschen Gegenwartssprache*. Mannheim: Bibliographisches Institut.
Duden (2008). *Deutsches Universalwörterbuch. (CD Version 5.0)*. Leipzig: Bibliographisches Institut & F.A. Brockhaus AG.
Duden (2009). *Die Grammatik*. Mannheim/Wien/Zürich: Dudenverlag.
Eisenberg, P. (1986). *Grundriss der deutschen Grammatik*. Stuttgart: J.B. Metzlersche Verlagsbuchhandlung.
Engel, U. (1988). *Deutsche Grammatik*. Heidelberg: Julius Groos Verlag.
Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung*. München: Iudicium.
Garriga, C., Bargalló, M. (2000). Gramáticas y diccionarios: el régimen del adjetivo. En *Euralex 2000 Proceedings*, pp. 601-609.
Helbig, G., Buscha, J. (1972). *Deutsche Grammatik*. Leipzig: VEB Verlag Enzyklopädie.
Langenscheidt (2003). *e-Großwörterbuch Deutsch als Fremdsprache* (CD Version 4.0). Berlin/München: Langenscheidt.

Llácer, I. et al. (1996). *Lengua Española. Nivel COU*. Fuente del Jarro-Paterna: Ecir Editorial.
Moliner, María (2008). *Diccionario de uso del español* (CD versión 3.0). Madrid: Gredos.
Pavón, M.V. (2012). Estructuras sintácticas en la subordinación adverbial. Madrid: Arco Libros.
Pons. *Deutsch als Fremdsprache online*. Acceso: http://www.pons.de/produkte [10/03/2014].
Real Academia Española (1973). *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa Calpe.
Real Academia Española: Banco de datos (CREA) (en línea). Corpus de referencia del español actual. Acceso: http://www.rae.es [10/03/2014.]
Real Academia Española – Asociación de Academias de la Lengua Española (2009). *Nueva gramática de la lengua española*. Madrid: Espasa Libros.
Sommerfeldt, K-E., Schreiber, H. (1974). *Wörterbuch zur Valenz und Distribution deutscher Adjektive*. Leipzig: VEB Bibliographisches Institut.

Reconocimientos

La investigación que subyace a este trabajo ha sido parcialmente financiada por el Proyecto FFI2012-32705.

Esame storico dei “realia” nei dizionari bilingui italiano/ungheresi

Zsuzsanna Fábián
Università “Eötvös Loránd” di Budapest, Istituto di Romanistica
fabian.zsuzsanna@btk.elte.hu

Abstract

Lo scopo del presente lavoro (di carattere lessico-contrastivo ma anche storico-lessicografico) è quello di analizzare l’evoluzione del trattamento dei “realia” (parole culturo-specifiche, *lexical/referential gaps*) in otto dizionari bilingui italiano-ungheresi e in sei ungherese-italiani. Per l’esame sono stati scelti 16 esponenti italiani e altrettanti ungheresi appartenenti al lessico della gastronomia (i lemmi sono presentati nelle due Appendici). Le singole equivalenze registrate nei dizionari sono state assegnate alle varie categorie delle possibili strategie traduttive quali l’uso di una parola di origine straniera (forestierismi, prestiti o calchi), quello di un equivalente (sintagmatico) esplicativo (distinto ev. anche graficamente, p.es. col corsivo o con la collocazione tra parentesi, da un “vero” equivalente), o di un equivalente generalizzante (iperonimo), l’adattamento o la sostituzione con un equivalente approssimativo/parziale di L2, ecc. Nella parte conclusiva dell’analisi sono state indagate le problematiche seguenti: Quali sono le strategie traduttive più frequenti nei dizionari bilingui tra italiano e ungherese? E’ possibile individuare, nei singoli autori dei dizionari analizzati, una preferenza verso certe strategie traduttive? Nel caso di eventuali errori si possono registrare “copiature” e/o “correzioni”? Sussiste una connessione tra le strategie dominanti nei dizionari bilingui e la direzione delle lingue/tipo del vocabolario (attivo o passivo)?

Keywords: bilingual dictionaries; lexical gaps; Italian & Hungarian

1 I “realia” nella traduzione e come problema lessicografico

A partire da Vlahov & Florin,[1] il concetto dei “realia” è stato studiato da vari punti di vista e numerose sono le definizioni e le categorizzazioni (Rey 1986; Lendvai 1986; Tellinger 2003; Klaudy 2005 ecc.), spesso elaborate anche su basi empiriche (analisi contrastive di testi concreti tradotti). In questo saggio il termine *realia* sarà inteso sia nel senso di ‘referente caratteristico in un dato ambiente culturale’, sia come ‘segno linguistico indicante un referente di questo tipo’ (v. p.es. *Realienlexeme* in Kujamäki 2004: 920; *reálialexém*a in Heltai 2013: 33; e ancora Klaudy 2005: 60). Nella letteratura internazionale il fenomeno è spesso designato, oltre che dal termine riassuntivo già evocato, anche da espressioni au-

1 *Vlahov, S. & Florin, S. (1970). Neperovodimoe v perevode. Realii. In S*ovetskij pisatel’, Mosca, pp. 432-456.

tonome per i due diversi concetti quali *referential gap* (Svensén 2009: 271), *cultural gap* risp. *lexical gap* (v. p es. Svensén 2009: 271-275).

Le difficoltà che emergono a proposito dei realia si manifestano prima di tutto nel processo di mediazione tra due culture, e quindi, tipicamente, nelle traduzioni. È perciò evidente che una descrizione dettagliata e una ricca sistematizzazione, oggi generalmente condivisa dagli studiosi, sono state elaborate nell'ambito della traduttologia (Heltai 2001, House 2004, Kujamäki 2004, Klaudy 2005, 2013). Nello stesso tempo sono stati studiati anche i risvolti lessicografici dei problemi di equivalenza che le lacune lessicali pongono agli autori dei dizionari bilingui (p.es. Tomaszczyk 1984; Rey 1986; Marello 1989: 52-54; Rey 1991; Uzonyi 2006; Éber & Fata 2009-2010; Fata 2010). Alcuni studiosi analizzano anche i tipi delle commutazioni adoperate sia nella traduzione che nella prassi lessicografica (Szabó 2008; Fábián 2013).

In questo saggio le categorie dell'esotizzazione (ingl. f*oreignization*) e della localizzazione (o adattamento, ingl. *domestication*) saranno rispettivamente intese come "strategia traduttiva consistente nell'introdurre o conservare elementi culturali appartenenti a una cultura diversa da quella ricevente" (Osimo 2003: 201), e come "strategia traduttiva in base alla quale i traduttori eliminano gli elementi culturali appartenenti a una cultura diversa da quella ricevente [...] sostituendovi elementi della cultura ricevente" (Osimo 2003: 210; Varga-Mujzer 2009). Esaminando invece le diverse classificazioni dei possibili cambiamenti traduttivi (descritti nei dettagli da Osimo 2003; Klaudy 2005: 43-171, e ancora in Heltai 2001: 12; Uzonyi 2006: 121; Varga-Mujzer 2009: 6; Éber & Fata 2010: 41 sulle orme di Tellinger 2003; ecc.) possiamo ribadire che nei dizionari sono adoperati, più frequentemente, i seguenti metodi: uso di una parola straniera, conio e uso di un prestito, conio e uso di un calco, uso di un equivalente (sintagmatico) esplicativo, uso di un equivalente generalizzante (iperonimo), adattamento o sostituzione con un equivalente approssimativo o parziale di L2. Szabó (2008: 62) considera i procedimenti elencati "surrogati di equivalenza".

2 Descrizione del corpus scelto per l'analisi e della metodologia

Il presente lavoro si pone il duplice scopo di indagare i tipici trattamenti dei realia nei dizionari bilingui italiano / ungheresi e di osservare, nello stesso tempo, se e come i cambiamenti traduttivi siano mutati nell'arco dell'ultimo secolo, periodo in cui i vocabolari presi in'esame sono stati pubblicati. Si tratta di otto dizionari italiano-ungheresi e di sei ungherese-italiani, a partire dal 1912. (Per la storia dei vocabolari italiano / ungheresi v. Fábián 2011. Le sigle che contraddistinguono i dizionari nel presente lavoro sono elencate nella parte bibliografica.)

I vocaboli scelti per il corpus (16 parole-esponenti sia nell'una che nell'altra direzione) appartengono al lessico della gastronomia: un campo in cui la presenza dei realia culturo-specifici è prevedibile. Si tratta dei seguenti vocaboli italiani: *barbera, cannelloni, chianti, gorgonzola, grissino, lasagna, marsala,*

mascarpone, mortadella, pandoro, panettone, panforte, pastasciutta, pizza, raviolo, tortellino. Le parole ungheresi (seguite da una breve spiegazione del significato tra virgolette) sono invece le seguenti: *beigli* ('dolce tipico di Natale, a forma di rotolo, con ripieno di noci o di semi di papavero macinati'), *fröccs* ('una specie di spritz preparato con vino e acqua di seltz'), *gulyás* ('gulasch: zuppa a base di carne di manzo e verdure'), *halászlé* ('zuppa di pesce alla paprica'), *kadarka* ('tipo di vino rosso, da tavola'), *kovászos uborka* ('cetrioli interi preparati attraverso fermentazione'), *lángos* ('leggera pasta lievitata, fritta in abbondante olio, dalla forma rotonda o quadrata'), *paprikás csirke* ('pezzi di pollo in umido, alla paprica'), *pogácsa* ('pasticcino/foccaccina rotondo/a, salato/a, ev. al formaggio o di altri gusti'), *rétes* ('strudel'), *szaloncukor* ('cioccolatino riempito di creme di vari gusti, tipico di Natale perché – incartato – adorna l'albero'), *tejberizs* ('riso al latte'), *tejföl* ('panna acida'), *tokaj(i)* ('vino di Tokaj'), *töltött káposzta* ('crauti/cavoli farciti'), *túrós csusza* ('pasta cotta a forma di rettangoli cc. 3 x 3 cm, con sopra ricotta e panna acida').
Quanto alla presenza dei lemmi selezionati per l'analisi possiamo affermare che nella direzione italiano-ungherese sono tutti compresi in Kő, H e H & J. Il fatto che in U mancano invece 12, e in KK1 11, delle parole scelte per il nostro corpus,[2] può essere spiegato dal diverso formato dei dizionari: i primi tre sono i "grandi" dizionari, gli ultimi due invece di mole molto minore. Nella direzione ungherese-italiana tutti i vocaboli scelti sono presenti in G & S; nei "grandi" mancano invece due lemmi in KKJ e solo uno in KKJ & J.[3] Interessante è la situazione relativa ai due dizionari bidirezionali per turisti: mentre in F & V si trovano 13-13 lemmi nelle due direzioni, in H & I abbiamo 12 lemmi nella direzione italiano-ungherese ma solo 6 in quella ungherese-italiana,[4] fatto che attesta la non completa bidirezionalità di questa opera.
Le equivalenze indicate nei singoli vocabolari sono esposte in dettaglio nelle due Appendici finali.
In questa analisi gli esponenti appaiono in corsivo (e possono risultare in corsivo anche le equivalenze, se presenti in questo carattere nei dizionari); le equivalenze sono precedute dal segno dell'uguaglianza (=). Le traduzioni delle parti in ungherese sono state collocate tra due apici ('…').

3 Risultati dell'analisi

3.1 I cambiamenti traduttivi più frequenti nei dizionari bilingui

La scelta tra i diversi cambiamenti traduttivi operata dal lessicografo è connessa prima di tutto con la mole del dizionario: in uno più ridotto essa sarà vincolata ad una gestione più parsimoniosa dello

2 In U sono presenti solo gorgonzola, lasagna, mortadella, tortellino; in KK1 invece lasagna, marsala, mortadella, panforte, pizza.
3 In KKJ töltött káposzta e túrós csusza e in KKJ & J túrós csusza.
4 Sono presenti in H & I fröccs, gulyás, halászlé, rétes, tejberizs, tejföl.

spazio (si tenderà, per esempio, ad evitare l'uso delle circoscrizioni esplicative, a favore degli adattamenti o dei prestiti, delle soluzioni generalizzanti con un iperonimo, ecc.).
Non tutti i cambiamenti traduttivi possono essere adottati anche nel lavoro lessicografico. Non è qui possibile, per esempio, ricorrere al metodo del tralascio perché a questo equivarrebbe, nel dizionario, la cancellazione del lemma stesso. D'altronde, è possibile solo nel dizionario la combinazione di più cambiamenti traduttivi: è molto frequente, per esempio, la combinazione dell'uso di un prestito o di un adattamento localizzante con una spiegazione esplicativa.
Nei seguenti paragrafi saranno elencati i cambiamenti traduttivi più frequentemente adoperati nei dizionari, illustrati da esempi tratti dai dizionari in esame.

3.1.1 Uso di una parola di origine straniera

Ci sono più possibilità per l'uso di una parola di origine straniera anche come equivalenza in L2, ma viste le peculiarità delle due lingue e per il fatto che il lessicografo basa il suo lavoro più su materiali scritti che sul parlato, nella direzione italiano-ungherese sarebbe difficile, se non impossibile annoverare[5] le equivalenze di questo tipo nelle classiche sottocategorie delle parole di origine straniera (forestierismo: senza adattamento a L2, prestito: adattato formalmente a L2, calco: traduzione). Gli esponenti italiani che appaiono anche come equivalenti in L2 sono *chianti* e *mortadella* (in 5 dizionari), *gorgonzola* (in 4), *pizza* (in 3), *tortellini* (in 2), *lasagna, mascarpone, panettone* (in 1). È H & J il dizionario in cui questo tipo è maggiormente presente (infatti, vi appaiono tutte le equivalenze indagate), ed è sempre questo dizionario ad usare per la prima volta *pizza* e *tortellini* anche in funzione di equivalenti ungheresi. Le equivalenze nel caso di *marsala* rientrano invece in una categoria mista: l'aggettivo derivato (con *-i*) dal nome della città siciliana appare sia nella forma originaria (= *marsalai* KK1, Ki) che in forma trascritta (= *marszalai* H, H & J), ed accompagna in tutti i casi il nome *bor* 'vino' (iperonimo). Potrebbe essere considerato un "falso calco" anche il caso di *pastasciutta* = száraztészta 'pasta non ancora cotta' (H, H & J). Nella direzione ungherese-italiano sono i due nomi di vini (*tokaji* in 4 e *kadarka* in 2 dizionari) e l'internazionalismo = gulasch per *gulyás* (in 4 dizionari) a rientrare in questa categoria delle equivalenze.

3.1.2 Uso di un equivalente sintagmatico esplicativo

In mancanza di un equivalente lessicale in L2 il lessicografo può scegliere di offrire una spiegazione sintagmatica che descriva i tratti essenziali del denotatum in L1. Questo metodo è adoperato molto spesso dai lessicografi, come attestato anche dagli innumerevoli casi riscontrati nel nostro corpus. Gli

5 Il caso di pizza, ormai radicato nell'ungherese, presenta i seguenti problemi. Quanto alla pronuncia, convivono oggi [pitstsa] (come in it. alzo) e [pizza] (come in it. viso), ma appare anche [pidzdza] (come in it. gazza), di cui fortemente concorrenziali le prime due. Se nel tempo dovesse vincere la seconda, la scrittura potrebbe rimanere inalterata (ma nascerebbe "un falso amico", nella pronuncia, tra it. e ungh.). Se dovesse invece vincere la prima variante di pronuncia (quella "all'italiana"), per poterla annoverare tra i classici prestiti bisognerebbe trascriverla secondo le regole dell'ungh. in picca (difficilmente ipotizzabile, vista anche la ormai globale diffusione dell'originaria forma italiana).

equivalenti esplicativi sono spesso differenti tra loro. *Panettone* è, secondo Kő, = lisztből, cukorból, sáfrányból készített és sörrel kelesztett sütemény ('dolce preparato con farina, zucchero, zafferano e lievitato con la birra'), per H è = mazsolás, cukrozott gyümölcsös milánói kalács" ('focaccia milanese con uva sultanina e canditi') e per H & I = *kuglófhoz hasonló, milánói eredetű karácsonyi sütemény*" ('dolce natalizio di origine milanese simile al Gugelhupf[6]'); *paprikás csirke* è = pollo in umido colla paprica" in G & S, = pollo con intingolo di paprica, spezzatino con paprika" in KKJ, e ancora = pollo alla paprika, spezzatino di pollo in KKJ & J e = pollo all'arrabbiata, pollo all'ungherese in F & V.

Le esplicazioni appaiono spesso anche come aggiunte agli adattamenti localizzanti (*panettone* = kuglóf *(karácsonyi édesség* 'dolce natalizio') (F & V)), agli iperonimi generalizzanti (*raviolo* = étel (vagdalthús- és túrósgombócból) 'pietanza (fatta da gnocchi di carne macinata o di ricotta'[7]) (Ki), e alle parole di origine straniera in L2 (*panettone* = panettone [karácsonyi milánói kuglóf 'Gugelhupf natalizio milanese'] (H & J).

3.1.3 Uso di un equivalente generalizzante (iperonimo)

Questo metodo appare molto frequentemente nei dizionari bilingui, prima di tutto in quelli di piccola mole. Un iperonimo può apparire come unico equivalente (spesso p. es. in Király: *gorgonzola* = sajtfajta 'tipo di formaggio', *panforte* = édestésztafajta 'tipo di dolce'). Siccome un unico equivalente iperonimico dice spesso troppo poco, come indicazione orientativa i lessicografi possono aggiungervi un attributo (*pizza* = zsíros sütemény 'un dolce che contiene grasso' Kő). Si cerca di indicare l'iperonimo più vicino, ma non mancano i casi in cui questo risulta troppo generalizzante (forse il lessicografo non conosce il denotatum?) perché situato, nella gerarchia tassonomica, diversi livelli sopra e l'equivalente ha dunque bisogno di un'aggiunta esplicativa (Ki: *raviolo* = étel (vagdalthús- és túrósgombócból) 'pietanza (fatta da gnocchi di carne macinata o di ricotta'). Nel caso dei nomi di vini (ad eccezione di *tokaji*) viene quasi sempre aggiunto l'iperonimo *vino* risp. *bor*.

Un iperonimo può servire anche come completamento di altri tipi di equivalenza, come nel caso di forestierismi non ancora diffusi in L2 (*gorgonzola* = gorgonzóla (sajt) 'formaggio' Kő, U; *mortadella* = mortadella (felvágott) 'affettato' H, H & J, F & V) o nel caso di eventuali fraintendimenti *tortellino* = húsos táska[8] (étel) 'fardellino di pasta farcito di carne (pietanza)' Ki).

3.1.4 Adattamento o sostituzione con un equivalente approssimativo o parziale di L2

Nel caso dell'adattamento il lessicografo cerca di trovare l'equivalente che più somigli all'originale, che riesca a rievocarlo, ad avvicinarsi ad esso. Nel nostro corpus appartengono a questo sottotipo i casi seguenti: *cannelloni* = palacsinta 'crêpe' (H, F & V), *grissino* = sütemény 'dolce' (Kő), = ropi 'stick salati' (H & J), *mortadella* = disznósajt 'sopressata' (Kő, KK1, Ki), = májas hurka'salsiccia al fegato' (U), *panettone*

6 Siccome i dizionari ungherese-italiani traducono ungh. *kuglóf* proprio con *panettone*, conviene usare qui il corrispondente tedesco.

7 Sic!

8 Il sintagma *húsos táska* potrebbe infatti significare, per la ricca polisemia della parola *táska*, 'una borsa con dentro la carne' o 'destinata al trasporto, alla custodia della carne' ecc.

= sütemény 'dolce' (Kő), = kalács 'pane dolce, lievitato, fatto a trecce' (H), *pastasciutta* = makaróni 'maccheroni' (Kő, Ki), *pizza* = lángos 'leggera pasta lievitata fritta in abbondante olio' (KK1, Ki, H, H & J), *lángos* = crostino (G & S), = pizza (KK2, KKJ), *rétes* = millefoglie (KKJ, KKJ & J).

In base ai casi riscontrati va ribadito che il metodo dell'adattamento può essere considerato "pericoloso" nei dizionari perché indirizza l'utente verso false direzioni. L'adattamento è quindi da evitare nei dizionari bilingui.

3.2 Preferenze dei singoli lessicografi quanto ai cambiamenti traduttivi

Nel primo "grande" vocabolario italiano-ungherese l'autore Kőrösi predilige ampie spiegazioni esplicative. Il metodo armonizza con lo stile generale del suo dizionario, ritenuto da Rezső Honti, anch'egli italianista e contemporaneo di Kőrösi, un'opera "di stile vivace, loquace" che vale la pena non solo di usare come vocabolario ma di leggere per puro diletto.[9] (Trattandosi di nomi di pietanze, è lecito dire che in molti casi le definizioni lessicografiche sfociano in vere e proprie ricette.[10]) In questo dizionario in più della metà dei lemmi esaminati viene usato il metodo esplicativo, e anche gli adattamenti e i forestierismi sono spesso accompagnati da aggiunte chiarificatrici (*pastasciutta* = (leves helyett tálalt) makaróni 'maccheroni (che si mangiano come prima portata')).

Nei dizionari di Urbanek, Gelletich & Sirola e di Koltay-Kastner 1 e 2, forse anche a causa della mole minore, domina invece il metodo dell'adattamento (*tortellino* U = táska, *pizza* KK1 = lángos, *gulyás* G & S = guazzetto).

Quanto a Király, colpisce una marcata preferenza per una soluzione generalizzante che viene usata all'incirca nella metà delle equivalenze (*gorgonzola* = sajtfajta 'tipo di formaggio', *panettone* = édestésztaféle 'tipo di pasta dolce'): nei suoi lemmi troviamo infatti spesso gli elementi *-féle, -fajta* 'un tipo di' che seguono l'iperonimo.

Herczeg adopera tutti i metodi possibili per le equivalenze, spesso utilizzandoli anche in combinazione. In H & J l'approccio diventa teoricamente più fondato e sistematico, nel senso che le esplicazioni sintagmatiche saranno distinte dalle vere e proprie equivalenze lessicali: le esplicazioni sintagmatiche metalinguistiche sono inserite tra i segni < >, le aggiunte agli adattamenti solo approssimativi sono invece collocate tra parentesi quadre: *chianti* = <toszkán vörösbor> chianti 'vino rosso della Toscana', *cannelloni* = [hússal, sajttal stb. töltött] tésztarolád 'rotolo di pasta [riempito di carne, formaggio ecc.]'.

Nei "grandi" dizionari tra ungherese e italiano di Koltay-Kastner (KKJ e KKJ & J) ritornano i metodi considerati sopra a proposito dei dizionari di Herczeg: *fröccs* = <vino e acqua gasata>. Tale affinità metodologica può essere spiegata con il semplice fatto che i dizionari di Herczeg e di Koltay-Kastner sono stati redatti e pubblicati all'incirca negli stessi anni e dalla stessa casa editrice.

9 Honti, R. (1912). Recensione a *Olasz-magyar szótár* di Kőrösi. In *Magyar Nyelvőr* X (3), p. 471.

10 P.es. *raviolo* = vajjal, túróval, fűszerszámmal és vagdalt hússal töltött toklóhúsból készült étel ('pietanza a base di carne di agnello, con burro, ricotta e spezie')

Infine, nei due dizionari destinati ai turisti dominano le spiegazioni esplicative in corsivo. F & V è il primo dizionario tra italiano e ungherese in cui queste sequenze metalinguistiche sono segnalate in corsivo, per distinguerle nettamente dalle equivalenze vere e proprie: *panforte = (sienai édesség)* 'dolce tipico di Siena', *fröccs = (due decilitri di vino con un decilitro di acqua di seltz).* Il metodo del corsivo (ma senza parentesi) sarà adoperato anche nella direzione italiano-ungherese in H & I: *panforte = gyümölcskenyérhez hasonló sienai karácsonyi sütemény* 'dolce di Natale simile al panfrutto tipico di Siena'.
Sarebbe fuorviante dedurre da questa breve panoramica sulle strategie traduttive adottate nei dizionari bilingui che esse siano rimaste immutate nell'arco dei cento anni trascorsi: pare che fossimo partiti, quanto ai realia, dal metodo esplicativo (usato con virtuosismo dal Kőrösi) e che sia proprio questo anche oggi quello adoperato di preferenza dai lessicografi (v. i due dizionari per turisti). Bisogna tuttavia considerare l'ultima fase del processo descritto come risultato dell'evoluzione della teoria lessicografica, in quanto le equivalenze esplicative appaiono nei dizionari di oggi come sequenze metalinguistiche (e quindi in corsivo e/o tra parentesi ecc.) e non come equivalenze sintagmatiche di tipo circoscrittivo.

3.3 Connessioni tra il tipo del referente e la strategia traduttiva

Considerato il numero relativamente esiguo degli esempi presi in esame, possiamo solo ribadire che i nomi dei vini risalenti a toponimi, quali i deonomastici chianti, risp. tokay, tokaj ecc. sono adoperati, in ambedue le direzioni, come forestierismi in L2.

3.4 Errori di equivalenza: "copiature" e "correzioni"

Il carattere diacronico dell'analisi non solo rende palesi le equivalenze poco fortunate o addirittura sbagliate, che sono state eventualmente riprese dai posteri, ma mette in luce anche le innovazioni o le correzioni degli errori. Le prime caratterizzano tipicamente i prodotti precoci della direzione italiano-ungherese e sono connesse all'utilizzo, da parte del lessicografo, della strategia dell'adattamento localizzante (spesso accompagnato da aggiunte esplicative).
Di seguito sono elencati alcuni altri casi che mostrano anche le fasi della correzione delle equivalenze.
Per *grissini* Kő adopera l'equivalente generalizzante = sütemény 'dolce', e anche se vi aggiunge l'attributo *ropogós* 'croccante', si tratta di una soluzione sbagliata; Ki introduce un altro equivalente fuorviante: = rúdkenyér 'pane a forma allungata, filoncino, baguette', che viene solo mutato nell'inesistente e sempre sbagliato = kenyérrúd 'ibid.' in H (e ripreso ancora in F & V nell'equivalente sintagmatico esplicativo); in H & J e H & I, invece, gli equivalenti esplicativi risultano finalmente corretti. (È da notare come *grissini* stia, nel frattempo, diventando prestito nell'ungherese.)

La parola *lasagna* viene interpretata da Kő come 'pasta per minestra,[11] in forma di larghe tagliatelle', e – forse sulla sua scia – anche i posteri (U, KK1, Ki, H) la fanno equivalere a = metélt 'tagliatelle, fettuccine'. Senza ulteriori passaggi intermediari l'errore viene poi corretto in H & J (siamo nell'anno 2000), dove appare per la prima volta l'equivalente *lasagna* come prestito dall'italiano. Più cauta è invece la posizione adottata in F & V e H & I, dove si fa ricorso a descrizioni esplicative metalinguistiche.

La *pizza* è un'altra volta = sütemény 'dolce' per Kő (con l'aggiunta *zsíros* 'grasso'); per KK1 e Ki si tratta invece di = lángos 'una specie di focaccia' (che nelle cucine ungheresi viene fritta in abbondante olio e ha solo una parca guarnizione); H riprende la parola = lángos aggiungendo anche = lepény 'schiacciata'; è con H & J che la *pizza* diventa = pizza, quindi la parola arriva in Ungheria (anno 2000); nell'uso della sola parola italiana come equivalente in F & V e H & I si vede, infine, che *pizza* è ormai saldamente radicato nell'ungherese come prestito e non ha quindi bisogno di adattamenti o di spiegazioni.

Nella direzione contraria è la parola gulyás a mostrare un consolidamento dello stesso tipo: dopo i tentativi di equivalenza con = guazzetto (G & S), = spezzatino (KK2, KKJ e con l'aggiunta alla ungherese ancora in KKJ & J) è la volta di = gulasch, che può essere ritenuto oggi un prestito ormai radicato nell'italiano.[12]

Nel caso di pogácsa 'piccola focaccia rotonda e salata' (F & V) sono fuorvianti le soluzioni con = focaccia (G & S, KK2, KKJ, senza le precisazioni necessarie) e con = schiacciata (KKJ, KKJ & J), anche se l'equivalenza in quest'ultimo dizionario è relativizzata, giustamente, dall'abbreviazione kb 'all'incirca'.

Anche nel caso di rétes 'strudel' è fuorviante l'uso (da parte di KKJ e KKJ & J) come equivalente di = millefoglie (tipo di pasta che viene usata per dolci di diverso tipo).

Gli esempi riportati in questo paragrafo e altri casi nel corpus attestano che un'analisi diacronica nell'ambito dei dizionari bilingui può rivelare anche le fasi del processo in cui una parola straniera entra nel lessico di un'altra lingua. I dati raccolti rivelano che nell'ungherese di oggi possono essere considerati prestiti dall'italiano pizza, chianti e (come calco-traduzione) marszalai bor, e sono probabili candidati mortadella e tortellino;[13] nell'altra direzione sono prestiti (gli ormai storici) gulasch e tokaj(i).[14]

11 Per Kőrösi *minestra* si riferisce ai 'vari tipi di pasta (come maccheroni ecc.) cotti in acqua, pietanza consumata come prima portata nei pranzi degli Italiani'.

12 La parola entra nell'italiano dal ted. (in cui è di origine ungh.) nel 1892 (*Vocabolario Zingarelli della lingua italiana* 2008).

13 Benché non ancora registrato dai dizionari analizzati, negli ultimi anni si nota una grande diffusione in Ungheria anche di *panettone*.

14 Ungh. *gulya* 'mandria di bovini' → *gulyás* 'mandriano' (con sottinteso *hús* 'carne'), quindi: 'la carne del mandriano'; nell'italiano attraverso il tedesco, 1892 (*Zingarelli* 2008). Ungh. *Tokaj* 'località vinicola sull'alto Tibisco' → *tokaji, tokaj* 'vino biondo oro molto pregiato ivi prodotto', 1709 (*Zingarelli* 2008).

3.5 Strategie traduttive e vocabolari attivi / passivi

In base ai dati dell'analisi può essere ribadito che il rapporto tra le strategie traduttive adoperate e la direzione delle lingue nei vocabolari risulta contraddittorio. Da una parte, il largo uso dei prestiti e dei forestierismi di origine italiana quali equivalenti nell'ungherese è, in base alle affermazioni della letteratura specifica, segno dell'esotizzazione, la quale si pone però in contrasto con lo scopo del dizionario passivo italiano-ungherese che dovrebbe essere orientato verso la localizzazione (e viceversa, v. Fábián 2013: 70). Dall'altra parte, la presenza e gli effetti di una cultura più prestigiosa (quale, nel nostro caso, l'italiana) nei confronti di un'altra (l'ungherese) sono sempre di tipo esotizzante: ciò è attestato dal grande numero di vocaboli di origine italiana nell'ungherese che, di conseguenza, sono presenti anche nei dizionari italiano-ungheresi (v. p.es. House 2004, Klaudy 2013: 88).

4 Conclusione

L'analisi diacronica sui realia nei dizionari italiano / ungheresi ha non solo presentato l'evoluzione delle strategie traduttive adoperate dai lessicografi ma anche le loro preferenze soggettive, ed ha inoltre contribuito alla descrizione delle fasi dell'addomesticamento dei rispettivi prestiti in L2. Per poter arrivare a conclusioni ancor più precise al riguardo, converrà estendere l'esame ad una cerchia tematicamente più vasta dei realia (p.es. espressioni del linguaggio amministrativo, nomi di parentela ecc.) e raffinare ulteriormente la metodologia dell'esame.

5 Bibliografia

5.1 I Vocabolari analizzati (in ordine cronologico)

Kő = Kőrösi S. (1912). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Budapest: Lampel.
G & S = Gelletich, V. & Sirola, F. (1914). *Magyar-olasz szótár* [Dizionario ungherese-italiano]. Fiume: Mohovich.
U = Urbanek, S. (1915). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Fiume: Mohovich.
KK1 = Koltay-Kastner, J. & Virányi, E. & Szabó, M. (1940^{2}). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Pécs: Danubia.
KK2 = Koltay-Kastner, J. & – Virányi, E. & Szabó, M. (1943^{2}). *Magyar-olasz szótár* [Dizionario ungherese-italiano]. Pécs: Danubia.
Ki = Király, R. (1944). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Budapest: Szent István Társulat.
H = Herczeg, Gy. (1978^{3}). *Olasz-magyar szótár* I-II. [Vocabolario italiano-ungherese]. Budapest: Akadémiai Kiadó.
KKJ = Koltay-Kastner, J. (1981). *Magyar-olasz szótár* I-II. [Vocabolario ungherese-italiano]. Budapest: Akadémiai Kiadó.

H & J = Herczeg, Gy. & Juhász, Zs. (2000). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Budapest: Akadémiai Kiadó.
KKJ & J = Koltay-Kastner, J. & Juhász, Zs. (2000). *Magyar-olasz szótár* [Dizionario ungherese-italiano]. Budapest: Akadémiai Kiadó.
F & V = Fábián, Zs. & Vásárhelyi, J. (2001). *Olasz-magyar és magyar-olasz útiszótár* [Vocabolario italiano-ungherese ed ungherese-italiano per turisti. Ed. riveduta ed aggiornata]. Budapest: Akadémiai Kiadó.
H & I = Hessky, E. & Iker, B. (2011). *Olasz-magyar útiszótár* [Vocabolario italiano-ungherese ed ungherese-italiano per turisti]. Szeged: Grimm Kiadó.

5.2 Saggi

Éber, B. & Fata, I. (2010). Reáliák és fordítói stratégiák vizsgálata német középfokú iskolatípusok és végzettségek magyarra fordítása kapcsán [Analisi dei realia nelle strategie traduttive dei concetti relativi ai tipi di scuola e ai nomi dei diplomi tedeschi]. In J. Dróth (ed.) *Szaknyelv és szakfordítás. Tanulmányok a szakfordítás és a fordítóképzés aktuális témáiról 2009–2010.* Gödöllő: Szent István Egyetem, pp. 40-56. Accesso: http://tti.gtk.szie.hu/datadir/content/file/idegen_nyelvi_k%C3%A9pz%C3%A9s_%C3%A1llom%C3%A1nyai/kiadvanyaink/szakford_book_2010_v%C3%A9gleges.pdf) [10/03/2014]
Fábián, Zs. (2011). La lessicografia ungherese / italiana. In Zs. Fábián (ed.) *Hungarian Lexicography I. Bilingual Dictionaries*. Budapest: Akadémiai Kiadó, pp. 93-108. [Serie Lexikográfiai füzetek 5.]
Fábián Zs. (2013). Reáliák a kétnyelvű szótárban régen és ma: megfigyelések és következtetések [I realia nei vocabolari bilingui ieri e oggi: osservazioni e conclusioni]. In V. Bárdosi (ed.) *Reáliák a lexikológiától a frazeológiáig.* Budapest: Tinta Könyvkiadó, pp. 61-72.
Fata, I. (2010). *Német nyelvű reáliák megjelenítése kétnyelvű szótárakban – esettanulmány*. [La presentazione dei realia tedeschi in dizionari bilingui]. Short paper al congresso della MANYE a Debrecen, 26-28 agosto 2010. Accesso: http://www.manyexx.unideb.hu/sites/default/files/manyexx_absztraktok.pdf [10/03/2014]
Heltai, P. (2001). Lexikai átváltási műveletek irodalmi és szakfordításban [Strategie traduttive lessicali nelle traduzioni letterarie e speciali] In *Fordítástudomány* X (1), pp. 5-17.
Heltai, P. (2013). Kultúraspecifikus kifejezések és reáliák [Espressioni culturo-specifiche e realia]. In *Fordítástudomány*, XV (1), pp. 32-53.
House, J. (2004). Culture-specific elements in translation. In H. Kittel et al. (eds.) *Übersetzung / Translation / Traduction.* Berlin: De Gruyter, pp. 494-504. [Serie HSK 26.1.]
Klaudy, K. (2005). *Bevezetés a fordítás gyakorlatába* [Introduzione alla prassi della traduzione]. Budapest, Scolastica.
Klaudy, K. (2013). Nyelvi és kulturális aszimmetria a reáliák fordításában [Assimetria linguistica e culturale nella traduzione dei realia]. In V. Bárdosi (ed.) *Reáliák a lexikológiától a frazeológiáig.* Budapest: Tinta Könyvkiadó, pp. 85-92.
Kujamäki, P. (2004). Übersetztung von Realienbezeichnungen in literarischen Texten. In H. Kittel et al. (eds.) *Übersetzung / Translation / Traduction.* Berlin: De Gruyter, pp. 920-925. [Serie HSK 26.1.]
Lendvai, E. (1986). *A „lefordíthatatlan elem" megfeleltetési lehetőségei* [Possibilità di equivalenza degli "elementi intraducibili"] Tesi per il titolo di "candidato" dell'Accademia delle Scienze Ungherese. Università di Pécs, HU.
Marello, C. (1989). *Dizionari bilingui*. Bologna, Zanichelli.
Mujzer-Varga, K. (2009). *Honosítás és idegenítés Örkény Egyperces novelláinak fordításaiban* [Localizzazione ed esotizzazione nelle traduzioni di "Racconti di un minuto" dello scrittore Örkény]. Tesi per il titolo di PhD. Università "Eötvös" di Budapest, HU. Accesso: http://doktori.btk.elte.hu/lingv/mujzervargakrisztina/tezis.pdf [10/03/2014]

Osimo, B. (2003). *Il manuale del traduttore.* Milano: Hoepli.
Rey, A. (1986). Les écarts culturels dans les dictionnaires bilingues. *Lexicographica* 2: 33-42.
Rey, A. (1991). Divergences culturelles et dictionnaire bilingue. In F. J. Hausmann, F.J. et al. (eds.) *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin: De Gruyter, pp. 2865-2870. [Serie HSK 5.3.]
Svensén, B. (2009). *A Handbook of Lexicography*. Cambridge: Cambridge University Press.
Szabó, H. (2008). A szótári ekvivalencia és a fordítási ekvivalencia viszonya. [Sul rapporto dell'equivalenza lessicografica e dell'equivalenza traduttiva] In *Fordítástudomány* X (1), pp. 61-70.
Tellinger, D. (2003). A reáliák fordítása a fordító kulturális kompetenciája szemszögéből. [La traduzione dei "realia" dal punto di vista della competenza culturale del traduttore]. In *Fordítástudomány* V (2), pp. 58-70.
Tomaszczyk, J. (1984). The culture-bound element in bilingual dictionaries. In R. Hartmann (ed.) *LEXeter '83 Proceedings*. Tübingen: Niemeyer, pp. 289-297. [Lexicographica. Series Maior.]
Uzonyi, P. (2006). A forrásnyelvi és célnyelvi adatok ekvivalenciájának nehézségeiről [Sulle difficoltà delle equivalenze tra i dati di partenza in L1 e quelli di arrivo in L2]. In T. Magay (ed.) *Szótárak és használóik*. Budapest: Akadémiai Kiadó, pp. 117-126. [Serie Lexikográfiai füzetek 2.]

Appendice 1

***barbera**:* Kő: piemonti vörös bor-faj, U: –, KK1: –, Ki: borfajta, H: Asti környéki vörös bor, H & J: *bor*[15] <piemonti vörösbor>, F & V: –, H & I: –

***cannelloni**:* Kő: a makaróninál vastagabb csöves tészta, U: –, KK1: –, Ki: –, H: (hússal, sajttal stb. töltött) tésztarolád, palacsinta, H & J: *konyh*[16] [hússal, sajttal stb. töltött] tésztarolád], F & V: húsos töltött palacsinta, H & I: *ujjnyi hosszú tésztacsöveh, melyeket töltve készítenek el*

***chianti**:* Kő: 2. Chianti-bor, U: –, KK1: –, Ki: toszkánai vörösborfajta, H: toszkán vörösbor, chianti, H & J: <toszkán vörösbor> chianti, F & V: chianti *(toszkán vörösbor)*, H & I: *(az azonos nevű toszkán vidékről származó vörösbor)* chianti

***gorgonzola**:* Kő: gorgonzóla (sajt), U: gorgonzola (sajt), KK1: –, Ki: sajtfajta, H: gorgonzola sajt, H & J: *konyh* gorgonzola sajt, F & V: *(sajtféle)*, H & I: *Lombardiából származó kékpenészes sajt*

***grissino**:* Kő: hosszúkás ropogós sütemény Piemontban, U: –, KK1: –, Ki: ropogós kenyérrúd, H: hosszú, vékony rúdkenyér, H & J: *konyh* ropi <hosszú, vékony ropogtatnivaló>, F & V: *(hosszú és vékony rúdkenyér)*, H & I: *kenyértésztából készült, ropogós pálcika*

***lasagna**:* Kő: széles szalagokra vagdalt levestészta v. más minesztrához való tészta, U: szalagmetélt, KK1: metélt, Ki: szélesmetélt, H: széles metélttészta, H & J: lasagna; *lasagne al forno* csőben sült lasagne v. rakott tészta, F & V: *(lerakott tésztaféleség)*, H & I: *lapos, széles, téglalap alakú tésztalapok, melyekből rakott ételeket készítenek*

***marsala**:* Kő: 2. marszálai bor, U: –, KK1: marsalai bor, Ki: marsalai bor, H: (magas szesztartalmú) marszalai fehér bor, H & J: [magas szesztartalmú] marszalai fehér likőrbor, F & V: –, H & I: –

***mascarpone**:* Kő: –, U: –, KK1: –, Ki: –, H: *észol*:[17] (Camemberthez hasonló) lombardiai puha sajt, H & J: mascarpone [lombardiai sajt], F & V: –, H & I: *tejszínes, fehér krémsajt*

***mortadella**:* Kő: disznósajt, U: májas hurka; mortadella, KK1: disznósajt, mortadella, Ki: disznósajtféle, H: mortadella (felvágott), H & J: *konyh* mortadella [felvágott], F & V: mortadella *(felvágottféle)*, H & I: –

***pandoro**:* Kő: –, U: –, KK1: –, Ki: édestésztafajta, H: veronai kuglóf, H & J: *konyh* <karácsonyi veronai kuglóf>, F & V: kuglóf *(karácsonyi édesség)*, H & I: *kuglófhoz hasonló, veronai eredetű karácsonyi sütemény*

***panettone**:* Kő: lisztből, cukorból, sáfrányból készített és sörrel kelesztett sütemény, U: –, KK1: –, Ki: édestésztaféle, H: mazsolás, cukrozott gyümölcsös milánói kalács, H & J: *konyh* panettone [karácsonyi

15 *bor* 'vino'
16 *konyh* 'termine gastronomico'
17 *észol* 'settentrionale'

milánói kuglóf], F & V: kuglóf *(karácsonyi édesség)*, H & I: *kuglófhoz hasonló, milánói eredetű karácsonyi sütemény*

***panforte**:* Kő: mandulából, kakaóból stb. készített borsos sütemény, U: –, KK1: sienai mézeskalács, Ki: édestésztafajta, H: lapos, kerek, sienai sütemény, H & J: *konyh* <lapos, kerek sienai sütemény>, F & V: *(sienai édesség)*, H & I: *gyümölcskenyérhez hasonló sienai karácsonyi sütemény*

***pastasciutta**:* Kő:[18] (leves helyett tálalt) makaróni, U: –, KK1: –, Ki: makaróni, H: száraz tészta, főtt tészta, H & J: száraztészta [élelmiszer], főtt tészta [fogás], F & V: főtt tészta, H & I: (száraz)tészta

pizza Kő: 2. zsíros sütemény, U: –, KK1: lángos, Ki: lángos, H: lángos, lepény, H & J: 1. pizza; ~ *alla napoletana* [...] 2. lángos; lepény, F & V: pizza, H & I: pizza

raviolo Kő: vajjal, túróval, fűszerszámmal és vagdalt hússal töltött toklóhúsból készült étel, U: –, KK1: –, Ki: étel (vagdalthús- és túrósgombócból), H: húsos táska; derelye, H & J: *konyh* húsos táska, derelye, F & V: húsos derelye/táska, H & I: *hússal, zöldséggel, ricottával töltött tésztabatyu*

tortellino Kő: tyúkhússal töltött, levesbe való táska, U: táska (levesbe), KK1: –[19], Ki: húsos táska (étel), H: darált hússal töltött táska, H & J: *konyh* tortellini [sonkával, hússal töltött tészta], F & V: tortellini *(töltött tésztagyűrű)*, H & I: *hússal, sajttal töltött gyűrű alakú tésztabatyu*

Appendice 2

***beigli**:* G & S: –, KK2: –, KKJ: –, KKJ & J: *konyh* rotolo dolce ai semi di papavero o ai noci [sic!], F & V: –, H & I: –

***fröccs**:* G & S: –, KK2: –, KKJ: due decilitri di vino con un decilitro di acqua di seltz, KKJ & J: <vino e acqua gasata>, F & V: *(due decilitri di vino con un decilitro di acqua di seltz)*, H & I: spritz

***gulyás**:* G & S: guazzetto, KK2: spezzatino, KKJ: gulasch, spezzatino alla ungherese, KKJ & J: gulasch, spezzatino alla ungherese, F & V: gulasch, H & I: gulasch

halászlé: G & S: salsa/zuppa di pesce, KK2: zuppa di pesce, KKJ: minestra/brodetto di pesce (all'ungherese), KKJ & J: *konyh* zuppa di pesce (all'ungherese), F & V: zuppa di pesce, minestra di pesce, H & I: zuppa di pesce

***kadarka**:* G & S: –, KK2: –, KKJ: kadarka (uva, vino leggiero [sic!] rosso ungherese), KKJ & J: kadarka <uva/vino rosso leggero ungherese>, F & V: –, H & I: –

***kovászos uborka**:* G & S: cetriolo sotto aceto, KK2: –, KKJ: cetriolo lievitato, KKJ & J: cetriolo in salamoia a base di lievito, F&V:–,H & I: –

***lángos**:* G & S: crostino, KK2: pizza, KKJ: pizza, *toszk*: schiacciata, *róm*: piada, KKJ & J: *toszk*: schiacciata <pasta lievitata fritta in abbondante olio>, F & V: *(una specie di schiacciata)*, H & I: –

***paprikás csirke**:* G & S: pollo in umido colla paprica, KK2: –, KKJ: pollo con intingolo di paprica, spezzatino con paprika, KKJ & J: pollo alla paprika, spezzatino di pollo, F & V: pollo all'arrabbiata, pollo all'ungherese, H & I: –

pogácsa: G & S: focaccia, KK2: focaccia, KKJ: schiacciata, focaccia, KKJ & J: *konyh kb*[20] schiacciata, focaccia, F & V: (piccola focaccia rotonda e salata), H & I: –

***rétes**:* G & S: pasta avvoltolata con ripieno, KK2: strudel, KKJ: *kb* millefoglie, KKJ & J: *konyh* strudel, millefoglie, F & V: strudel, H & I: strudel[21]

szaloncukor: G & S: –, KK2: –, KKJ: confetto/fondente incartato, KKJ & J: <confetto/fondente incartato per ornare l'albero di Natale>, F & V: –, H & I: –

***tejberizs**:* G & S: riso in latte, KK2: riso col latte, KKJ: riso col latte, KKJ & J: riso al latte, F & V: *(riso cotto in latte dolce)*, H & I: riso al latte

---

18 Nel lemma di *asciutto*.

19 Come lemma abbiamo: *tortello* húsos táska.

20 *kb* 'all'incirca'

21 Nello stesso tempo nei lemmi con collocazioni appaiono anche altre soluzioni: G & S: *káposztás rétes* pasticcio di cavolo, *almás rétes* dolce di mele, KKJ: *túrós rétes kb* pasta sfoglia di ricotta, KKJ & J: *túrós rétes* pasta sfoglia di ricotta, strudel alla ricotta.

***tejföl/tejfel**:* G & S: panna, crema, fior di latte, KK2: fiore di latte, KKJ: fiore di latte, crema, KKJ & J: fiore di latte, crema/panna acida, F & V: panna acida, H & I: panna acida
***tokaji**:* G & S: (bor) (vino) di Tokaj, KK2: –, KKJ: tokay, tocai, KKJ & J: tokaj, tocai, F & V: tocai , H & I: –
***töltött káposzta**:* G & S: –, KK2: –, KKJ: –, KKJ & J: cavoli/crauti ripieni, F & V: cavoli ripieni, H & I: –
***túrós csusza**:* G & S: –, KK2: –, KKJ: –, KKJ & J: –, F & V: *(tagliatelle con ricotta e panna acida)*, H & I: –

Quello che i dizionari possono fare: l'esempio dei Dizionari di Tedesco (Giacoma/Kolb – Zanichelli/Klett)

Luisa Giacoma
Università di Torino
luisa.giacoma@unito.it

Abstract

Partendo dall'affermazione di Hausmann (1993) che dichiara la non apprendibilità del lessico e che vede nei dizionari l'unica possibilità di salvezza per l'apprendente, si avverte l'esigenza di indagare su cosa possa fare la lessicografia per far fronte a questa gravosa responsabilità. Da una parte bisogna educare il lettore ad un uso proficuo del dizionario, mentre dall'altra è necessario innovare i dizionari, affinché possano aiutare l'apprendente anche nell'uso attivo della lingua. Per potersi esprimere adeguatamente in L2 egli deve infatti saper padroneggiare i legami semantici, sintattici e pragmatici esistenti tra le parole ed essere consapevole dei notevoli vincoli che questi pongono alla libertà combinatoria del parlante (Lo Cascio 1997). Se teoricamente il parlante può abbinare liberamente le parole, in realtà solo poche combinazioni sono possibili, poiché esse sottostanno a limitazioni sintattiche, semantiche, enciclopediche e combinatorie dettate dall'uso e dal contesto. Un'analisi dei principali problemi rimasti irrisolti nella lessicografia tradizionale, orientata per lo più verso l'uso passivo della lingua, ha portato allo sviluppo di un nuovo modello lessicografico alla base dei dizionari di Tedesco a cura di Luisa Giacoma e Susanne Kolb e pubblicati dalla casa editrice italiana Zanichelli e tedesca Klett, che costituiscono una risposta concreta alle nuove esigenze degli apprendenti.
Keywords: Dizionari; Tedesco; Italiano

1 Introduzione

Quando l'aereo della compagnia di bandiera italiana, poco dopo il decollo, abbandonò l'assetto della salita vertiginosa per assumere quello orizzontale della crociera, mi venne servito un tè che mi fece trasalire. Sul bicchierino di carta bianca campeggiavano a caratteri cubitali due scritte verdi e rosse a completamento dei colori nazionali, una delle quali recitava: IL NOME DELLA BEVANDA PIÙ DIFFUSA DEL MONDO PUÒ ESSERE SCRITTO IN ITALIANO IN DUE MODI TÈ O THÈ. Rilessi nuovamente, ma l'errore continuava a rimanere lì in bella evidenza a fine frase. Ora, ammettiamo pure che qualcuno possa non sapere che in italiano la più nota bevanda inglese si scrive "tè" oppure "the", e che quindi la "e" possa scegliere se andare a braccetto con l'"h" oppure portare un accento calcato sulle ventitrè come un cappello, ma se si decide di scrivere una frase come quella, dove l'oggetto è proprio la corretta grafia di una parola, non bisognerebbe come minimo controllarla sul dizionario? In un caso

come questo, ovviamente, nemmeno il migliore dei dizionari può fare qualcosa, ma in moltissimi altri casi può fare invece la differenza.

2 Il ruolo del dizionario nella didattica delle lingue

Se Hausmann (1993: 471) afferma che non si può imparare il lessico di una lingua perché è sterminato, totalmente idiomatico e caotico e vede nei dizionari l'unica via di salvezza, diventa allora cruciale disporre di strumenti ben costruiti, che forniscano per ogni lemma non solo le informazioni morfologiche, semantiche, sintattiche e i traducenti usualmente reperibili nei dizionari, ma anche una sorta di mappa delle possibilità combinatorie del lemma, con tutte le informazioni sull'uso solitamente appannaggio dei madrelingua. Il recente spostamento del focus lessicografico dalla descrizione della lingua in sé alle esigenze del lettore ha avuto come conseguenza evidenti vantaggi per l'apprendimento e il passaggio dell'interesse dall'uso prevalentemente passivo a quello attivo.

L'intero corpo di una lingua, pur entro limiti stabiliti, si trova solo in un dizionario, che va quindi sempre più concepito "come strumento centrale per la conoscenza e la descrizione sistematica della lingua nel suo pieno uso, superando la sua vecchia immagine di strumento per conoscere il 'significato' (o poco più) delle parole" (Sabatini 2008: 112).

Nonostante vi siano ancora posizioni antitetiche sull'apprendibilità del lessico, nella didattica delle lingue, dopo un periodo di ingiustificato ostracismo, è ormai indiscusso il ruolo centrale del dizionario bilingue come strumento per l'apprendimento. Resta però ancora molto da migliorare sul fronte delle possibilità del lettore di utilizzarlo efficacemente. Da una parte viene fatto troppo poco per insegnare a usare bene i dizionari, mentre dall'altra il contributo della metalessicografia non si è ancora sufficientemente diffuso nella pratica lessicografica e quindi sono ancora pochi i dizionari centrati sui bisogni effettivi dell'apprendente.

Come è già stato evidenziato in ricerche sul tema (Giacoma 2011, 45-46), la conoscenza "spontanea" dell'uso del dizionario è uno dei tanti pregiudizi molto diffusi su questi strumenti: varrebbe invece la pena di dedicare, possibilmente fin dall'inizio del corso, qualche ora per avviare gli studenti ad un utilizzo efficace del dizionario. Poiché non si può vedere quello che non si conosce, come affermò il fondatore della casa editrice per guide da viaggio Dumont, si potrebbe sacrificare qualche ora di spiegazione riservata ad argomenti grammaticali minori e sfruttare questo tempo a vantaggio di un uso migliore dei dizionari, dato l'impatto che essi hanno sull'apprendimento delle lingue straniere.

3 Dalla parte del lettore: alcuni problemi irrisolti nella lessicografia bilingue tradizionale

I maggiori problemi della lessicografia bilingue tradizionale del secolo scorso, ma ancora presenti nelle successive edizioni di dizionari di Tedesco-Italiano concepiti in quel periodo, sono da attribuirsi principalmente alla mancata ricezione dei risultati ottenuti negli studi di linguistica e di lessicografia contrastiva, risultando così inadeguata a soddisfare le sempre crescenti necessità dei lettori, soprattutto di quelli di livello più avanzato, che si trovano a usare la lingua in modo molto più (inter)attivo rispetto a quanto avvenisse in passato, quando i dizionari venivano consultati quasi esclusivamente per la decodificazione dalla L2 verso la propria. I limiti più evidenti dei dizionari tradizionali sono (Giacoma 2011):

- l'insufficiente o inadeguata registrazione delle collocazioni,[1] che non compaiono come tali, ma come esempi alle singole accezioni o nel blocco fraseologico al fondo della voce, se non addirittura in luogo delle spiegazioni
- l'inadeguata differenziazione tra i diversi equivalenti della L2, che vengono generalmente elencati uno dopo l'altro, senza alcun elemento che indichi quale sia da preferire all'interno di un certo contesto, come se fossero sinonimi intercambiabili in tutti i contesti, anche se gli apprendenti tendono a fare molti errori proprio nel momento in cui costruiscono frasi o parti di esse combinando tra loro elementi che non possono essere abbinati
- la mancanza di indicazioni sistematiche e ben evidenziate sull'intorno sintattico del lemma, nonostante la trasposizione in L2 delle strutture della propria lingua sia all'origine di molti errori
- l'inadeguata registrazione dei fraseologismi
- le scarse informazioni sulla morfologia delle parole tedesche, fonte di frequenti dubbi per l'apprendente italofono.

Naturalmente ci sarebbero anche altri problemi da discutere, ma si rimanda ad altro luogo per ragioni di spazio.

4 L'evoluzione della voce lessicografica nei dizionari di Tedesco-Italiano

L'avvento nella pratica lessicografica di corpora come *COSMAS II* dell'Institut für Deutsche Sprache di Mannheim, consultabile gratuitamente in rete, ha notevolmente migliorato le basi materiali sulle quali possono essere costruiti oggi i dizionari, permettendo l'accesso a dati statisticamente significa-

1 Qui le collocazioni vengono intese in senso ampio, adatto a scopi lessicografici. Si confronti a questo riguardo la definizione data nel dizionario per apprendimenti del tedesco *LGWDaF* = Dieter Götz et al., *Langenscheidts Großwörterbuch Deutsch als Fremdsprache*, Langenscheidt, Berlin 1993, S. XX „[…] collegamenti tipici di più parole che costituiscono un'unità sintattica […] poiché mostrano […] 'partner' con il quale il lemma occorre frequentemente„ [trad. LG].

tivi sull'uso reale della lingua. Anche le acquisizioni della metalessicografia sono potenzialmente di grande aiuto, ma il dialogo tra linguisti e lessicografi è una conquista relativamente recente. Nied Curcio (2006, 61), nella sua analisi dei dizionari di Tedesco, lamenta che i dizionari di tipo tradizionale come il Sansoni e il DIT "non si basano sugli studi della linguistica contrastiva, ma si limitano ad essere un insieme di informazioni utili solo per la consultazione". Già nel 1989 Marello criticava la scarsa ricaduta sui dizionari bilingui di Tedesco-Italiano del vivace dibattito metalessicografico sulle valenze che aveva avuto luogo in Germania.

5 Possibili soluzioni e loro realizzazione

I dizionari di Tedesco a cura di Luisa Giacoma e Susanne Kolb, vale a dire l'intera serie di dizionari scritti dalle autrici nell'arco degli ultimi vent'anni e pubblicati dalla casa editrice italiana Zanichelli e da quella tedesca Klett (d'ora in avanti GK), nati sulla base di approfondite riflessioni sui limiti della lessicografia bilingue di Tedesco-Italiano degli anni '90, hanno cercato di risolvere tali problemi con l'introduzione di alcuni strumenti innovativi, offrendo risposte concrete alle necessità degli apprendenti. A tal fine è risultato particolarmente utile poter contare sulla teoria della valenza come riferimento teorico e valutare le applicazioni pratiche in campo lessicografico, che proprio in quegli anni venivano realizzate in Germania.

Chi ascolta, scrive, legge, parla o traduce una lingua si relaziona nella quasi totalità dei casi con testi, mentre nel dizionario trova parole singole. A questo proposito Sabatini (2008: 112) afferma che "di questa globalità, accolta in un unico contenitore anche se scomposta nella lemmaticità alfabetica, c'è bisogno per cogliere il funzionamento della lingua, incrociando le diverse prospettive (morfologica, sintattica, semantica, testuale [...])".

Per riuscire a superare il gap tra il testo e le parole è necessario che la voce lessicografica dia il maggior numero di informazioni sull'intorno lessicale e sintattico della parola, su quali parole cioè si trovano frequentemente in compagnia del lemma (i *collocatori*) e su come lemma e *collocatori* si combinano tra loro (con o senza preposizione, ecc.). Il modello valenziale di descrizione del comportamento sintattico della singola parola ha mostrato enormi potenzialità nella sua applicazione alla voce lessicografica, con importanti risvolti per l'apprendente. Inoltre esso è particolarmente utile quando si mettono a confronto lingue diverse perché evidenzia le divergenze e/o convergenze, che sono spesso fonte di errore per gli apprendenti, soprattutto nei momenti di codificazione in L2 (Curcio 1999).

A questo proposito Fischer e Mollica (2012) osservano che l'approccio valenziale, così come quello costruzionistico, vede i lessemi e le costruzioni in stretta relazione tra loro, ma ciò che distingue il primo approccio dal secondo è la prospettiva "bottom-up", che parte cioè dal lessema per arrivare alla costruzione. Si può quindi trarre la conclusione che proprio questo sia ciò che serve alla lessicografia, che parte dai singoli lemmi per descrivere costruzioni, frasi, testi e la lingua in genere. In Germania si hanno i primi dizionari valenziali già alla fine degli anni '60 con il *Wörterbuch zur Valenz und Distributi-*

on deutscher Verben di Helbig/Schenkel del 1969 (Nied Curcio 2012, 175). Non stupisce pertanto che le prime applicazioni lessicografiche della teoria della valenza filtrino in Italia proprio grazie ai contributi di germanisti italiani (Bianco, Curcio, Soffritti) o romanisti attivi in Germania (Blumenthal, Rovere) sotto forma di dizionari valenziali su base contrastiva, come quello dei verbi tedeschi e dei loro equivalenti italiani (Bianco 1996), oppure quello dei verbi italiani con traducenti tedeschi (Blumenthal/Rovere 1998), o ancora della lingua (tedesca e italiana) parlata (Curcio 1999), degli aggettivi (Soffritti, DIVA 2005) e dei circa 3000 lemmi del lessico di base presenti in ELDIT. Ma una nuova generazione di dizionari generali si stava affacciando nel panorama della lessicografia bilingue di Tedesco-Italiano. I dizionari GK nascono in un periodo di grande innovazione lessicografica, come si evince dall'uscita in Germania nel 1993 del primo dizionario monolingue per apprendenti *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (LGWDaF) e del dizionario bilingue *Pons Großwörterbuch Französisch Deutsch* (PONS) nel 1996. Entrambi applicano la teoria della valenza per un pubblico di apprendenti stranieri e hanno costituito un ottimo modello di riferimento per i dizionari GK. Essi sono i primi nei quali la teoria della valenza trova spazio per la descrizione di *tutta* la lingua italiana e tedesca. Da questo punto di vista essi possono essere considerati, in quanto dizionari valenziali, *dizionari di studio*, vale a dire strumenti articolati e sistematici di riflessione sulle proprietà sintattiche e semantiche della lingua.
Seguendo il modello del LGWDaF e del PONS, essi introducono una sorta di *sintassi della parola* basata sulle teorie grossiane (1967, 1975) che, al compito svolto anche dai dizionari tradizionali di descrivere con puntigliosa precisione ogni singola parola, aggiungesse quello non meno importante di dare informazioni esplicite su come le parole possono o debbono essere combinate tra loro. Sono proprio queste informazioni sintagmatiche ad essere spesso carenti nei dizionari di tipo tradizionale. Il metodo scelto per realizzare questa *sintassi della parola* è stata l'introduzione sistematica di due nuove categorie: i *collocatori* e le *formule di struttura*, entrambe *desiderata* della lessicografia teorica (Fontenelle 1997).

5.1 I collocatori

Uno degli aspetti meno soddisfacenti dei dizionari di tipo tradizionale è la scarsa registrazione di *collocatori* e soprattutto il modo confuso col quale questa avviene. A questo riguardo Marello/Rovere (1999: 198) sottolineavano, due anni prima dell'uscita del primo dizionario GK, la necessità di indicare le collocazioni come tali e di registrarne un numero maggiore nei dizionari. Per il lettore i principali vantaggi della registrazione dei collocatori da un punto di vista lessicografico sono: la differenziazione dei traducenti, la registrazione di un maggior numero di contesti tipici, una migliore strutturazione della voce, nonché una maggiore possibilità di controllo. Nei dizionari GK si è tenuto conto fin dall'inizio del progetto della possibilità di consultazione elettronica ed è pertanto stato dedicato un campo di ricerca ai soli collocatori. Nel GK (2014[3]) è sufficiente digitare una parola, ad esempio *cavallo,* all'interno di tale campo nella maschera della ricerca avanzata per ottenere la lista delle parole che ad essa più frequentemente si accompagnano. Sulla sinistra della maschera di ricerca vi è la colonna a

scorrimento con l'elenco dei collocatori numerati in ordine progressivo. Il numero totale (193) è in basso. Facendo scorrere l'elenco si possono leggere tutti i 193 collocatori della parola *cavallo*. Sulla destra è invece aperta la voce del collocatore evidenziato nella lista. La parola per la quale è stata fatta la ricerca appare all'interno di un riquadro evidenziato in rosso. Dizionari così concepiti permettono di rispondere a domande che rimarrebbero quasi sempre senza risposta con la consultazione, anche elettronica, di un dizionario tradizionale (Fontenelle 1997). Se vogliamo ad esempio sapere che verso fa il cavallo in un dizionario senza collocatori possiamo solo sperare di trovare un esempio al lemma *cavallo* o al lemma *verso*, ma non avremmo altri aiuti per arrivare a *nitrito*. Anche la ricerca elettronica di *cavallo* non ha dato esiti né nel DIT, né nel Sansoni.

In questo caso particolare osserviamo che anche nel GK manca *nitrito* nella lista dei collocatori di *cavallo*, ma si può facilmente risalire ad esso grazie alla presenza del verbo *nitrire*. Il potenziale lessicografico dei collocatori è però molto più evidente se desideriamo sapere, ad esempio, quali azioni può compiere un cavallo (*ansimare, cadere, calciare, caracollare, disarcionare, galoppare, imbizzarrire, impennare, inalberare, nitrire, rinculare, rompere, sbuffare, scartare, sgroppare, stramazzare, trottare, trotterellare, volteggiare ...*), o quando il cavallo è l'oggetto dell'azione (*addestrare, addomesticare, allenare, ammaestrare, azzoppare, bardare, cavalcare, domare, dopare, dressare, ferrare, frenare, guidare, imbrigliare, incavezzare, legare, montare, sbalzare, scendere, scommettere, sfiancare, spronare, strigliare ...*). Possiamo inoltre elencare quali caratteristiche può avere un cavallo semplicemente selezionando tra i collocatori gli aggettivi (*addestrato, baio, balzano, belga, berbero, bizzarro, bizzoso, domabile, drogato, favorito, fuoriclasse, furioso, irrequieto, matto, ombroso, perdente, pezzato, piazzato, pomellato, purosangue, ricalcitrante ...*) o trovare i termini corrispondenti alle singole parti del cavallo (*callo, cantone, coda, criniera, fianco, garretto, groppa, manto, nodello, retrotreno, schiena, spalla, unghia, ventre, zampa ...*) o al suo frame (*ambio, bardatura, briglia, paraocchi, sperone, ...*).

5.2 Le formule di struttura

Le *formule di struttura* (o semplicemente *strutture*) vengono riprese dai due dizionari di lingua tedesca serviti da modello (LGWDaF e PONS) e modificate per adattarle alle esigenze di un pubblico di apprendenti italiani. Anche nei dizionari GK esse riducono a formula le *reggenze* della grammatica tradizionale dando informazioni esplicite su come i collocatori e il lemma si combinano tra loro, ma diversamente dal LGWDaF e PONS, rinunciano ad esempio all'uso della terza persona singolare nei verbi. Tale scelta è motivata dal fatto che l'apprendente italiano ha dimestichezza con la forma all'infinito *scrivere qc a qu* e non a quella flessa *qu scrive qc a qu*. Quest'ultima formula ha però l'indubbio vantaggio di dare informazioni anche sul soggetto. Come ovviare, allora, a tale perdita di informazione? Si è optato per dare il soggetto come collocatore e differenziarlo dagli altri attraverso l'uso del corsivo. L'uso del grassetto corsivo mette in evidenza nei dizionari GK le *formule di struttura* permettendo al lettore di rendersi immediatamente conto del contesto sintagmatico del lemma (alla voce *cadere*: ***cadere da qc, cadere in qc, cadere su qu/qc, cadere a qc***, ecc.), differenziando spesso anche un'accezione dall'altra.

In riferimento alla necessità dell'apprendente di avere informazioni sui possibili collegamenti sintattici del lemma per poter produrre frasi grammaticalmente corrette, Schafroth (2004: 119) sottolinea come questo non costituisca generalmente un problema se il dizionario registra sistematicamente *formule di struttura* in modo esplicito. L'inserimento delle strutture all'interno della voce sono un'ulteriore aiuto per orientare il lettore e obbligano il lessicografo ad un lavoro minuzioso di strutturazione della glossa lessicografica. Esse hanno anche il vantaggio di essere facilmente controllabili perché date in modo esplicito, anziché implicitamente, come negli esempi.
Come si può evincere da quanto sopra riportato, la presenza delle *formule di struttura* non è solo un modo diverso di organizzare le informazioni, ma porta in primo piano la sintassi della parola enucleandola dal contingente e facendone brillare la quintessenza. Inoltre, a differenza di quanto possibile fare con gli esempi, le *formule di struttura* segnalano anche la facoltatività o meno delle *Ergänzungen* sia nella lingua di partenza che in quella di arrivo, con evidenti vantaggi per il lettore soprattutto quando una *Ergänzung* è obbligatoria in una lingua e facoltativa in un'altra.
Nied Curcio (2012), a proposito dei vantaggi dell'introduzione delle strutture nella voce lessicografica, afferma che, fornendo indicazioni esplicite, esse permettono un utilizzo più efficace del dizionario, diminuendo così in misura considerevole le possibilità di errore da parte di chi lo utilizza.

5.3 Corredo di maggiori informazioni sulle espressioni idiomatiche

Per quanto riguarda invece la registrazione delle *espressioni idiomatiche* Dobrovol'skij (2009) afferma che la soluzione fraseografica tradizionale di giustapporre nei dizionari bilingui ad una *espressione idiomatica* della lingua di partenza una *espressione idiomatica* della lingua di arrivo è insufficiente se non addirittura fuorviante, per via delle differenze semantiche, sintattiche e pragmatiche che spesso si rilevano all'interno di queste coppie di espressioni. Ad esempio a ***etw im Auge haben*** nel DIT viene semplicemente fatta seguire la traduzione 'avere qcs in testa', senza nessuna ulteriore informazione.
Ai fini di permettere al lettore di comprendere e usare correttamente le *espressioni idiomatiche* è necessario corredare queste ultime di informazioni riguardanti la semantica, la sintassi e la pragmatica delle stesse. Le regole per l'uso delle *espressioni idiomatiche*, e le restrizioni alle quali esse sottostanno, devono costituire un commento all'equivalente, soprattutto in considerazione del fatto che le equivalenze sono quasi sempre parziali.
Ogni *espressione idiomatica* andrebbe quindi descritta nelle sue peculiarità semantiche, pragmatiche e combinatorie (tra le quali quelle sintattiche) per permetterne un uso corretto all'interno di un contesto.
Un contributo in tale direzione è quello fornito alla lessicografia bilingue italo-tedesca dal metodo utilizzato nel Giacoma/Kolb, sia dal punto di vista dell'ordinamento che delle informazioni sintattiche, semantiche e pragmatiche che accompagnano le *espressioni idiomatiche*. Si confronti ora, ad esempio, la registrazione dell'*espressione idiomatica* ***etwas im Auge haben*** nel DIT e nel Giacoma/Kolb.

(1) DIT
etw im Auge haben 'avere qcs in testa'

(2) Giacoma/Kolb
etw im Auge haben {INSEKT, STAUBKORN}, avere qc nell'occhio;
(etw im Sinn haben) {SEINEN EIGENEN VORTEIL}, guardare solo a qc*;*
{EIN BESTIMMTES MODELL}, aver pensato a qc di preciso

Si può osservare come, in quest'ultimo caso, oltre all'informazione standard generalmente registrata nei dizionari, e cioè la traduzione, ve ne siano numerose altre che vale la pena di commentare. Per cominciare, il significato proprio viene aggiunto in prima posizione e corredato dei collocatori {INSEKT, STAUBKORN} e della traduzione 'avere qc nell'occhio'. Successivamente il significato idiomatico viene introdotto da una spiegazione *(etw im Sinn haben).* Per concludere i due equivalenti suggeriti dal dizionario 'guardare solo a qc' e 'aver pensato a qc di preciso' vengono accompagnati dal loro contesto in forma di collocatori {SEINEN EIGENEN VORTEIL} e {EIN BESTIMMTES MODELL}. Da questo semplice esempio si può dedurre quanto rischiosa sia l'equivalenza ***etw im Auge haben*** = 'avere qcs in testa' se non viene accompagnata da altro. Il lettore può infatti essere tratto in inganno in più modi, sia essendo indotto a pensare che il traducente proposto 'avere qcs in testa' possa valere anche per il significato proprio, sia non impedendogli di selezionare, sul piano idiomatico, un equivalente inadeguato al contesto. L'equivalente 'avere qcs in testa' non si adatterebbe infatti alla maggior parte dei contesti citati in (2).

5.4 Le tabelle di flessione

Le indicazioni morfologiche che tradizionalmente accompagnano il lemma sono solo una minima parte delle informazioni necessarie all'apprendente quando deve esprimersi in L2.
A partire dalla seconda edizione, grazie alla collaborazione con la società Canoo di Basilea, il dizionario GK dispone nella sua versione elettronica della tabella di flessione di tutti i lemmi tedeschi, comprendente la coniugazione dei tempi semplici per i verbi regolari e irregolari, il nominativo, genitivo, dativo e accusativo singolare e plurale per i sostantivi e la flessione completa degli aggettivi, anche al grado comparativo e superlativo. Naturalmente questo è stato possibile grazie ai supporti elettronici, che permettono di memorizzare un'enorme quantità di dati.

6 Orizzonti vicini: uno sguardo verso un futuro realizzabile

Il mestiere del lessicografo è certamente difficile ed estremamente faticoso, ma possiede l'indubbio fascino di poter migliorare continuamente negli anni la propria opera, apportando cambiamenti, a volte anche rilevanti, da un'edizione all'altra.

Oltre alle soluzioni realizzate finora e sopra descritte, ve ne è una ulteriore piccola parte nella terza edizione del Dizionario di Tedesco. Sono state aggiunte infatti 600 note sui falsi amici, note d'uso e di civiltà come aiuto concreto per comprendere persone di altre culture, comunicare con loro, recepire e produrre testi in lingua straniera, partecipare attivamente alla vita sociale e culturale di altri paesi. La versione cartacea è così stata sfruttata al massimo delle sue potenzialità e non ha permesso purtroppo ulteriori aggiunte, ma nella versione elettronica è stato possibile utilizzare lo spazio disponibile e le caratteristiche del supporto per arricchire il dizionario con la pronuncia sonora di tutte le parole del tedesco.

Cosa sarebbe utile realizzare ancora in una quarta, quinta, ecc. edizione? Poiché i dizionari sono testi complessi e ricchi di abbreviazioni, un modo per renderli di più facile lettura sarebbe la possibilità di leggere al passaggio del mouse la forma estesa delle abbreviazioni. Potrebbe essere infatti di aiuto, per uno studente italiano, leggere "lessico giuridico" passando sull'etichetta *jur*, che di per sé potrebbe essere non del tutto trasparente. Nella mia esperienza di docente mi è capitato spesso di constatare come proprio le etichette tendano ad essere saltate dal lettore, soprattutto se inesperto. Esse sono invece di grande importanza perché orientano nella lettura, evidenziando la struttura della voce e permettono una lettura verticale della glossa per arrivare rapidamente al punto che interessa, senza essere costretti a leggerla tutta.

Altro punto che, a mio avviso, dovrebbe essere migliorato data la sua importanza all'interno della lingua, è il trattamento della fraseologia. Oltre a quanto già fatto finora nei dizionari GK, sarebbe necessario innanzitutto evidenziare le espressioni idiomatiche come categoria a sé stante e chiaramente individuabile. Sarebbe inoltre auspicabile aggiungere informazioni ed esempi attivabili con un clic. Ecco sull'esempio dell'espressione idiomatica tedesca *etw im Auge haben* un confronto tra la situazione attuale e quella che potrebbe essere realizzata in future edizioni:

Giacoma/Kolb 2014[3]

> ***etw im Auge haben*** {Insekt, Staubkorn}, avere qc nell'occhio; *(etw im Sinn haben)* {seinen eigenen Vorteil}, guardare solo a qc; {ein bestimmtes Modell}, aver pensato a qc di preciso

Giacoma/Kolb (future edizioni)

> ***etw im Auge haben*** {Insekt, Staubkorn}, avere qc nell'occhio;
> ***jdn/etw im Auge haben*** *(jdn/etw im Sinn haben)*

{SEINEN EIGENEN VORTEIL}, guardare solo a qu/qc;
{EIN BESTIMMTES MODELL}, aver pensato a qu/qc di preciso

> Fenati/Rovere/Schemann (2009, 2011[2])
jn./etw. (immer/...) **im Auge haben** 1. tenere d'occhio qu, + non perdere d'occhio qu, stare con gli occhi addosso a qu 2. avere sott'occhio qc
3 avere ben presente qc

> 1. Die Kleine mußt Du beständig im Auge haben, sonst läuft sie auf die Straße und es gibt einen Unfall.
2. Seien Sie unbesorgt, ich habe das Karrussel immer im Auge. Da passiert nichts, was ich nicht sofort bemerke.
3. Wenn Du den Kerngedanken der Sache nicht immer im Auge hast, verlierst du dich in Einzelfragen.

Link 1

Institut für Deutsche Sprache (2010): COSMAS II. Corpus Search, Management and Analysis System. http://www.ids-mannheim.de/ cosmas2/projekt /referenz/korpora.html

Link 2

http://it.bab.la/dizionario/tedesco-italiano/im-Auge-haben

In questo caso specifico sarebbe da migliorare la differenziazione tra il significato proprio e quello idiomatico anche attraverso l'evidenziazione delle strutture diverse (come una semplice ricerca in *COSMAS II* dimostra, il significato idiomatico prevede infatti anche un complemento oggetto animato e quindi questo andrebbe aggiunto). Si potrebbe inserire poi subito dopo un campo apribile con un clic nel quale trovare materiale proveniente dai dizionari fraseologici bilingui come ad esempio quello di Fenati/Rovere/Schemann. In questo primo campo potrebbero esserci solo i traducenti, mentre in un successivo, apribile se fosse necessario proseguire la ricerca, potrebbero trovarsi anche i relativi esempi d'uso. Al fondo potrebbero essere aggiunti link a banche dati, come ad esempio *COSMAS II* o il sito http://it.bab.la/dizionario/ tedesco- italiano nelle quali il lettore possa trovare ulteriori informazioni.

7 Conclusioni

I dizionari bilingui sono centrali nella didattica delle lingue ma, per essere veramente utili, devono essere pensati per le esigenze specifiche dell'apprendente. Il dialogo da poco iniziato tra linguisti e lessi-

cografi si è rivelato un'enorme spinta per quell'innovazione lessicografica difficile ma necessaria che non poteva essere ulteriormente procrastinata. Accogliendo l'invito rivolto dai linguisti ai lessicografi, Luisa Giacoma e Susanne Kolb hanno introdotto nella lessicografia bilingue italo-tedesca la registrazione sistematica dei collocatori in spazi ben identificati anche graficamente (carattere maiuscoletto tra parentesi graffe nella versione cartacea con l'aggiunta del colore rosso in quella elettronica). Hanno reso altresì possibile, nella versione elettronica disponibile sia su cd-rom che on line, la consultazione per collocatori attraverso un campo di ricerca ad essi dedicato.

Ai collocatori sono state affiancate anche oltre 40.000 formule di struttura che descrivono in modo esplicito come il lemma e i collocatori possono combinarsi tra loro. In questo modo è stato possibile superare i confini della lessicografia tradizionale costruendo dizionari di nuova generazione che forniscono informazioni esplicite e sistematiche sul contesto nel quale il lemma compare. A proposito del metodo sviluppato nel dizionari Giacoma/Kolb Gobber (2003, 448) afferma: "[...] thanks mainly to the methodology adopted, this could well be considered one of the best bilingual dictionaries of any language pair anywhere. The editors are to be congratulated on such a work, which will be relevant and useful for years to come".

Nelle versioni elettroniche più recenti sono state anche aggiunte le tabelle di flessione per tutta la lingua tedesca, compiendo un altro importante passo verso il lettore, che frequentemente ha dubbi proprio sulla morfologia del tedesco.

Naturalmente le rapide evoluzioni non solo tecnologiche del futuro permetteranno di migliorare continuamente i dizionari, non senza però la quotidiana collaborazione tra lessicografi, linguisti e case editrici per poter essere sempre di più dalla parte del lettore.

8 Bibliografia

Bianco, M. T. (1996). Valenzlexikon Deutsch – Italienisch, Dizionario della valenza verbale. Heidelberg: Julius Groos Verlag.

Blumenthal, P., Rovere, G. (1998). Wörterbuch der italienischen Verben. Konstruktionen, Bedeutungen, Übersetzungen. Stuttgart: Klett.

Curcio, M. L. (1999). Kontrastives Valenzwörterbuch der gesprochenen Sprache Italienisch-Deutsch. Grundlagen und Auswertung. Mannheim: Institut für Deutsche Sprache.

DIT = (2008[4]). *Dizionario Tedesco – Italiano, Italiano – Tedesco*. Torino: Paravia. Berlin [u.a.]: Langenscheidt.

DIVA = Soffritti, M., Heinrich, W. (2005). Dizionario valenziale degli aggettivi/Deutsch – Italienisches Valenzwörterbuch der Adjektive. (Applicazione software).

Dobrovol'skij, D. (2009): Zur lexikografischen Repräsentation der Phraseme (mit Schwerpunkt auf zweisprachigen Wörterbüchern). In C. Mellado Blanco (Hg.), *Theorie und Praxis der idiomatischen Wörterbücher*. Tübingen: Niemeyer, 149-168.

ELDIT. (1999ss.) *Elektronisches Lern(er)wörterbuch Deutsch – Italienisch/ Dizionario elettronico per apprendenti italiano – tedesco*. Bolzano: Europäische Akademie (www.eurac.edu/eldit).

Fischer, K., Mollica, F. (2012). *Valenz, Konstruktion und Deutsch als Fremdsprache*. Frankfurt a.M. [u.a.]: Peter Lang.

Fontenelle, T. (1997). Turning a bilingual dictionary into a lexical-semantic database. Tubingen: Niemeyer.

Giacoma, L. (2011). Übersetzungsfehler und Gebrauch von zweisprachigen Wörterbüchern Deutsch – Italienisch: ein Erfahrungsbericht. In S. Bosco, M. Costa, L. Eichinger edd. *Deutsch/Italienisch: Sprachvergleiche*. Heidelberg: Winter Verlag, pp. 45-65.

Giacoma, L., Kolb, S. (2006). L'utilità dell'introduzione sistematica delle collocazioni nella voce lessicografica bilingue. L'esempio del *Dizionario di Tedesco* (GIACOMA/KOLB, Zanichelli/Klett, 2001). In E. Corino, C. Marello, C. Onesti edd. *Atti del XII Congresso Internazionale di Lessicografia – Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso, pp. 967-978.

GK = Giacoma, L., Kolb, S. edd. (2014[3], 2009[2], 2001). *Il Nuovo dizionario di Tedesco*. Bologna: Zanichelli/Stuttgart: Klett.

Giacoma, L., Kolb, S. edd. (2010). PONS Wörterbuch Studienausgabe Italienisch-Deutsch. Deutsch-Italienisch. Stuttgart: Klett.

Giacoma, L., Kolb, S. edd. (2011). *Il Tedesco smart*. Bologna: Zanichelli/Stuttgart: Klett.

Gobber, G. (2003). Review – *Dizionario Tedesco Italiano – Italiano Tedesco/Wörterbuch Deutsch Italienisch – Italienisch Deutsch,* hrsg. v./a cura di Luisa Giacoma e Susanne Kolb. Bologna: Zanichelli/Stuttgart: Pons Klett. 2001. In *International Journal of Lexicography* 16/4, pp. 445-448.

Gross, M. (1967). *Analyse formelle comparée des complétives en français et en anglais*, Thèse de troisième cycle. Université de la Sorbonne. Paris.

Gross, M. (1975). *Méthodes en syntaxe*. Paris: Hermann.

Hausmann, F. J. (1993). Ist der deutsche Wortschatz lernbar? In *Informationen Deutsch als Fremdsprache* 20, 471-485.

LGWDaF = Götz, D. et al. edd. (1993). *Langenscheidt Großwörterbuch Deutsch als Fremdsprache*. Berlin [u.a.]: Langenscheidt.

Lo Cascio V. (1997). Semantica lessicale e i criteri di collocazione nei dizionari bilingui a stampa e elettronici. In T. De Mauro, V. Lo Cascio (a cura di), *Lessico e grammatica. Teorie linguistiche e applicazioni lessicografiche. Atti del convegno interannuale della Società di Linguistica Italiana*. Roma: Bulzoni.

Marello, C. (1989). *Dizionari bilingui*. Bologna: Zanichelli.

Marello, C., Rovere, G. (1999). Mikrostrukturen in zweisprachigen Wörterbüchern Deutsch – Italienisch/ Italienisch – Deutsch. In H. E. Wiegand, *Germanistische Linguistik* 143-144 (Studien zur zweisprachigen Lexikographie mit Deutsch IV), pp. 177-206.

Nied Curcio, M. (2006). La lessicografia tedesco–italiana: storia e tendenze. In: F. San Vicente, *Lessicografia bilingue e traduzione: metodi, strumenti, approcci attuali*. Monza: Polimetrica International Scientific Publisher, pp. 57-70.

Nied Curcio, M. (2012). Die Valenz in der zweisprachigen Lexikographie Italienisch-Deutsch. Wohin führt der Weg? In *Studi germanici* 1/12, pp. 175-191.

PONS = (1996). *Pons Großwörterbuch Französisch Deutsch.* Stuttgart: Klett.

Sabatini, F. (2008). La grammatica in un dizionario. In: S. Vanvolsem, L. Lepschy, Nell'officina del dizionario. Atti del Convegno Internazionale organizzato dall'Istituto Italiano di Cultura. Lussemburgo 10 giugno 2006. Stuttgart: Ibidem Verlag, pp. 111-112.

Sansoni = (2006[6]). Dizionario Sansoni Tedesco–Italiano, Italiano–Tedesco, Milano: RCS Libri.

Schafroth, E. (2004). Anmerkungen zur lexikographischen Dimension der Lernersprachen Italienisch und Deutsch. In: *daf-werkstatt* 3, pp. 109-124.

Bilingual Dictionary Drafting. The Example of German-Basque, a Medium-density Language Pair

David Lindemann[1], Iker Manterola[2], Rogelio Nazar[3], Iñaki San Vicente[2], Xabier Saralegi[2]
[1]UPV-EHU University of the Basque Country, [2]Elhuyar Foundation,
[3]Pontificia Universidad Católica de Valparaíso
david.lindemann@ehu.es, i.manterola@elhuyar.com, rogelio.nazar@ucv.cl,
i.sanvicente@elhuyar.com, x.saralegi@elhuyar.com

Abstract

This paper presents a set of Bilingual Dictionary Drafting (BDD) methods including manual extraction from existing lexical databases and corpus based NLP tools, as well as their evaluation on the example of German-Basque as language pair. Our aim is twofold: to give support to a German-Basque bilingual dictionary project by providing draft Bilingual Glossaries and to provide lexicographers with insight into how useful BDD methods are. Results show that the analysed methods can greatly assist on bilingual dictionary writing, in the context of medium-density language pairs.
Keywords: bilingual dictionary drafting; comparable corpora; Natural Language Processing; open lexical resources; parallel corpora

1 Introduction

For a bilingual dictionary project that starts from scratch, from no or little previous lexicographical work and no or little bilingual glossaries (BG) existing on their language pair, a lexicographer lacks a useful set of guidelines for Bilingual Dictionary Drafting (BDD) strategies. A Dictionary Draft, i.e., lexicographical data obtained by automatic or semi-automatic methods, is useful in the lexicographical process as it may ease the editing of macro- and microstructural lexicographical data and save human resources.

In this article, we present a set of BDD methods and their evaluation on the example of German-Basque as language pair: Direct extraction of bilingual glossaries from existent lexicographical databases and Corpora based Natural Language Processing extraction methods. The evaluation of the obtained glossaries is done (1) quantitatively for the covering of German lemmata against a corpus based frequency lemma list adapted from DeReWo-40.000 (IDS 2009) as gold standard, (2) quantitatively for the amount of Basque Translation Equivalents (TE) obtained, and (3) qualitatively for the adequateness of the TE pairings, against manually edited German-Basque dictionary entries from EuDeLex, the lem-

malist of which is adapted from DeReWo-40.000, as gold standard (3406 German lemmata starting with A), and for the adequateness of the TE's part of speech (POS)[1].

Our aim is twofold: to give support to the bilingual dictionary project EuDeLex by providing draft BG and to provide useful information related to BDD methods for lexicographers working on medium-density language pairs.

A Word about Density

Density, understood as "the availability of digitally stored material" in a language (Varga et al. 2005) is a factor not to be neglected in corpus-based lexicography. In most cases, the number of speakers of a language and its size on the web serve as approximation indicators for density, and the availability of electronic language resources is also a factor to be considered. Following Varga et al. (*op. cit.*), we group languages according to density as follows:

(1) High-density languages: languages with a hundred million speakers or more (about 12)
(2) Low-density languages: small languages with less than half a million speakers (more than 5000)
(3) Medium-density languages: languages that lie between these two extremes (about 500)

Basque is one of the latter ones; Table 1 carries a comparison of density approximation indicators for German, one of the high-density languages, and Basque.

In the bilingual context, it is the density of the smaller language of the pair that determines by which methods Dictionary Draft data can be gathered and to what extent those methods lead to useful results.

Approaches for obtaining BG that rely on statistical Natural Language Processing methods (part 2.2) and that provide reliable results in higher density language pairs, in our case may lead to a much more limited success, and we shall ask whether the reasons for a more limited success of NLP methods are the quantitative and qualitative limitations of parallel and comparable corpora available for our language pair, or whether a lack of performance is explained also by the employed NLP tools themselves, which do lead to good results for the (high density) languages they were designed for. In the case of Basque, as it is not official in EU, we can not recur to parallel corpora based on EU legal documents, as we would in the case of other medium-density European languages (cf. Steinberger et al. 2006). Independently from this fact, German-Basque parallel corpora compiled from movie subtitles and software localization files may reach considerable sizes in a near future, as promised in the OPUS Corpus project (cf. Tiedemann 2012).

1 EuDeLex is currently being developed at UPV-EHU (cf. Lindemann 2014). The manual editing of German letter A (around 10% of the planned lemmalist based on DeReWo) has been finished. The intersecting set of EuDeLex and DeReWo (German Letter A) covers more than 90% of both. EuDeLex is available at http://www.ehu.es/eudelex/.

	German	Basque
Speakers	98 million	0,8 million
Biggest Corpus (token counts)	5,4 billion	0,12 billion
Wikipedia Pages	4,5 million	0,37 million
Web contents	5.7%	< 0.1%
ELRA Products	444	6

Table 1: Some density approximation indicators for German and Basque.

Approaches that rely on lexicographical databases maintained by human lexicographers (part 2.1) also presumably suffer from a density-bias: Wikimedia content is crowd-edited by the collaborating communities of volunteers, and those of a high-density language like German largely outnumber voluntaries in Basque[2]. On the other hand, lexical databases maintained in an academic context by human lexicographers like WordNet, may be on a par in terms of quantity and quality, disregarding density across languages.

2 Bilingual Dictionary Drafting Methods

2.1 Extraction of BG from existent lexicographical databases

2.1.1 Wikimedia

For the page titles that match to our gold-standard lemmalist, we extract the whole Wikipedia / Wiktionary page content if existent. Redirect pages are also taken into account. From the page content, we extract the Interlanguage-Link for Basque (Wikipedia) and the Basque translation links (Wiktionary).

2.1.2 WordNet

In this experiment, we align German WordNet lexical units (GermaNet 8.0, see Hamp & Feldweg 1997) with Basque WordNet (EusWN 3.0, see Pociello 2007) lexical units using Princeton WordNet (PWN 3.0, see Fellbaum 1998) as pivot. GermaNet synsets (to which *n* lexical units belong) are referred to PWN synsets in the GermaNet Interlingual Index records (ILI). On the other hand, the EusWN datasets carry links to PWN. By parsing the WordNet data into the same XML file, we get a structure like the one shown in Fig. 1. From each of those aligned datasets, German-Basque glossary entries are extracted by pairing all German lexical units to all Basque lexical units present in the synset.

2 For instance, German *wiktionary* counts with 78634 user accounts and 199 active members over the last month, while the Basque *wiktionary* only has 1982 accounts with 11 active members (statistics from 15.09.2013).

```
<EngSynset Synset="eng-30-00042757-n">
     <PWN30 EngLexUnit="departure_1"></PWN30>
     <PWN30 EngLexUnit="going_1"></PWN30>
     <PWN30 EngLexUnit="going_away_1"></PWN30>
     <PWN30 EngLexUnit="leaving_1"></PWN30>
     <GermaNet80 GerLexUnit="Abfahrt"></GermaNet80>
     <GermaNet80 GerLexUnit="Weggang"></GermaNet80>
     <GermaNet80 GerLexUnit="Aufbruch"></GermaNet80>
     <GermaNet80 GerLexUnit="Distanzierung"></GermaNet80>
     <EusWN30 EusLexUnit="irteera_6"></EusWN30>
     <EusWN30 EusLexUnit="joanaldi_1"></EusWN30>
     <EusWN30 EusLexUnit="joate_1"></EusWN30>
</EngSynset>
```

Fig. 1: Aligned data of 3 WordNets.

2.2 NLP Methods

In this work, three different tools were used in order to extract lexical correspondences. Each tool depends on different NLP methods and resources. On the one hand, we applied a tool called *Pibolex,* which relies on pivoting over existing bilingual dictionaries and combines their structure and comparable corpora based methods for selecting correct translations of source words. On the other hand, we made use of two tools for bilingual lexicon extraction from parallel corpora: Giza++ and Bifid. The following sections describe those tools and the experiments we conducted with them.

2.2.1 Pibolex: Pivot techniques + comparable corpora word-alignment

Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries which share a common language (e.g. LA-LB, LB-LC) in order to create a dictionary for a new language pair (e.g. LA-LC). In our case, we merged the German-English *Beolingus*[3] dictionary (Lde-en) with the English-Basque *Elhuyar*[4] dictionary (Len-eu), obtaining Lde-eu. However, this process may include wrong translations due to the polysemy of words. A pivot word can lead to wrong translations corresponding to senses not represented by the source word. These senses can be either completely different or related but with a narrower or wider meaning. For pruning the wrong translations, in this work we apply the Pibolex tool (Saralegi, Manterola & San Vicente 2011) which uses two different methods adequate for medium-density language pairs because they depend on resources that can be easily obtained:

(a) Inverse Consultation (IC1) (Tanaka & Umemura 1994): this algorithm uses the structure of the source dictionaries to measure the similarity of the meanings between a source word and its translation candidates. The IC1 method counts the number of pivot words in language B between a source word in LA and its TE candidate in LC. The more pivot words found, the stronger is the evidence for the candidate to be correct.

(b) A pruning method based on cross-lingual distributional similarity (DS) computed from a bilingual comparable corpus. Different authors (e.g. Fung 1995; Rapp 1999) have proposed to extract bilingual

3 http://dict.tu-chemnitz.de
4 http://hiztegiak.elhuyar.org

equivalents from monolingual or comparable corpora because, despite offering lower accuracy than those extracted from parallel corpora, they can be an alternative for medium and low density language pairs where parallel corpora are scarce. The underlying idea is to identify as TEs those words which show similar distributions or contexts across two corpora of different languages, assuming that this similarity is proportional to the semantic distance. The method we apply here is described in detail in Saralegi, San Vicente & Gurrutxaga (2008). Following the "bag-of-words" paradigm, a word *w* is represented by a vector composed of weighted collections of words. Those words are extracted from the contexts where the word *w* appears in the corpus. The context words are weighted with regard to *w* according to the Log-likelihood ratio measure. Once we have vector representations of the words in both languages, the algorithm computes for each source word in LA the cosine similarity between its context vector and the context vectors of all TE candidates in LC. However, we can not directly compare vectors in different languages. In order to overcome this problem, we translate vectors of words in LA to LC by means of the noisy bilingual dictionary Ldeeu, which is the only bilingual dictionary available at this stage of the process.

The IC1 algorithm suffers from low recall, which makes it rather inadequate for the task at hand. But the combination of it with the DS based method may be a way to tackle this problem. DS results vary depending on the corpora used for computing the cross-lingual similarities. The more comparable the corpora, the better. With that in mind, experiments were conducted over two different comparable corpora:

(4) News comparable corpus: the first experiment was conducted using a comparable corpus composed of news articles extracted from Die Zeit[5] newspaper in German (29M tokens) and from Berria[6] newspaper in Basque (36M tokens). No effort was done to match news topics or publications dates.

(5) Wikipedia comparable corpus: Wikipedia has been extensively exploited with NLP methods a comparable corpus (eg, Tomás et al. 2008; Paramita et al. 2012). In this case we constructed a corpus by gathering all articles that have both Basque and German versions, connected through wiki interlanguage links. The corpus has 61.484 articles per language and 91M tokens (72.5M tokens in German and 18M tokens in Basque). Although this corpus is highly comparable with respect to the topics (each article has its counterpart), it is important to note that the amount of tokens per language is unbalanced. This can lead to a decrease in the comparability degree of the corpus, because the German part holds much more information.

5 http://www.diezeit.de
6 http://berria.info

Table 2 shows the dictionaries used in the process and their statistics.

	#entries	#pairs
Lde-en (A) - Beolingus	146,451	171,775
Len-eu (B) - Elhuyar	17,672	43,201
Lde-eu (A+B, no pruning)	12,939	48,097
Lde-eu (IC1)	4,305	7,211
Lde-eu (IC1+DS wiki)	7,878	18,641
Lde-eu(IC1+DS news)	7,821	20,014

Table 2: Pivot dictionary process.

2.2.2 Parallel Corpus word-alignment

There is a large tradition of parallel corpus processing in computational linguistics, starting with the work of Gale & Church (1991), Brown, Lai & Mercer (1991), McEnery & Oakes (1995) and others (see Véronis 2000 for an overview). Different methods and tools have been proposed to align parallel texts and extract lexical correspondences from them. In this investigation, we used two word alignment tools based on these methods: Giza++ (Och & Ney 2000) and Bifid (Nazar 2012).

The fact that Basque is a medium density language unlike German represents an added difficulty for any attempt in the line of Resnik (1999), who proposed to download parallel corpora by mining the web for translated pages. In our experiments we used two parallel corpora of different sizes: a German-Basque Literary Corpus created in the context of recent research (Sanz Villar 2013; Zubillaga 2013) and another built by aligning Basque and German translations of the Bible[7]. The first one was compiled using the content of 81 digital or digitized and OCR-ed literary German originals and their official direct translations into Basque (146,457 segment pairs). In the case of the second, after removing the books not included in both Bible versions, a parallel corpus of 30,440 segment pairs was obtained using the verse as segment unit. This is an easily built resource for medium-density language pairs because the Bible is available for a wide variety of languages (Lardilleux, Gosme & Lepage 2010; Resnik, Olsen & Diab 1999) and is, therefore, an adequate baseline parallel resource for evaluating other extraction methods over this kind of language pairs.

The aforementioned extractions tools (Giza++ and Bifid) were used in order to obtain translations pairs. The resulting figures are shown in table 3.

7 Basque (Elizen Arteko Biblia 1994) and German (1984 revision of the Luther Bible).

	# of seg.	# of EU tokens	# of DE tokens	# of candidates GIZA++ p(b|g)>0.1	# of candidates BIFID
Literary Corpus	146,457	1,948,504	2,203,307	266,678	4,838
Bible Corpus	30,440	639,581	810,671	49,443	2,926

Table 3: DE-EU parallel corpora.

For word alignment with Giza++, the default sequence of models were used (IBM model 1, HMM-based model, IBM model 3 and IBM model 4). The German corpus was lemmatised and POS tagged by using Treetagger (Schmid 1995) and the Basque one using Eustagger (Ezeiza et al. 1998). Then, each word of the source corpus was substituted by a chain including the corresponding lemma and POS category. Punctuation marks and words with POS regarded as possible source of noise in the alignment process were removed from both corpora. Specifically, words excluded from the German part were those with POS tags such as APPR (preposition), APPRART (preposition with article), ART (article), KOKOM (particle of comparison), KOUS (subordinating conjunction), PRELS (relative pronoun), VAFIN (auxiliary verb, finite form) and VAINF (auxiliary verb, infinitive). From the Basque corpus, in turn, the excluded units were those with the ADL tag (auxiliary verb)[8].

Giza++ returns two files of word alignments (DE-EU and EU-DE) including a probability for each word alignment. For the draft BG, BibleGiza and LitGiza word alignments with a probability p(b|g) greater than 0.1 were selected from both Bible and Literary corpora.

In order to reduce noise, the BG obtained by Giza++ was then submitted to a filtering process using a stoplist consisting of the 150 most frequent Basque words[9]. Two versions of these BG have been evaluated: (1) Giza++ BG after stoplist filtering, and (2) after stoplist and a filtering that only allows BG entries with the POS-tags mapping to each other in one of the following ways (see table 4):

TreeTagger flag	**Eustagger flag**
NN (noun)	IZE (noun)
VV (verb)	ADI (verb)
ADV (adverb)	ADB (adverb)
AD (adjective)	ADB (adverb)[1]
AD (adjective)	ADJ (adjective)

Table 4: mapping of POS-tags TreeTagger (German) and Eustagger.

The other alignment tool, Bifid, is part of a larger project comprising the analysis of language pairs where no prior knowledge is available, which means that all forms of external resources are excluded from the processing. This tool incorporates modules for the integral process of analysing a set of do-

8 VAFIN and VAINF would be German equivalents of Basque ADL. The other removed German POS would be a morpheme in Basque words in almost all cases.

9 The Basque stoplist has been obtained from Basque ETC Corpus data (UPV-EHU, Sarasola et al. 2013).

cuments in unknown languages with the only assumption that such set consists of a parallel corpus in two languages. In its original version, this tool separates the set of documents in the two languages, aligns each document with its most probable translation and then proceeds to align the segments inside the documents (assuming that the newline character is the segment separator). Finally, from this segment alignment, it extracts an initial bilingual vocabulary which is then used for a realignment of the corpus at the segment level. The process is iterated in this way *n* times to improve the quality of the alignment at all levels.

In the case of this paper, however, we only used the bilingual lexicon extraction module because our parallel corpora were already aligned at the sentence level. The corpora were also lemmatized with the above mentioned tools, but no mapping exploiting the POS tags was used because this information is not used by the algorithm[10]. The bilingual vocabulary extraction module of Bifid uses a combination of strategies that include co-occurrence statistics as well as length and orthographic similarity metrics. As in its original version this extracted vocabulary was intended to be used for realignment, the program is very conservative in its lexical alignment in order to avoid the reproduction of errors in subsequent steps. As a consequence, it favors precision over recall, with fewer aligned pairs having a higher probability of being correct. Further experimentation will determine the right thresholds for the best compromise between noise and silence, meaning larger sets of aligned pairs with the maximum possible purity.

3 Evaluation

3.1 Comparison of German Lemma lists

Germanet offers the best recall on DeReWo lemmata. Wikipedia and Wiktionary on their own offer a similar recall; the intersections of both of these with DeReWo also reaches a very high level (see Table 5 below):

10 As the motivation behind Bifid is to be a language independent alignment tool, it does not use any kind of language-specific resources such as lemmatization or POS-tagging.

	Derewo		GermaNet		Wikipedia		Wiktionary	
∩ Derewo			32,199	33,73%	19,461	2,22%	22,028	7.01%
∩ GermaNet	32,199	80,50%			47,588	5,43%	34,309	10.93%
∩ Wikipedia	19,461	48,65%	47,588	49,86%			46,968	14.96%
∩ Wiktionary	22,028	55,07%	34,309	35,94%	46,968	5,36%		
∩ WikiORWikt	29,164	72,91%	59,995	62.86%				
Lemma total	39,998		95,449		876,309		314,016	

Table 5: DeReWo and existing lexicographical databases: German lemma counts (A-Z) and intersecting sets.

The best recall is offered by LitGiza, with a notable difference regarding the rest of drafts, even Bible-Giza. Pibolex recall is second best but far from LitGiza. BibleBifid and LitBifid offer a very low recall (see table 6).

	Derewo	Bible Giza Stop	Bible Giza StopPos	LitGiza Stop	LitGiza StopPos	Bible Bifid	Lit Bifid	Pibolex Wiki	Pibolex News
Derewo		4,639 (34.97%)	3,500 (38.69%)	15,775 (23.30%)	12,846 (24.93%)	1,007 (40.84%)	2,995 (67.27%)	5,812 (77.34%)	5,753 (77.14%)
BibleGiza Stop	4,639 (11.60%)		9,047 (100.00%)	5,001 (7.39%)	3,879 (7.53%)	2,372 (96.19%)	1,122 (25.20%)	1,868 (24.86%)	1,851 (24.82%)
BibleGiza StopPos	3,500 (8.75%)	9,047 (68.19%)		3,763 (5.56%)	3,248 (6.30%)	1,699 (68.90%)	957 (21.50%)	1,518 (20.20%)	1,504 (20.17%)
LitGiza Stop	15,775 (39.44%)	5,001 (37.70%)	3,763 (41.59%)		51,533 (100.00%)	1,125 (45.62%)	4,389 (98.58%)	4,571 (60.83%)	4,549 (60.99%)
LitGiza StopPos	12,846 (32.12%)	3,879 (29.24%)	3,248 (35.90%)	51,533 (76.12%)		935 (37,92%)	3,864 (86.79%)	3,960 (52.69%)	3,955 (53.03%)
Bible Bifid	1,007 (2.52%)	2,372 (17.88%)	1,699 (18.78%)	1,125 (1.66%)	935 (1.81%)		526 (11.81%)	544 (7.24%)	546 (7.32%)
Lit Bifid	2,995 (7.49%)	1,122 (8.46%)	957 (10.58%)	4,389 (6.48%)	3,864 (7.50%)	526 (21.33%)		1,404 (18.68%)	1,398 (18.74%)
Pibolex Wiki	5,812 (14.53%)	1,868 (14.08%)	1,518 (16.78%)	4,571 (6.75%)	3,960 (7.68%)	544 (22.06%)	1,404 (31.54%)		7,309 (98.00%)
Pibolex News	5,753 (14.38%)	1,851 (13.95%)	1,504 (16.62%)	4,549 (6.72%)	3,955 (7.67%)	546 (22.14%)	1,398 (31.40%)	7,309 (97.26%)	
Lemma total	39,998	13,267	9,047	67,699	51,533	2,466	4,452	7,515	7,458

Table 6: DeReWo and BG German entries intersections (A-Z, NLP methods).

3.2 Quantitative and Qualitative Evaluation of BG

Table 7 shows results of the qualitative evaluation carried out manually by a human lexicographer. EuDeLex is set as gold standard for comparison, regarding lemmalist and evaluation of TE appropriateness, in terms of (a) a full matching as suitable Basque TE for one of the word senses of the German BG headword (OK), (b) a semantic mismatch (FALSE), (c) a semantic (fuzzy) matching without being the TE a valuable equivalent to cite in a dictionary entry (NEAR), or (d) as PART, when a BG entry is a correct TE for a lemma as part of a Multi Word Expression in the other language. A second variable, part of speech (POS) is evaluated as matching (OK) or mismatching (FALSE).

The BG obtained from aligned WordNet synsets offers a relatively high recall on GS lemmata and, in absolute figures, the largest proportion of correct TEs. Wikipedia and Wiktionary extraction is less effective in terms of recall, but more effective with regard to TE adequateness.

Among the NLP methods, Giza and Pibolex BG offer the largest number of TE evaluated as correct, far ahead of Bifid, which on the other hand returned very little false TE and POS among the results. Giza BG, and, to a lower extent, Pibolex BG, suffer from a high percentage of inadequate, noisy TE.

	LitGiza Stop	LitGiza StopPos	BibleGiza Stop	BibleGiza StopPos	LitBifid	BibleBifid	Pibolex News	Pibolex Wiki	Wikipedia	Wiktionary	WordNet	ANY DRAFT
DE Letter A Lemma with EU TE	5,361	3,826	1,349	818	265	230	682	730	5,457	396	1,076	
Intersection with EuDeLex GS	1,276	696	434	323	177	49	559	567	265	147	774	2,928
Recall on 3406 GS lemmata	**35,41%**	**19,31%**	**12,04%**	**8,96%**	**4,91%**	**1,36%**	**15,51%**	**15,73%**	**7,35%**	**4,08%**	**21,48%**	**81,24%**
Ø TE per Lemma	2,27	1,73	2,35	1,46	1,07	1,22	2,64	2,82	1,00	1,33	2,57	2,28
Evaluated Lemma	1,276	696	434	323	177	49	559	567	265	147	774	4,248
Evaluated TE	2,901	1,206	1,019	470	189	60	1,476	1,601	265	195	1,988	9,694
• OK	**746**	**537**	**256**	**215**	**178**	**43**	**939**	**1,007**	**236**	**189**	**1,654**	**5,248**
• FALSE	1,923	519	728	230	3	15	414	513	6	3	188	3,793
• PART	186	122	11	6	6	1	14	7	4		17	246
• NEAR	46	28	24	19	2	1	109	74	19	3	129	407
Evaluated POS	2,901	1,206	1,019	470	189	60	1,476	1,601	265	195	1,988	9,694
• OK	1,071	965	448	422	186	52	1,321	1,286	245	195	1,983	6,787
• FALSE	1,830	241	571	48	3	8	310	315	21		5	3,063
Lemma with 1+ TE OK	609	439	213	189	168	39	443	454	236	147	700	2,081
Lemma with all TE OK	315	294	111	149	168	35	316	322	236	142	601	1,123
Lemma with something usable	444	232	129	62	7	6	165	164	23	5	143	1,056
Lemma with all TE FALSE	517	170	194	112	2	8	78	82	6		30	371
TE OK	**25,72%**	**44,53%**	**25,12%**	**45,74%**	**94,18%**	**71,67%**	**63,62%**	**62,90%**	**89,06%**	**96,92%**	**83,20%**	**54,14%**
POS OK	36,92%	80,02%	43,96%	89,79%	98,41%	86,67%	89,50%	80,32%	92,45%	100,00%	99,75%	70,01%
Lemma with 1+ TE OK	**47,73%**	**63,07%**	**49,08%**	**58,51%**	**94,92%**	**79,59%**	**79,25%**	**80,07%**	**89,06%**	**100,00%**	**90,44%**	**48,99%**
Lemma with all TE OK	24,69%	42,24%	25,58%	46,13%	94,92%	71,43%	56,53%	56,79%	89,06%	96,60%	77,65%	26,44%
Lemma with something usable	34,80%	33,33%	29,72%	19,20%	3,95%	12,24%	29,52%	28,92%	8,68%	3,40%	18,48%	24,86%
Lemma with all TE FALSE	40,52%	24,43%	44,70%	34,67%	1,13%	16,33%	13,95%	14,46%	2,26%	0,00%	3,88%	8,73%

Table 7: Manual Qualitative Evaluation.

4 Conclusions

4.1 Discussion

Combining all BDD methods presented here, we obtain a BG that covers more than 80% of a dictionary lemmalist based on DeReWo, and provides one or more correct TE for about a half of those.
For BDD purposes, correct TE must be separated from not suitable (noisy) BG entries; TE adequateness has to be the key criterion for lexicographical needs, before the amount of gathered data. In the ongoing editing process of EuDeLex, draft data will be divided in three groups, (1) methods with no or very little results evaluated as FALSE; the data obtained by those may be included in dictionary entries and published, without manual post-editing, (2) methods with high precision results (low degree of noise); the BG obtained by these methods could be pasted into the bilingual lexicographical database for manual post-editing, and (3) methods with a larger proportion of noisy results; the BG obtained by those will have to be post-processed in order to reduce false TEs; the POS-mapping approach presented here for Giza is a first step in that direction, enriching a Basque stoplist for results proposed by Giza from the list of Basque TEs that repeatedly have been evaluated as FALSE will be the next.
We propose to group the methods presented in this paper according to the criteria mentioned above as follows:

(1) Wiktionary, Wikipedia, LitBifid
(2) BibleBifid, WordNet
(3) LitGizaStop, BibleGizaStop, LitGizaStopPos, BibleGizaStopPos, Pibolex News, Pibolex Wiki

As we found out in this investigation, more than two thirds of the DeReWo list, on which a lemmalist for EuDeLex that covers the whole alphabet will base on, are linked to a dataset in Wiktionary and/or Wikipedia. The high rates in our qualitative evaluation reached by these methods encourage us to make use of them, and thanks to their open licence, it is possible. While other draft data needs human post-editing before inclusion in published bilingual dictionary entries, relevant data from those sources may be directly included in a dictionary search result webpage[11]. The recall these sources offer for German-Basque TE is still limited; it will mainly depend on the growth and activity of the Basque editor communities to increase it, which is, supposedly, a matter of time. Measurements like those proposed in this investigation may serve to monitor that process.
The Basque WordNet EusWN has been actively developed by human lexicographer teams at UPV-EHU, and it is today the largest and trustworthiest Basque lexical resource available with open data sources. The approach to align its synsets with GermaNet synsets using Princeton WordNet as pivot has been the one which delivered the largest proportion of correct TE for more German lemmata.
Pibolex overall results are similar to those obtained for other language pairs, confirming the tool performs robustly across languages. With respect to the corpora used in the experiments, the news corpus achieves slightly better results. This means that the Wiki corpus, although more comparable in terms of topics, suffers from the difference in amount of text between languages. The results obtained by both word alignment tools from two parallel corpora of different size show, as was to be expected,

11 Data found in these sources relevant to a bilingual dictionary is not only a TE, but also encyclopedic, phonetic, morphological, syntactic and pragmatic information or audiovisual material about a lemma.

that recall rates relate to corpus size, and the same is true for result precision. A further development of German-Basque parallel corpora is strongly desired.

In spite of German-Basque being a medium-density language pair with limited bilingual lexical and corpus resources, the amount of adequate BG entries gathered during the presented experiments is considerably high, and it will help saving human resources in dictionary writing. There is no need to say, however, that human lexicographers are still the key factor for a German-Basque dictionary writing that would meet acceptable quality standards.

4.2 Future Work

A future line of work would be to create higher comparability degree corpora, taking care of maintaining balance in terms of size, topics and genres across languages, without decreasing the overall size of the corpora. Further research about BDD would also include rendering optimization of the applied corpus based methods for the language pair German-Basque (enhancements of corpus tagging, word alignment stoplists and parameter tuning), a sophistication of these (e.g., by making use of syntactic information), as well as reproducing these experiment sets for other language pairs, which would allow for a comparison of results. Other goals not achieved at the present stage are the inclusion of multiword expressions in the experiments and measurements of polysemy covered by draft BG. We are now centering our efforts in developing a new method for the exploitation of Wikipedia as a comparable corpus using the frequency distribution of lexical units in the articles. We are representing the relative frequency of words in the articles as curves, and then comparing the curves in a purely geometrical fashion using Euclidean distance. We assume that German and Basque words with similar frequency curves will be equivalent, however we still need to find the way to make up for the already mentioned asymmetry in the amount of text in the corresponding articles in both languages.

5 References

Brown, P.F., Lai, J.C. & Mercer, R.L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 91, Stroudsburg, PA.: Association for Computational Linguistics, pp. 169–176.

Ezeiza, N., Alegria, I., Arriola, J.M., Urizar, R. & Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics*. Association for Computational Linguistics, pp. 380–384.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*. Boston, MA., pp. 173–183.

Gale, W.A. & Church, K.W. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the ACL Workshop on Speech and Natural Language*. ACL 91, Stroudsburg, PA: Association for Computational Linguistics, pp. 152–157.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid: Association for Computational Linguistics, pp. 9–15.

Lardilleux, A., Gosme, J. & Lepage, Y. (2010). Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*. LREC 2010, Malta, pp. 252–256.

Lindemann, D. (2014). Zweisprachige Lexikographie des Sprachenpaares Deutsch-Baskisch. In Domínguez Vázquez, M.J., Mollica, F. & Nied, M. (eds.) Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik, Lexicographica Series Maior. De Gruyter.

McEnery, A.M. & Oakes, M.P. (1995). Sentence and word alignment in the CRATER project: methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop: from texts to tags, Issues in Multilingual Language Analysis (ACL)*. Dublin, pp. 77–86.

Nazar, R. (2012). Bifid: un alineador de corpus paralelo a nivel de documento, oración y vocabulario. In *Linguamática*, 4, 45–56.

Och, F.J. & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 00, Hongkong: Association for Computational Linguistics, pp. 440–447.

Paramita, M.L., Clough, P., Aker, A. & Gaizauskas, R.J. (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 790–797.

Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. In *Language Resources and Evaluation*, 45, 121–142.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 99, College Park, MD.: Association for Computational Linguistics, pp. 519–526.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 99, College Park, MD.: Association for Computational Linguistics, pp. 527–534.

Resnik, P., Olsen, M.B. & Diab, M. (1999). The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues." In *Computers and the Humanities*, 33, 129–153.

Sanz Villar, Z. (2013). Hacia la creación de un corpus digitalizado, paralelo, trilingüe (alemán-español-euskera). In Sinner, C. & Van Raemdonck, D. (eds.) Fraseología contrastiva del alemán y el español. Traducción y lexicografía, Études linguistiques Linguistische Studien. München: Peniope, pp. 43–58.

Saralegi, X., San Vicente, I. & Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of the 1st workshop on Building and using Comparable Corpora (BUCC)*. LREC 2008, Marrakech.

Saralegi, X., Manterola, I. & San Vicente, I. (2011). Analyzing Methods for Improving Precision of Pivot Based Bilingual Dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP 2011, Edinburgh.

Sarasola, I., Landa, J. & Salaburu, P. (2013). Egungo Testuen Corpusa. UPV-EHU University of the Basque Country

Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, pp. 47–50.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. LREC 2006, Genoa.

Tanaka, K. & Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 297–303.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 2214–2218.

Tomás, J., Bataller, J., Casacuberta, F. & Lloret, J. (2008). Mining wikipedia as a parallel and comparable corpus. In *Language Forum*, 34.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. & Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*. Borovets, pp. 590–596.

Véronis, J. (2000). Parallel Text Processing: Alignment and use of translation corpora. Kluwer.

Zubillaga, N. (2013). Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez. PhD Thesis. Vitoria-Gasteiz: UPV-EHU.

Acknowledgements

This study has been supported by the following projects: IT665-13, Zubiak (Saiotek-SA-2013/00308) and Ber2tek (Etortek-IE12-333), funded by the Basque Government; and project EC FP7/SSH-2013-1 AThEME (613465), funded by the European Commission. Funding is gratefully acknowledged.

Illustrative Examples and the Aspect of Culture: The Perspective of a Tshivenḓa Bilingual Dictionary

Munzhedzi James Mafela
University of South Africa
mafelmj@unisa.ac.za

Abstract

Few Tshivenḓa dictionaries make use of illustrative examples to define lexical entries. Among those which use illustrative examples to define lexical entries is *Venda Dictionary: Tshivenḓa – English,* which makes use of simple sentences, idioms, riddles and proverbs. Through the use of illustrative examples lexicographers show the headword in use in typical contexts (Katzaros 2004). In many instances illustrative examples in *Venda Dictionary: Tshivenḓa -- English* reflect on the culture of the Vhavenḓa; information which is valuable to Tshivenḓa learners. Learners of a foreign language find it difficult to understand meanings of cultural words. Bartholomew (1995:4) writes that a cursory reading of the illustrative sentences not only helps the reader to understand the entry word, but it also gives him a feel for the kind of life led by the speakers. This paper seeks to highlight the importance of illustrative examples in familiarising dictionary users with the culture of the Vhavenḓa. Data from *Venda Dictionary: Tshivenḓa – English* will be used to facilitate the discussion of illustrative examples and the aspect of Tshivenḓa culture.

Keywords: Culture; illustrative example; phrase; proverb; idiom

1 Introduction

Bilingual dictionaries deal with the translation of entry words from the source language into the target language. In addition to the provision of words with equivalent meaning in the target language, a bilingual dictionary defines the entry word in order to make the meaning clearer to dictionary users. Sometimes such a definition of the entry word will be difficult to comprehend unless it is explained in context. This poses a challenge to dictionary users, especially those who are learners of the language, who may not find the meaning they are looking for. In order to avoid such misunderstandings, some lexicographers make use of illustrative examples to make the meaning of a word clearer. As Kavanagh (2000:101) writes: "We need to know about words and their meanings, but we also need to know about attitudes, manners, and social norms". The functional inclusion of examples, illustrating actual everyday language usage, is of prime importance because it adds to the quality and user-friendliness of a dictionary (Prinsloo & Gouws 2000:139). Illustrative examples give dictionary users the opportunity to learn more about the way of life of the source language community, including its cultu-

ral aspects. Such knowledge of the community's culture provides a key to understanding the meaning of lexical entries. Illustrative examples may serve as vehicles that clarify the meaning of culturally specific words.

An examination of the cultural setting and the equivalence in bilingual dictionaries reveals that language does not exist in a vacuum but occurs in context. Hence, relationships between the various languages as well as within a language become clear. One of the important aspects is the cultural context in which a word is used.(Gangla-Birir 2005:40)

The inclusion of cultural aspects in the definition of a lexical entry in a dictionary would be of great benefit to language learners in South Africa, a country with many racial and ethnic groups. Kavanagh (2000:101) asserts: "In South Africa, a country where people are being positively encouraged to learn another language and to communicate across cultures, the need to develop cultural as well as linguistic awareness should be emphasised".

The purpose of this paper is to highlight the importance of illustrative examples in familiarising dictionary users with the culture of a community. As Whitcut (1995:255) argues: "... it is in the use of examples that our cultural presuppositions become most obvious". Examples from the *Venda Dictionary: Tshivenḓa – English* will be cited to illustrate how learners of Tshivenḓa can be helped to acquire information on cultural aspects of the lives of the Vhavenḓa.

2 Illustrative Examples

Illustrative examples can take the form of sentences or phrases. On the use of sentences as illustrative examples, Bartholomew (1995:3) cites Bernard and Salinas Pectraza (1989) who state: "The dictionary sentences provide a series of snapshots of the daily life of the Mezquital Otomi Indians. They don't constitute ethnography as such but the information they supply is consistent with a fuller ethnography written by an Otomi speaker." The message in this quotation applies not only to Mezquital Otomi Indians but to all communities. Good illustrative sentences help to define the entry word by using it correctly in a typical context (Bartholomew 1995:3). When dictionary users read illustrative examples they can imagine the kind of life led by the source language community. Learners get to know how the example works and how it collocates (Katzaros 2004). Language has a close connection to the culture that produced it.

A language is not solely composed of words interacting with one another; it has a close connection to the culture that produced it. Thus, learning a language is also getting to know the particular aspects of the culture behind it. Examples can play an essential part in showing culture by illustrating words that have a specific cultural dimension. (Katzaros 2004:491)

A bilingual dictionary with illustrative examples is user-friendly because such examples show the entry word in context, and distinguish one meaning from another, illustrating grammatical patterns and typical collocations, and indicating appropriate registers (Drysdale 1987 cited by Al-Ajmi 2008:16).

3 Illustrative Examples and the Aspect of Culture in Venda Dictionary: Tshivenḓa – English

Illustrative examples in the *Venda Dictionary: Tshivenḓa – English* take several different forms. Some of the examples are in the form of proverbs, whereas others are in the form of phrases, riddles and idioms. This discussion will focus on proverbs, idioms and phrases as they are used to highlight cultural information about a language community. Illustrative examples in the *Venda Dictionary: Tshivenḓa* – English are informative because they illustrate the use of the word and enhance the dictionary user's understanding of its semantic range and the culture of the Vhavenḓa (Mafela 2008:31). Kavanagh (2000:112) comments thus on the presentation of cultural items in the definition of a lexical item in Van Warmelo's dictionary: "In the *Tshivenḓa – English Dictionary* edited by Van Warmelo cultural snippets follow the translation". Through these illustrative examples, dictionary users learn more than the mere meaning of a lexical entry.

3.1 Phrases

Van Warmelo (1989) has a tendency to use phrases to present the lexical entry in context. However, only those entry words whose meanings are figurative or cultural are explained in context. One such word is the lexical entry *-bva.* This entry is partly defined as follows:

(1) -bva 1 go out, come out, emerge, issue, come from. -- malofha bleed. -- ngomu escape. -- phaḓi get scabies (which come out of the skin). Nḓila i -- nga vhukati ha tsimu path goes right through the garden. –
ṱhangu go to consult a diviner.

The lexical entry *-bva* can be defined in three senses, i.e. going out or coming from, persevere, and pass. However, the English equivalents of *-bva* in the sense of going out or coming from are: go out, come out, emerge, issue, and come from. A number of illustrative examples in this regard have been provided to show its use in context: *-bva malofha* (bleed),-bva ngomu (escape), *-bva phaḓi* (scabies coming out of the skin). Only the last of these, *-bva ṱhangu* (go to consult a diviner) reveals an aspect of culture. The phrase *-bva ṱhangu* has been used as an illustrative example to distinguish its meaning from any other. According to the Tshivenḓa culture, if a member of a family dies, his/her relatives go out to consult a diviner to find out the cause of the death. This is a popular belief among the Vhavenḓa, as it is in some other African communities. These groups believe that people do not die of natural causes; they believe that death is brought about by another person or by the ancestors. This action is taken not only when a death takes place, but also when the family encounters misfortunes. A diviner is an important person in the lives of the Vhavenḓa in particular, and in African communities in general. Through this

illustrative example, learners of Tshivenḓa come to understand some of the cultural beliefs and religious activities of the Vhavenḓa.

The use of a phrase as an illustrative example to reveal the culture of the Vhavenḓa is also realised in the definition of the lexical entry *dzekiso;* which is defined this as follows:

(2) dzekiso 5 bride-price for a wife, (or, usually, one animal towards it), given by one's father or other close relative, see -dzekisa. Musadzi wa -- great wife, whose son will be heir (Van Warmelo 1989:35)

There is only one English equivalent for the lexical entry *dzekiso,* and that is bride-price of a wife. However, this is not an ordinary bride-price; it is the price paid for a wife who is highly regarded in the family. Polygamy is not a foreign practice among the Vhavenḓa and it is perfectly acceptable to the community for a man to marry more than one wife. However, a woman whose bride-price is paid by the father of the husband or a close relative in the family is known as *Musadzi wa dzekiso* (the great wife). This woman is the one who gives birth to the heir in the family. In the case of the royal house, *Musadzi wa dzekiso* (the great wife) gives birth to a future king/chief. Normally, *dzekiso* (bride-price) will be paid in the form of animals. This knowledge, beliefs, and practices of a particular society are reflected in a language (Kavanagh 2000:103).

The definition of the entry word *-kuvha* shows, too, how supplementary information regarding the meaning of a word can be provided through the use of illustrative examples:

(3) -kuvha 1 wash, as clothes. 2 (Kar. Idem) stop bearing (of trees at end of season), cease to be in season (of fruit). Mapapawa a vho -- zwino pawpaw are coming to an end. 3. finish payment, make up the balance outstanding, pay in full. -- misho pay all the presents due to bride's parents. Ndi ḓo -- dzoṱhe I shall take all. Engedza ṱhanu uri dzi kuvhe zwoṱhe pay a further five to make up the total

Muthu o felaho kule na haya u a kuvhiwa. If a man dies far from home and cannot be buried there, a sheep is killed and its head buried near his ancestors' graves by a doctor and the family. This is a substitute, and they say: "Mudzimu o vhuya hayani" (Van Warmelo 1989: 122).

Four meanings of the lexical entry *-kuvha* have been identified in the definition above. Meanings 1, 2 and 3 are clearly spelt out. However, the illustrative example in 3 is accompanied by information on a cultural aspect. The example *-kuvha misho* and the word *misho,* in particular, means to *pay all the presents due to a bride's parents.* Through this illustrative example, dictionary users learn that a bride-

groom is expected to produce all the presents before he is allowed to take his bride. If not, the bride will not be given permission to join the in-laws. *Kuvha* in this instance means pay.

Another interesting aspect of culture is revealed in illustrative example 4 (*Muthu o felaho kule na haya u a kuvhiwa* - If a man dies far from home and cannot be buried there, a sheep is killed and its head buried near his ancestors' graves by a doctor and the family). This illustrative example is not numbered and the meaning of the entry -*kuvhiwa* is not easily identified. Its use in this context reveals a cultural aspect of the lives of the Vhavenḓa. The Vhavenḓa believe that when a man dies far from home, he should be buried with his ancestors. This is not possible if the corpse is not brought back home. Instead, family members and the doctor will kill a sheep as a substitute. If this does not happen, something bad will befall the family. This is the reason a sheep is killed and buried near his ancestors' graves. In this regard, the Vhavenḓa say "Mudzimu o vhuya hayani", which means "God has come back home". Through this illustrative example, dictionary users are shown an aspect of Vhavenḓa culture they may not have realised if the example had not been provided. *Kuvhiwa* in this regard can be explained as: a sheep is killed and its head buried near the graves of the deceased's ancestors. *Kuvhiwa* in means replaced. The vocabulary of a language can provide some evidence of what is considered culturally important (Kavanagh 2000:103).

3.2 Idioms

Van Warmelo uses not only simple phrases to illustrate the meaning of lexical entries. He also makes use of idioms. In defining idioms Guma writes:

Idioms are characteristic indigenous expressions, whose meanings cannot be ordinarily deduced or inferred from a knowledge of individual words that make them up. They are native to a language, and have the stylistic effect of giving it a typical native ring that is characteristic of its mode of expression (Guma 1977:66).

Many idioms in Tshivenḓa depict the way the Vhavenḓa live, including their cultural beliefs. Even if their meanings cannot be ordinarily deduced from the knowledge of individual words that make them up, a number of idiomatic expressions have been used to supplement the information on the meaning of a lexical entry in a definition. Some idiomatic expressions used by Van Warmelo are *u bva dzimamuḓi* (go out of the village) and *u tevhela maḓamu* (collect fees due to a chief for Vhusha rites). The lexical entry *dzimamuḓi* is defined as follows:

(4) dzima-muḓi (or dzimu-muḓi, dzimuḓi) in -bva -- go outside the village (sc. to stay there during an illness, to get fresh air, to be undisturbed, or for other reasons); vha mu bvisa -- they are taking the patient outside the village, away to relatives (Van Warmelo 1989:40)

Dzimamuḓi is a compound noun meaning "going outside the village". One may go outside the village for various reasons, such as to stay away during an illness, to get fresh air, or to be undisturbed. However, the Vhavenḓa believe that when one is ill, one is more likely to get better if one goes outside the village. The illustrative example that serves as a vehicle to teach the traditional way of doing things is *vha mu bvisa muḓi* (they take him/her outside the village). According to the Vhavenḓa, when a person is seriously ill, he/she will not be healed if kept in the village. This is attributed to many factors, among them witches that might live in the village. The Vhavenḓa believe that a person is bewitched by someone closest to him/her, i.e. a relative. When he/she leaves the village, relatives will not have the opportunity to exacerbate the illness. The Vhavenḓa even have a proverb to support this belief: *Mutsinda ndi khwine shaka ndi bulayo* (A stranger is better than a relative; a kinship may be the death of one). This means that one will be safer outside the village than within it. The Vhavenḓa trust strangers more than they do kinsmen.

Another example of an idiomatic expression used as an illustrative example to supplement the information in the definition of a lexical entry is *U tevhela maḓamu* (collect fees due to a chief for Vhusha rites). The lexical entry *ḓamu* is defined as follows:

(5) ḓamu 5 1 udder, female breast, hence e.g. thungo ya ḓamuni the mother's side (of family, of one's relatives). 2 node or hand of bananas, i.e. those growing from the same level on the stem of the bunch. 3 -tevhela maḓamu (follow the breasts) collect fees due to a chief for Vhusha rites, from girls who have gone through these rites elsewhere. 4 -vha na -- be in calf, in foal (ruminants, equines) (Van Warmelo, 1989:20).

The English equivalents of *ḓamu* are udder and female breast. However, the word can be used in different contexts to mean different things. Our concern in this discussion is with the context in which it has been used in meaning 3. *-tevhela maḓamu*, loosely translated as follow the breasts. This reveals a cultural aspect of the life of the Vhavenḓa. When Vhavenḓa children grow up they undergo various rites to prepare them for the challenges of adult life; in this case it is the *Vhusha* rite. *Vhusha* is a puberty rite for girls. In order to attend this right, girls must pay a fee. *U tevhela maḓamu* means to collect fees due to the chief for the *Vhusha* rite from those girls who have undergone these rites elsewhere. Girls who attend these rites in other villages are made to pay a fee to their chief, which is referred to as *maḓamu*. The most important point here is that learners of Tshivenḓa learn about this cultural activity through the definition of the lexical entry *ḓamu.* Besides the meanings given, a learner receives this added information.

3.3 Proverbs

Proverbs express a community's commonly held ideas and beliefs. They are concerned with those things that the people know in their daily lives, and not with things that fall outside the scope of their experience (Guma 1977:66). However, proverbs are found in various communities of the world. Comparisons of proverbs found in various parts of the world show that the same kernel of wisdom may be gleaned under different cultural conditions and languages (McHenry 1992:749). As indicated above, many proverbs have been used as illustrative examples in *Venda Dictionary: Tshivenḓa – English.* Some of these are discussed below.

(6) -vhanga 1 wrongfully lay claim to what is another's inherited right or entitlement, as vhukoma chieftainship, headship of family, ifa inheritance; dispute, challenge or fight over such right. 2 make oblique cut with axe, as to facilitate straightening a pole; make oblique cuts towards one another to remove wedge & get at inside of wood (Mbani i vhangwa mitandani the mbani bee is got out of wood with oblique cuts); hammer and flatten to make thinner, as wire. 3 cause, bring about, aggravate, as sickness or accident. Goloi yo mbangela khombo the cart did me a mischief (as by overturning). Zwi mbangela mushumo this makes work for me (that I don't want)

Prov: Ha sa vhangwa, a vhu lalami "Unless a chief has had to fight for his position, he does not keep it long" (Van Warmelo 1989:451)

The lexical entry *-vhanga* has three meanings. All these meanings have been well defined; they do not confuse the dictionary user. The lexicographer has gone a step further by explaining meaning 1 by providing an illustrative example in the form of a proverb. The proverb comments on chieftainship, that is, if his position is not fought for, the chief will not keep it long. It is a well-known fact that among the Vhavenḓa a chief is born into the position. However, according to Vhavenḓa culture, the heir should fight for this position. This is why, before the installation of a chief/king, there will be conflicts where different groups will fight for their favourite to be installed as chief/king of the people. Such conflicts make the winner a stronger person. According to Van Warmelo (1989:151), "No chieftainship that succeeds, has not survived the test of force at its inception. This quarrel over the succession forces the leading spirits to show their colours, and the victor can then eliminate his enemies. His victory convinces the passive majority that he is indeed the right man. He has "got something", and his ancestors clearly support him." At the end of the quarrels the rightful person is the victor, and his reign lasts a long time.

Another illustrative example which takes the form of a proverb is *Wa kokodza luranga, mafhuri a a tevhela* (You pull at the calabash creepers and the calabashes follow). This illustrative example is used to add meaning to the lexical entry *luranga* which is defined as follows:

(7) luranga 11 (pl. thanga 10) cultivated cucurbit, plant of any var. of pumpkin, calabash, melon
Prov. Wa kokodza luranga, mafhuri a a tevhela "You pull at the calabash creepers and the calabashes follow," i.e. to get at the important facts of a case one discusses irrelevant details (Van Warmelo 1989:151)

This proverb makes a comment on the importance of details that may seem irrelevant but that can be valuable in supporting the important facts. Irrelevant details are not supposed to be ignored when one is discussing important facts. This proverb is usually used by the Vhavenḓa to comment on the actions of individuals, mainly men. In Tshivenḓa culture when a man gets married to a woman who has children by another man, he is expected to take all her children as his. These children are not to be ignored or left behind with the in-laws; they are to be treated as his own and they in turn should address him as "father". He is not called "stepfather", nor are the children called "stepsons/daughters" because the concepts of "stepfather" and "stepson/daughter" do not exist among the Vhavenḓa. This is such an important aspect of the culture that the children change their surname/ last name and adopt that of their new father.

Among the Vhavenḓa women are traditionally regarded as minors. This is supported by some of the proverbs in use in their daily life. In the definition of the lexical entry *tsadzi,* a proverb *Khuhu tsadzi a i imbi (ambi) mutsho, i imbaho (ambaho) ndi ya nduna* sums it all up. The entry word is defined as follows:

(8) tsadzi 9, 10 of -sadzi female; of female nature, origin. O dzula thaka -- (= ya tshisadzini) he inherited from the female side, i.e. from his mother.

Prov: Khuhu tsadzi a i imbi (ambi) mutsho, i imbaho (ambaho) ndi ya nduna "A hen does not announce the dawn, it is the cock that crows" i.e. a woman is a minor and may not discuss matters in public proceedings (Van Warmelo 1989: 385)

As women are considered minors, they are not usually allowed to participate at the *khoro* where public discussions are held and cases heard. Only men participate in such gatherings and they decide and hear cases on behalf of women. Women, on the other hand, accept this cultural aspect and abide by all decisions made by men.

4 Conclusion

The above exposition has shown that the *Venda Dictionary: Tshivenḓa – English* has a wealth of illustrative examples which have been used to convey cultural information to learners of the language. Cultural words exist in context. The knowledge, beliefs and practices of a society are revealed in the defi-

nitions of lexical entries. As a result, when looking for meanings of lexical entries, users of this dictionary also learn about the beliefs and practices of the Vhavenḓa. Cultural information merits inclusion in dictionaries and this can be achieved through the use of illustrations (Gangla-Birir, 2005:39). Prinsloo and Gouws (2000:139) have this to say: "Illustrative examples play a vital role in dictionaries, and the dictionary conceptualisation plan of any new lexicographic project should make provision for a systematic presentation of this data type in the data distribution structure". Bilingual dictionaries, as tools for learning foreign language and culture should include illustrative examples for the purposes of enhancing users' understanding of cultural words.

5 References

Al-Ajmi, H. (2008). The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos 18.* pp.15 – 26.

Bartholomew, D. (1995). Otomi Culture from Dictionary Illustrative Sentences. In Kachru, B.B. & Kahane, H. Tubingen: Max Niemeyer Verlag. pp.3 – 7.

Gangla-Birir, L. (2005). The Use of Pictorial Illustrations in Africa Language Dictionaries. *Lexikos 15.* pp. 38 – 51.

Guma, S.M. (1977). *The Form, Content and Technique of Traditional Literature in Sesotho.* Pretoria: Van Schaik.

Katzaros, V. (2004). The Different Functions of Illustrative Examples in Learners' Bilingual Dictionaries. In Williams, G. and Vessier, S. (eds.), *Proceedings of the Eleventh EURALEX International Congress, July 6 – 10, 2004. Volume 11.* Lorient: Université de Bretagne-Sud. pp. 487 - 494.

Kavanagh, K. (2000). Words in a Cultural Context. *Lexikos 10.* pp.99 – 118.

McHenry, R. (ed). (1992). *The New Encyclopaedia Britannica. Volume 9.* Chicago: Encyclopaedia Britannica, Inc.

Mafela, M.J. (2008). Proverbs as Illustrative Examples in a Tshivenḓa Bilingual Dictionary: A Reflection of Meaning and Culture. *South African Journal of African Languages, Volume 28, No. 1.* pp. 30 - 35

Prinsloo, D.J. & Gouws, R.H. (2000). The Use of Examples in Polyfunctional Dictionaries. *Lexikos 10.* pp.138 – 156.

Van Warmelo, N.J. (1989). *Venda Dictionary: Tshivenḓa – English.* Pretoria: J.L. Van Schaik.

Whitcut, J. (1995). Taking it For Granted: Some Cultural Preconceptions in English Dictionaries. In Kachru, B.B. & Kahane, H. Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta. Tubingen: Max Niemeyer Verlag. pp.253 – 257.

Corpus, Parallélisme et Lexicographie Bilingue

Adriana Zavaglia, Gisele Galafacci
Université de São Paulo
zavaglia@usp.br, gisele.galafacci@usp.br

Résumé

Dans le cadre de la lexicographie bilíngue brésilienne, surtout sur le statut de la relation entre l'entrée et ses traductions, il n'est pas rare que les dictionnaires proposent à l'usager une liste de traductions possibles pour une entrée sans contextualisation ni rubrique d'usages. L'usager de ses ouvrages se pose fréquemment des questions sur l'usage le plus approprié, sur la frontière entre les différents sens des mots proposés et leur champ notionnel. Les traductions répertoriées et décontextualisées n'aident donc pas à résoudre les ambiguïtés concernées. Pour essayer de collaborer là-dessus, nous proposons une approche différentielle (portugais - français) de la lexicographie bilingue à l'aide de la linguistique de corpus. Composé d'originaux et de traductions, le corpus parallèle utilisé dans ce travail est à la base de la contextualisation authentique de la polysémie des entrées, des définitions et des exemples bilíngues. Compte tenu de ces aspects, ce travail fournit des commentaires sur les méthodologies des dictionnaires bilingues consultés en considérant le mot grammatical « bem » du portugais brésilien et présente la méthodologie que nous appelons « parallèle » et les conséquences pour l'usager concernant l'entrée en question, en particulier dans les contextes d'apprentissage du français langue étrangère et de la traduction.
Mots-clés: traduction; corpus parallèle; lexicographie bilingue

1 Introduction

Dans le cadre de l'élaboration de dictionnaires et glossaires bilingues, il n'est pas rare que les dictionnaires disponibles - monolingues, bilingues ou multilingues - soient la principale source du lexicographe, en plus de ses propres connaissances linguistiques, méta-linguistiques et extra-linguistiques. Cependant, cette méthodologie de travail, encore courante aujourd'hui au Brésil pour les langues portugais-français, objet de notre intérêt, présente des résultats insuffisants. Comme l'a déjà dit Baldinger (1973: 63) dans le cadre de la lexicographie historique, et ceci peut être adapté à notre objectif, « les dictionnaires se copient [et se traduisent] mal… ». Par conséquent, les références aux méthodologies adoptées pour la constitution de la macro- et la micro-structure des œuvres lexicographiques bilingues ne sont pas claires, pour ne pas dire nébuleuses, dans les paratextes de ces œuvres (introductions, avant-propos, préface, etc.) ; les critères de sélection et d'enregistrement n'étant donc pas mis en évidence.

De ce fait résulte une persistance des problèmes observés dans la micro-structure de ces ouvrages, surtout dans le cas où l'utilisateur se voit proposer une liste de traductions possibles pour une entrée sans contextualisation ni rubrique d'usages. Face à cet éventail d'options pour traduire l'entrée, l'usager est dès lors confronté à un dilemme: quel est l'usage le plus approprié? Comment délimiter la frontière entre les différents sens des mots? Dans quel champ notionnel ces traductions possibles s'insèrent-elles?

En appliquant les concepts de correspondance et d'équivalence (Delisle 2003) à la micro-structure des articles bilingues, il peut être déduit que les traductions répertoriées et décontextualisées correspondraient à une entrée dans un environnement statique. Mais comment rendre compte de la polysémie mouvante des mots dans les dictionnaires bilingues étant donné l'indétermination du langage? Sans oublier le facteur statique de leur description par le lexicographe et en même temps le facteur dynamique de l'énonciation du traducteur ou de l'utilisateur des dictionnaires / glossaires bilingues, comment en choisir un seul mot correspondant, en l'introduisant dans la traduction, dans un environnement dynamique d'équivalence où la question énonciative de l'aspect intersubjectif des interprétations se pose ? Et que dire encore des mots grammaticaux?

Pour essayer de collaborer là-dessus, notamment sur le statut de la relation entre l'entrée et leurs traductions dans ce contexte, nous avons proposé une approche différentielle de la lexicographie bilingue (portugais - français) à l'aide de la linguistique de corpus (Zavaglia 2004, 2006, 2008, 2009, 2010, 2012) pour traiter un problème difficile pour la lexicographie en général et pour la lexicographie bilingue plus spécifiquement: la description des mots grammaticaux. Composé d'originaux et de traductions, le corpus s'est avéré être un bon allié. Pour les dictionnaires semi-bilingues, le corpus parallèle offre aux lexicographes une contextualisation authentique de la polysémie des entrées, c'est-à-dire une contextualisation du point de vue documentaire des sources. Considérant également la grande facilité actuelle pour la construction de dictionnaires ou lexiques bilingues électroniques, des définitions et des exemples bilingues peuvent être inclus facilement dans la micro-structure des articles, ce qui diminue considérablement les ambiguïtés et ainsi, augmente le degré de fiabilité.

Les avantages d'une telle approche ne sont pas nouveaux. Comme disait déjà Langlois (1996, Chapitre I: «Les corpus et les bitextes»)[1] :

> La valeur des corpus en linguistique n'est pas à prouver. En effet, les langues étant particulièrement complexes à décrire, les chercheurs ne peuvent se fier à l'introspection seule pour les étudier. C'est pourquoi l'étude des corpus présente de nombreux avantages, que Jan Svartvik (1992: 8) et Geoffrey Leech (1992: 106) résument bien.

Et l'auteur ajoute à la fois qu' « il est clair, cependant, que les corpus ne livrent pas que des renseignements intéressants ; il faut que le lexicographe analyse ces données empiriques pour en tirer le meilleur parti possible. » (Langlois 1996, Chapitre I: "Les corpus et les bitextes"). De toute façon, les diction-

1 Les pages de la version électronique de la thèse de Lucie Langlois ne sont pas numérotées, c'est pourquoi nous n'indiquons ici que le chapitre du travail d'où la citation a été retirée.

naires bilingues portugais-français brésiliens ne sont pas encore fabriqués à partir de corpus, ce qui fait l'intérêt de notre approche.

Dans ce cadre, la traduction d'une entrée donnée n'est ni considérée comme un correspondant, ni comme un équivalent, mais plutôt comme un parallélisme. Ceci est le résultat d'une méthodologie différente de celle adoptée dans la construction des articles de dictionnaires bilingues français-portugais brésilien disponibles. Compte tenu de ces aspects, ce travail (1) fournit des commentaires sur les méthodologies des dictionnaires bilingues consultés et leurs conséquences pour l'usager en considérant le mot grammatical *bem* du portugais brésilien et (2) présente la méthodologie que nous appelons «parallèle» et les conséquences pour l'usager concernant l'entrée en question, en particulier dans les contextes d'apprentissage du français langue étrangère et de la traduction.

2 Observations Générales sur les Dictionnaires Bilingues Consultés

Pour la réalisation de ce travail, nous avons sélectionné les ouvrages disponibles aux étudiants de langue française ou de traduction au Brésil, ceux qui sont utilisés comme outil d'apprentissage dans les cours de langue ou de traduction. Pour l'analyse, nous avons travaillé sur les contenus informationnels de l'entrée *bem* présentés dans les dictionnaires sélectionnés (voir tableau 1).

D'emblée, il est possible de repérer la grande variation de l'extension de contenu informationnel présenté par les différents ouvrages. En plus, ceux-ci ne fournissent pas d'informations précises aux usagers quant à leurs critères de sélection et d'enregistrement de données mis en œuvre lors de la production de ces outils d'apprentissage. En général, il n'existe aucune indication sur leurs pages indiquant la source des options offertes (textes authentiques et leurs traductions, dictionnaires monolingues ou bilingues?). Les lexicographes n'assument pas non plus la responsabilité des traductions de leurs dictionnaires. Nous pouvons dire que ces œuvres conservent une posture traditionnelle, plus impressionniste que scientifique, en ce qui concerne les options de traduction apparaissant dans la microstructure des articles: il s'agit d'apparier un mot avec un autre de façon directe. Cette perspective, très simpliste et aussi illusoire, que nous ne pouvons répertorier ici par manque d'espace, est un euphémisme pour l'équivalence, car il n'y a pas, dans la relation entre deux unités linguistiques, de correspondance directe au niveau des langues.

Cependant, après avoir observé les micro-structures de *bem*, nous pouvons avancer qu'aucun dictionnaire n'a présenté des citations pour exemplifier l'usage des correspondants proposés. De ce fait, il y a un grand désaccord entre les objectifs de conception de ces outils et les critères utilisés dans leur production. Dès lors, si la conception de dictionnaire bilingue est considérée comme outil d'apprentissage, ce désaccord est inadéquat aux buts pédagogiques intrinsèques. Selon Tarp:

> Un dictionnaire d'apprentissage est un dictionnaire dont le vrai but est de répondre aux besoins d'informations lexicographiques qui sont importantes aux étudiants dans une série de situations extra-lexicographiques pendant leur processus d'apprentissage d'une langue étrangère. (Tarp 2006: 300, notre traduction)

Si les dictionnaires sélectionnés sont utilisés au Brésil dans des contextes d'apprentissage (de langue, de traduction), il est en effet moins productif de proposer une liste de correspondants sans une contextualisation de leur usage aux étudiants que de leur proposer un seul correspondant. Ceci s'explique par le fait que le manque d'exemples les empêche de choisir l'option la plus adéquate pour la construction du sens envisagé, surtout dans les contextes de production oraux et écrits. Même si ces dictionnaires n'ont pas été conçus avec ce but précis, ils sont utilisés dans ce contexte, et c'est à partir de ce contexte, apprentissage/usager, que nous avons repéré les problèmes déjà mentionnés.

Ainsi, en considérant chacun des dictionnaires consultés, avec quelques exceptions concernant Gálvez (2008) et Burtin-Vinholes (2003), qui présentent une quantité raisonnable d'exemples parmi les ouvrages analysés, nous pourrions dire qu'ils ne se fondent pas sur de vraies bases lexicographiques pour ce qui est de la conception de leurs micro-structures. Cependant, ceci ne remet pas en cause le fait que ces dictionnaires sont des sources importantes d'information pour les étudiants et les professionnels de la langue en général au Brésil. Mais, à présent, avec les progrès scientifiques et technologiques applicables à la lexicographie, les dictionnaires ont besoin de se construire sur une base théorique solide qui dirige sa constitution interne, avec une base de données consistante permettant l'extraction du corpus en tenant compte des valeurs de fréquence, d'usage, de complexité et de cohérence linguistique. Ceci est également le cas pour l'emploi d'outils informatiques robustes qui conduisent à une meilleure utilisation des données à analyser, à la fois en termes de temps et de qualité. C'est pour cette raison que nous nous concentrons sur ce problème méthodologique en essayant, à partir de certaines expériences, de présenter des solutions possibles.

Avolio & Faury 2002	bem sm 1 bien. 2 bens pl biens. adv bien. ela dança muito bem/elle danse très bien. bem feito! bien fait! muito bem! bravo! nem bem nem mal ni bien ni mal. se bem que bien que. Veja nota em mieux.
Burtin-Vinholes 2003	BEM, s. m. Bien, bon, utile, avantageux. Bienfait, grâce; service; avantage; profit. POSSUIR ALGUNS –NS, posséder des biens. PAGAR O – COM O MAL, rendre le mal pour le bien. PESSOA DE –, honnête personne. FAZER O –, faire le bien, être charitable. FAZER –A ALGUÉM, rendre service à quelqu'un. QUERER O – DE ALGUÉM, désirer le bien à quelqu'un. QUERER – A SEUS PAIS, avoir de l'amour, de l'attachement pour ses parents. LEVAR A –, approuver. MEU –!, mon amour! POR –,volontiers.
Florenzano n.d	BEM, adv. Bien, beaucoup, fort, très. Convenablement, comme il faut. Certainement, assurément. ESTÁ –,c'est bien, c'est bon. AINDA –, à la bonne heure. É – FEITO, c'est bien fait. OS NEGÓCIOS VÃO –, les affaires vont bien. ELE ESTÁ MUITO –, il est dans l'aisance, il a de la fortune. MUITO –, très bien. NÃO VI MUITO –, je n'ai pas bien vu. NÃO A CONHEÇO –, je ne la connais pas beaucoup. NÃO SEI LÁ MUITO – O QUE ELE DISSE, je ne sais pas au juste ce qu'il a dit. – QUE, quoique, bien que.
Gálvez 2008	bem, s.m. e adv. bien.

Marote 2004	bem ◊ m bien m; praticar o bem faire le bien. ◊ adv 1. [ger] bien; dormiu bem? tu as bien dormi?; fez bem? tu as bien fait?; sente-se bem? tu te sens bien?; estar bem être bien; [de saúde] aller bien; queria uma bebida bem gelada je voudrais une boisson bien glacée; quero um quarto bem quente je veux une chambre bien chaude; é um quarto bem grande c'est une chambre bien grande; é um lugar bem bonito c'est un endroit bien beau; foi bem ali
Rónai 1989	bem. S.m. 1. Bien. bens. Pl. 2. Chose matérielle susceptible d'appropriation, propriété, possession, domaine. Bens de raiz. Biens-fonds..Fazer uma coisa por bem. Faire une chose pour le bien. Haver por bem. Daigner, vouloir bien. Levar a bem. Trouver bon, prendre en bonne part. Meu bem. Mon chéri, ma chérie. Pagar o bem com o mal. Rendre le mal pour le bien. Adv 3. Bien, comme il faut. Bem mais. Beaucoup plus. Ainda bem! À la bonne heure! Está bem! C'est bon! c'est bien! Falar bem de alguém. Dire du bien de quelqu'un. Pois bem! Eh bien! Querer bem a alguém. Aimer bien quelqu'un, avoir une grande estime pour quelqu'un. Se bem que Loc. conj. Bien que, quoique.
Signer 1998	bem adv. Bien ; – que loc. conj. quoique, bien que; ainda –! à la bonne heure!; adj. bien; s.m. bien; por – ou por mal bon gré mal gré; pl. biens, fortune f.; bens imóveis biens immeubles.
Valdez 2000	bem adv bien; bon; juste. está bem = c'est bien/c'est bon. falar bem = dire du bien. bem no meio = au beau milieu. bem falante = beau parleur. é bem do lado = c'est juste à côté. loc conj se bem que = encore que. nm bens de consumo = biens de consommation. bem imóvel = bien-fonds. bem de raiz = immeuble. bens trazidos pelo cônjuge, pelo sócio, etc. à sociedade = apport. pl avoirs; patrimoine. Interj muito bem! = bravo!. Fam. tudo bem = ça colle/ça roule.

Tableau 1: Entrée *bem* dans les dictionnaires bilingues.

3 Corpus et Parallélisme

Les outils d'informatique dédiés à la manipulation de données linguistiques rendent possible le renouvellement de la méthode de production de dictionnaires. Ils permettent le stockage des corpus constitués d'exemples d'usage des unités lexicales qui peuvent être récupérés par les lexicographes avec le but de contextualiser les propositions de correspondance suggérés dans les entrées des œuvres lexicographiques. De plus, le travail sur les occurrences dans un corpus bien adapté au type de dictionnaire qui se développe peut aussi aider dans le choix des entrées qu'il en aura et dans la composition de leurs micro-structures.

Cette possibilité de manipulation et récupération de donnés linguistiques authentiques de manière rapide et fiable a contribué au développement d'un nouveau concept de dictionnaire bilingue, nommé dictionnaire bilingue contrastif. Selon Durão (2009: 18), un dictionnaire bilingue contrastif diffère d'un dictionnaire bilingue car il ne se limite pas à des propositions d'équivalence, mais fournit en plus des informations qui permettent à l'usager de confronter les caractéristiques constitutives, les règles d'utilisation et les traits sémantiques des unités lexicales dans les deux langues. Cette approche permet d'exploiter les possibilités de transfert et offre des avertissements sur des possibles in-

terférences d'une langue à l'autre, car il est destiné aux personnes qui apprennent une langue étrangère avec la même langue maternelle ayant comme but de reconnaitre les différences entre les deux langues.

Ainsi, nous proposons un travail lexicographique pour les langues portugais brésilien-français qui utilise des corpus parallèle, c'est-à-dire, comme nous l'avons déjà dit, une base de données composée de textes authentiques en portugais et leurs respectives traductions en français. Ce corpus, nommé CorPPoFranco – Corpus Parallèles du Portugais au Français de la Francophonie, est en train d'être développé pour ce projet de recherche. Avec le but d'être utilisé dans la production de dictionnaires bilingues, ce corpus aura l'adéquation nécessaire à la réalisation de cet objectif, car, selon Sardinha (2004: 29), la recherche avec l'appui de corpus a du sens seulement si les questions posées et les corpus sont adéquats aux objectifs du travail.

Cette méthodologie permet de traiter de façon plus efficace l'hétérogénéité qui entoure la relation complexe entre deux langues, ce qui permet tant l'observation de la distribution morphologique et structurale d'une lexie, que la visualisation de son profil, de ses collocations et de sa prosodie sémantique. L'élaboration de paradigmes adoptés dans le développement de la recherche sur la lexicographie bilingue dans cette perspective différentielle inclut des procédures spécifiques de la linguistique de corpus (comme par exemple la concordance et l'alignement), de la linguistique énonciative, de la lexicographie et des études de traduction. De ceci a émergé une nouvelle entité empirique (les traductions offertes dans les micro-structures) et théorique (la nature et la fonction des traductions). Cette nouvelle entité de la lexicographie bilingue demande un nom, une définition et une conceptualisation, que voici :

> Le Parallélisme est la relation entre deux lexies, complexes ou non, décrite dans la micro-structure d'articles de dictionnaires bilingues de langue générale construits à partir de corpus parallèle (originaux et traductions publiées) ayant la fonction d'expliciter la variation sémantique du mot-vedette, avec ses usages spécifiques, en contexte bilingue authentique, avec ses traductions possibles.

4 Illustrations

Pour illustrer cette méthodologie parallèle, nous avons utilisé un échantillon du CorPPoFranco, dont nous avons sélectionné quelques occurrences de la lexie *bem* pour réfléchir sur les propositions de traductions offertes par les dictionnaires bilingues décrits sommairement avant et les parallélismes qui résultent de l'utilisation de cette méthodologie.

Nous n'allons prendre, pour manque d'espace, que quelques aspects sur la relation entre *bem* et *bien* remarqués sur cette traduction directe suggérée par tous les dictionnaires bilingues consultés: les lexicographes n'ont pas pris en compte dans leurs articles les informations sur les usages spécifiques, les

questions de syntaxe et de collocation concernant cette relation. Les exemples ci-dessous confirment leur relevance par rapport:

- à la morphologie :

(1) À mesa do café eu me enquadrava com meu robe branco, meu rosto limpo e bem esculpido, e um corpo simples

(1a) Assise à ma table, dans ma robe de chambre blanche, avec mon visage net et bien sculpté, mon corps tout simple

(2) Revi o rosto preto e quieto, revi a pele inteiramente opaca que mais parecia um de seus modos de se calar, as sobrancelhas extremamente bem desenhadas, revi os traços finos e delicados que mal eram divisados no negror apagado da pele.

(2a) Je revis le visage noir et placide, la peau parfaitement opaque qui semblait davantage une autre façon de se taire, les sourcils extrêmement bien dessinés, je revis les traits délicats et gracieux, mais à peine lisibles dans la noirceur éteinte de la peau.

Dans les exemples ci-dessus, avec *bem* adverbe traduit par *bien*, on peut voir, au-delà de sa fonction d'intensif, son positionnement, devant le participe en fonction d'adjectif, mais surtout son caractère invariable, car le parallélisme (1)/(1a) le montre avec un participe adjectif au singulier et le (2)/(2a), au pluriel.

- au positionnement :

(3) Mas era tão inegável sentir aquele nascimento de dentro da poeira - que eu não podia senão seguir aquilo que eu bem sabia que não era loucura, era, meu Deus, uma verdade pior, a horrível.

(3a) Mais c'était si incomparable de sentir cette naissance au sein de la poussière que je ne pouvais que suivre ce dont je savais bien que ce n'était pas de la folie; c'était, mon Dieu, une vérité pire, l'horrible vérité.

(4) Queria saber se era verdade que ele lhe arrancara a calcinha com os dentes, queria saber isto e aquilo, queria saber o que já sabia muito bem

(4a) Il voulait savoir s'il était vrai qu'il lui avait arraché sa petite culotte avec les dents, il voulait savoir ceci et cela, il voulait savoir ce qu'il savait déjà parfaitement

Dans les exemples ci-dessus, nous voyons une question d'usage en ce qui concerne les différentes positions de *bem* avec le verbe *saber*: devant le verbe, avec un usage absolu; après le verbe, dans une construction avec un adverbe qui l'intensifie. Dans les traductions, il est possible de remarquer tout de suite les différences en français: *bien* avec *savoir* après le verbe, avec la possibilité d'un autre parallélisme quand il est intensifié, comme dans (4a).

- à la polysémie :

(5) De meu próprio mal eu havia criado um bem futuro.

(5a) J'avais, avec mon propre mal, crée un bien à venir.

(6) A partilha dos bens foi realizada como previsto.

(6a) Le partage des biens a été réalisé comme prévu.

(7) No quarto de Ariela, Benjamim pagaria a Lorna para prosseguir falando "tu és meu bem, tu és meu bem, tu és meu bem"

(7a) Dans la chambre d'Ariela, Benjamin paierait Lorna pour qu'elle continue à dire « tu es mon ange, tu es mon ange, tu es mon ange »

Dans les parallélismes ci-dessus, *bem*, nom, est traduit par *bien* ; et pourtant, les contextes montrent clairement que, du point de vue du sens, la relation n'est pas biunivoque, mais polysémique, car dans (5)/(5a), le « bem », au-delà d'avoir son origine ancrée dans le « mal », ce qui exclut par elle-même toute relation simpliste d'antonymie, est un bien abstrait parmi d'autres qui ne peut pas être déterminé, à part son caractère futur ; dans (6)/(6a), *bens*, pluriel, couramment utilisé dans le vocabulaire juridique, se rapporte aux biens concrets, meubles ou immeubles. En plus, le parallélisme (7)/(7a) montre que *bem* concret peut faire référence à une personne, généralement quand il est précédé d'un possessif, avec une possibilité de traduction, *ange*, dans « mon ange ».

D'autres parallélismes ont été trouvés dans le corpus. Cependant, leurs usages spécifiques sont attachés à des structures particulières, lesquelles valident la construction du sens envisagé. Pour cela, la simple suggestion de ces mots dans un article de dictionnaire dans une liste de possibilités se révèle improductive pour l'usager qui a besoin d'informations plus détaillées sur chaque traduction proposée. Les parallélismes montrent que, en fonction de l'adverbe que *bem* intensifie, une traduction différente est convoquée (« bem devagar », « tout doucement » ; « bem mais », « beaucoup plus »), de même dans d'autres relations, *bem/bon* (« Bem, além de fixar... » / « Bon, en plus de fixer ... »), ou formations (lexies complexes: « bem-estar »/« bien-être » ; locutions : « se bem que »/« pourtant » ; binômes : « bem ou mal »/« bien ou mal »). Voici quelques exemples extraits de l'ouvrage *A paixão segundo G.H*, de Lispector (1964/1998) et de sa respective traduction:

(8) Aquelas pessoas que, só elas, entenderão bem devagar que este livro nada tira de ninguém.

(8a) Ces personnes, et elles seules, comprenont tout doucement que ce livre n'enlève rien à personne.

(9) Desculpa eu te dar isto, eu bem queria ter visto coisa melhor.

(9a) Pardonne-moi pareil cadeau, je préférerais tellement avoir vu une chose plus agréable.

(10) É bem mais que uma elegância.

(10a) C'est beaucoup plus qu'un snobisme.

(11) Mas tendo aos poucos, por meio de dinheiro razoavelmente bem investido, enriquecido o suficiente, isso impediu-me de usar essa minha vocação: não pertencesse eu por dinheiro e por cultura à classe a que pertenço, e teria normalmente tido o emprego de arrumadeira numa grande casa de ricos, onde há muito o que arrumar.

(11a) Mais, pour avoir su placer adroitement mon argent, j'ai acquis peu à peu une certaine alliance, ce qui m'a empêchée de réaliser ma vocation profonde: si je n'appartenais pas, par la culture et l'argent, à la classe à laquelle j'appartiens, j'aurais dû normalement être femme de chambre chez des gens riches, dans une grande maison où il y a beaucoup à ranger.

(12) Animei-me com uma idéia: aquele guarda-roupa, depois de bem alimentado de água, de bem enfartado nas suas fibras, eu o enceraria para dar-lhe algum brilho, e também por dentro passaria cera pois o interior devia estar ainda mais esturrado.

(12a) Une idée me réconforta: cette armoire, une fois complètement imbibée, gorgée d'eau dans toutes ses fibres, j'allais la cirer pour lui donner un peu de brillant et je passerais de la cire à l'intérieur aussi car l'intérieur devait être encore plus calciné.

(13) De encontro ao rosto que eu pusera dentro da abertura, bem próximo de meus olhos, na meia escuridão, movera-se a barata grossa.

(13a) Tout contre mon visage passé dans l'ouverture de la porte, tout près de mes yeux, dans la demi-obscurité, un énorme cafard avait bougé.

(14) Até que - enfim conseguindo me ouvir, enfim conseguindo me comandar – ergui a mão bem alto como se meu corpo todo, junto com o golpe do braço, também fosse cair em peso sobre a porta do guarda-roupa.

(14a) Jusqu'à ce que – parvenant enfin à m'entendre, parvenant enfin à me donner un ordre – je levai la main très haut comme si mon corps tout entier, entraîné par le mouvement du bras, allait tomber lui aussi de tout son poids sur la porte de l'armoire.

(15) Bem, além de fixar as dunas com eucaliptos, eu tinha que não esquecer, se viesse a ser necessário, que o arroz prospera em solo salobre, cujo alto teor de sal ajuda a desbastar; disso eu também me lembrava das leituras de antes de dormir que eu, de propósito, procurava que fossem impessoais para me ajudarem a adormecer.

(15a) Bon, en plus de fixer les dunnes avec des eucalyptus, il ne fallait pas que j'oublie, le cas échéant, que le riz pousse sur un terrain saumâtre dont la haute teneur en sel facilite le sarclage; j'avais aussi retenu cela de mes lectures d'avant de m'endormir que je choisissais, exprès, impersonnelles pour m'aider à trouver le sommeil.

(16) O tédio profundo – como um grande amor – nos unia. E na manhã seguinte, de manhã bem cedo, o mundo se me dava.

(16a) L'ennui profond – comme un grand amour – nous unissait. Et le lendemain, le matin très tôt, le monde s'offrait à moi.

(17) Usarei, sim, o vestido azul novo, que me emagrece um pouco e me dá cores, telefonarei para Carlos, Josefina, Antônio, não me lembro bem em qual dos dois percebi que me queria ou ambos me queriam, comerei crevettes ao não importa o quê", e sei porque comerei crevettes, hoje de noite, hoje de noite vai ser a minha vida diária retomada, a de minha alegria comum, precisarei para o resto dos meus dias de minha leve vulgaridade doce e bem-humorada, preciso esquecer, como todo o mundo.

(17a) Je mettrai, oui, ma nouvelle robe bleue qui m'amincit un peu et me donne des couleurs, je téléphonerai à Carlos, Joséfina, Antonio, je ne me souviens qu'il me désirait, ou si c'était tous les deux, je mangerai des crevettes « à la je ne sais quoi », et je sais pourquoi je mangerai des crevettes, ce soir, ce soir je vais reprendre ma vie quotidienne, celle de ma joie ordinaire ; j'aurai besoin

pour le restant de mes jours de ma légère vulgarité douce et bon enfant, j'ai besoin d'oublier comme tout le monde.

Cependant, comment inclure dans la micro-structure bilingue portugais-français de *bem* toutes les informations des observations et analyses faites à partir de cette méthodologie parallèle? Une fois que l'usager cherche une solution immédiate dans un dictionnaire bilingue? Dans notre article, qui sera électronique et disponible en ligne, cette solution immédiate est présentée après la transcription phonétique et la classification morphologique, par ordre de fréquence dans le corpus. Par exemple:

bem [bẽ]

adv. bien, très, tout, beaucoup, complètement, tellement, adroitement.

n. bien, pl. bens, biens.

ap. bon.

Pour avoir plus de détails (de positionnement, d'usage ou de contexte), l'usager peut cliquer sur l'une des traductions présentées. Par exemple, en cliquant sur « beaucoup », une nouvelle fenêtre s'ouvre avec les informations suivantes :

[DEFINIÇÃO] Intensificador comparativo [DEFINITION] intensif comparatif

[COLOCAÇÃO] bem mais [COLLOCATION] beaucoup plus

[SINÔNIMO] pt. muito mais [SYNONYME] fr. bien plus

[POSIÇÃO] bem + adverbe [POSITIONNEMENT] beaucoup + adverbe

[USO] neutro [USAGE] neutre

[CONTEXTO] É bem mais que uma elegância. [CONTEXTE TRADUIT] C'est beaucoup plus qu'un snobisme.

5 Conclusion

Implicitement, l'utilisateur aura une description bilingue de *bem* en raison du co-texte et du contexte dans lequel cette lexie apparaît. Cela lui permettra également la possibilité de saisir le caractère idiomatique qui tourne autour de ses parallélismes, à savoir les relations authentiquement bilingues entre *bem* et ses traductions. Le développement du concept de parallélisme, comme mentionné plus haut, a une origine naturelle lors de la mise en place de la méthodologie présentée, radicalement différente de celle observée dans les dictionnaires bilingues portugais-français utilisées au Brésil. De plus, le rapport entre l'entrée et les traductions proposées dans la micro-structure des articles, il faut le dire, n'est pas fixe. En effet, ce rapport se déplace et se modifie en fonction du co-texte et du contexte, comme deux lignes parallèles qui affluent vers l'infini sans envahir le domaine l'une de l'autre, conservant chacune ses propres caractéristiques, et met en avant les valeurs de la fréquence de l'utilisation de la lexie ou de sa complexité et, plus précisément, l'explicitation de sa polysémie, qui se construit et se réalise toujours dans un contexte véritablement bilingue.

6 Références

Avolio, J.C., Faury, M.L. (2002). *Michaelis Minidicionário Francês-Português, Português-Francês*. São Paulo: Melhoramentos.

Baldinger, K. (1973). Le DEAF en tant que dictionnaire diachronique. Problèmes théoriques et pratiques, *Meta, 18,* 61-85.

Burtin-Vinholes, S. (2003). *Dicionário Francês-Português/Português-Francês*. São Paulo: Ed. Globo.

Delisle, J. (2003). *La traduction raisonnée*. Ottawa: University of Ottawa Press.

Durão, A.B.A.B. (ed.) (2009). *Por uma Lexicografia Bilíngue Constrastiva*. Londrina: UEL.

Florenzano, E. (n.d.) Dicionário Ediouro Francês-Português/Português-Francês.

Gálvez, J.A. (ed.) (2008). Dicionário Larousse Francês-Português, Português-Francês: mini. São Paulo: Larousse do Brasil.

Langlois, L. (1996). *Bitexte, bi-concordance et collocation*. Thèse de doctorat. Université d'Ottawa. En ligne http://www.dico.uottawa.ca/theses/langlois [13/08/2013].

Lispector, C. (1998). *A paixão segundo G.H*. Rio de Janeiro: Rocco. (Original work published 1964).

Lispector, C. (1978). *La passion selon G.H*. Trad. Claude Farny. Paris: Des Femmes.

Marote, J.T.O. (2004). Minidicionário Francês-Português, Português-Francês. São Paulo: Ática.

Rónai, P. (1989). *Dicionário Francês-Português/Português-Francês*. Rio de Janeiro: Nova Fronteira.

Signer, R. (1998). Dicionário Brasileiro Francês-Português, Português-Francês. São Paulo: Oficina de Textos.

Tarp, S. (2006). Lexicografia de Aprendizaje. In *Cadernos de Tradução, n.18*. Florianópolis: UFSC, pp. 295-317.

Valdez, J.F. (2000). *Dicionário Francês-Português/Português-Francês*. Belo Horizonte: Garnier.

Sardinha, T.B. (2004). *Linguística de Corpus*. Barueri: Manole.

Zavaglia, A. (2004). Lingüística de córpus e lexicografia bilíngüe: o caso experimental de *como* e suas traduções para o francês. In *Crop, v. 10.* São Paulo: FFLCH-USP, pp. 211-224.

Zavaglia, A. (2006). Lexicografia bilíngüe e corpora paralelos: procedimentos e critérios experimentais. In *Cadernos de Tradução, v. XVIII*. Santa Catarina: UFSC, pp. 19-39.

Zavaglia, A. (2008) Apresentação das bases do Dicionário Relacional (português-francês) DIRE. In Isquierdo, A.N., Finatto, M.J.B. (eds.). *Ciências do Léxico: Lexicologia, Lexicografia e Terminologia, v. 4*. Campo Grande: Ed. UFMS, pp. 233-254.

Zavaglia, A. (2009). Linhas gerais para a elaboração do Dicionário Relacional – DIRE (português-francês). In Rezende, L.M., Silva, B.C.D. & Barbosa, J.B. (eds.). *Léxico e Gramática: dos sentidos à construção da significação*. São Paulo: Cultura Acadêmica, pp. 185-201.

Zavaglia, A. (2010). Sinonímia e lexicografia. In Isquerdo, A. N., Barros, L. A. (eds.). *As Ciências do Léxico: Lexicologia, Lexicografia e Terminologia, v. V*. Campo Grande: Editora da UFMS, pp. 189-199.

Zavaglia, A. (2012). Por uma descrição esquemática do léxico: o caso de *mediante.* In Isquerdo, A.N., Seabra, M.C.TC. (Eds.). *As Ciências do Léxico: Lexicologia, Lexicografia, Terminologia, v. VI*. Campo Grande: Editora da UFMS, pp. 73-84.

Lexicography for Specialised Languages, Technology and Terminography

EcoLexicon

Pamela Faber, Miriam Buendía Castro
University of Granada
pfaber@ugr.es, mbuendia@ugr.es

Abstract

EcoLexicon (http://ecolexicon.ugr.es) is an environmental knowledge base which is based on the premises of Frame-based Terminology (FBT) (Faber 2009, 2011, 2012). EcoLexicon represents the conceptual structure of the specialized domain of the Environment in the form of a visual thesaurus in which environmental concepts are configured in semantic networks. The various terminological designations for a concept are offered in six languages: Spanish, English, German, French, Russian, and Greek. So far, EcoLexicon contains 3,527 concepts and 18,617 terms. It provides conceptual, linguistic, and administrative information for each entry, and we are now beginning to include phraseological information for terms as well (Buendía 2013; Buendía & Sánchez 2012, Sánchez & Buendía 2012). EcoLexicon is designed to meet the needs of different user types, such as a student of science wishing to acquire specialized knowledge about a certain concept, a translator seeking translation correspondences in a language, or a specialist interested in text production.
Keywords: knowledge base; specialized language; environment

1 Introduction

EcoLexicon (http://ecolexicon.ugr.es) is a visual online thesaurus of environmental science, which currently contains 3,527 concepts and 18,617 terms in English, Spanish, German, French, Russian and Modern Greek. It provides conceptual, linguistic, and administrative information for each entry. This information as well as the corpus of specialized texts is stored in a private database, which allows members of EcoLexicon to add, eliminate, and/or modify conceptual and terminological information. In EcoLexicon it is assumed that up to a certain level, its potential users are familiar with scientific language and its usage in English or Spanish at least, since these are the interface languages. Potential users should thus possess a good command of any of the six languages in the knowledge base, as well as a minimum of scientific knowledge (López, Buendía & García 2012: 62).

EcoLexicon is based on the premises of Frame-based Terminology (FBT) (Faber 2009, 2011, 2012), a cognitive approach to Terminology. The FBT approach to Terminology applies the notion of frame, defined as a schematization of experience (a knowledge structure), which is represented at the conceptual level and relates elements and entities associated with a particular culturally embedded scene, situation or event from human experience.

The public version of EcoLexicon is freely available online. The new version of EcoLexicon includes the following new features, which are an improvement over the previous version.

- It is compatible with all modern browsers and does not require Java.
- It has more interactive and configurable maps, which allow the users to do the following:
  - Change the scale of the map.
  - Select the relations to be represented.
  - Eliminate nodes in order to make other nodes more prominent.
  - Adjust the position of nodes.
  - Go backwards or forwards.
  - Establish a direct link to a concept or term.
  - Look up a concept or term on Google or Google Images.
- It has new representation modes:
  - Hierarchical tree structure
  - Map of the shortest path between concepts
- It allows user registration to do the following:
  - Personalize results.
  - Store user preferences between sessions.

2 The micro and macrostructure of entries in EcoLexicon

Figure 1 displays the entry for *alluvial fan* in EcoLexicon. As shown in Figure 1, when EcoLexicon is opened, three zones appear:

(1) The top bar that allows access to different functionalities, such as the term/concept search or changing the language of the interface. It also permits to personalize the search by means of the contextual domains generally associated with a specialized knowledge field (e.g. Geology, Coastal Engineering, Environmental Law, etc.). This allows users to focus on the knowledge area or specific domain and to eliminate irrelevant information. It also allows users to create an account. This permits the storage of their preferences and options, independently of the computer used to access the database.

(2) The central area that includes a dynamic network that displays the search concept/term and links it to all related items in terms of a closed inventory of conceptual relations. When users click on any of the terms or concepts in the map, the network rearranges itself. In this new map, the term/concept that was clicked on is at the center and is connected with all of the entities directly related to it. As shown (Figure 1), in the lower left corner of the map, there is a text box with captions that allow users to identify the three categories of conceptual relations in EcoLexicon: (i) hyponymic (generic-specific) relations; (ii) meronymic (part-whole) relations; (iii) non-hierarchical relations. The conceptual relations used in EcoLexicon include a set of 17 hierarchical (hypernymic

and meronymic) and non-hierarchical relations, some of which are domain-specific. The 17 relations included are the following: AFFECTS, ATTRIBUTE_OF, CAUSES, COMPOSED_OF (MATERIAL), OPPOSITE_OF, DELIMITED_BY, STUDIES, PHASE_OF, MEASURES, PART_OF, REPRESENTS, RESULT_OF, EFFECTED_WITH, HAS_FUNCTION, TAKES_PLACE_IN, TYPE_OF, LOCATED_AT. In addition, a series of tabs are included in this central area that allow the user to access: (i) the history of concepts/terms visited; (ii) the results of the most recent search, which allows users to go backwards and forwards on this list, in the same way as on web browsers; (iii) the shortest path between two concepts; (iv) concordances for a term

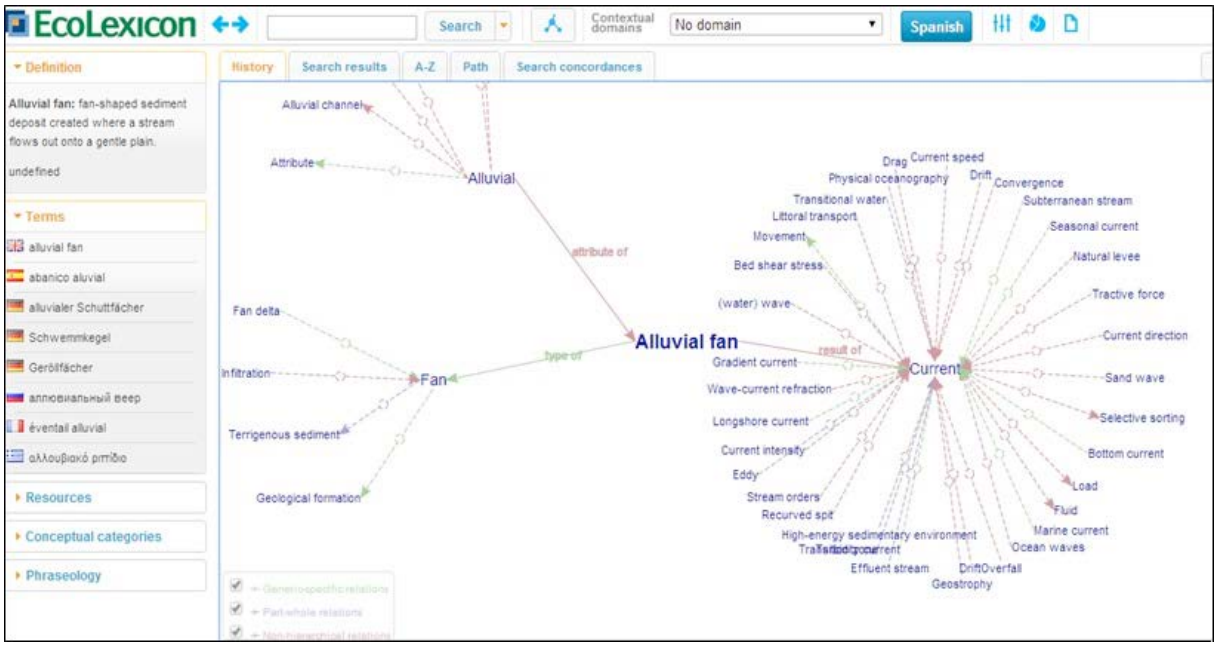


Figure 1: Entry of *alluvial fan* in EcoLexicon.

(3) The side bar (Figure 2) that shows a wide range of information about the search item:

- The terminographic definition of the concept, which encodes the most basic conceptual relations in the category template;
- The associated terms designating the concept in the six languages contained in EcoLexicon. The list is ordered by language and term type (main entry term, synonym, variant, acronym, etc.). At the left of each term, there is a small flag representing the country where its language is spoken. By clicking on each term, a new window is displayed which provides users with the following linguistic information: (i) morphosyntactic information regarding grammatical category (noun, verb, adjective or adverb); term type (main term, synonym, geographical variant, and acronym); gender (masculine, feminine, and neuter); (ii) contexts of use; (iii) phraseology.
- The associated resources, (images, documents, URLs, audiovisual material, etc.) associated with each concept/term. Regarding images, depending on the type of knowledge activated, three different kind of images are distinguished, namely, iconic, abstract, or dynamic images (Prieto, 2009; Prieto &

López 2009). The complete list of the resources for each concept is shown in this box. Users can easily identify the type of resource by means of the icons beside each of the resource listed. In order to access more information regarding the resource (title, description, source, etc.) users can place the cursor on the resource and a new window will open with all this information.

- Conceptual categories. Each concept in EcoLexicon is associated with one or more conceptual categories, which are shown as a list. If users click on one of these categories, this opens a window with a list of all the concepts included. Furthermore, this box includes a *Category hierarchy* icon, which, when clicked on, shows the concepts in a hierarchical format in which nodes can expand or retract. If one of the categories in the hierarchy is clicked on, a window appears with all the concepts associated with that category.

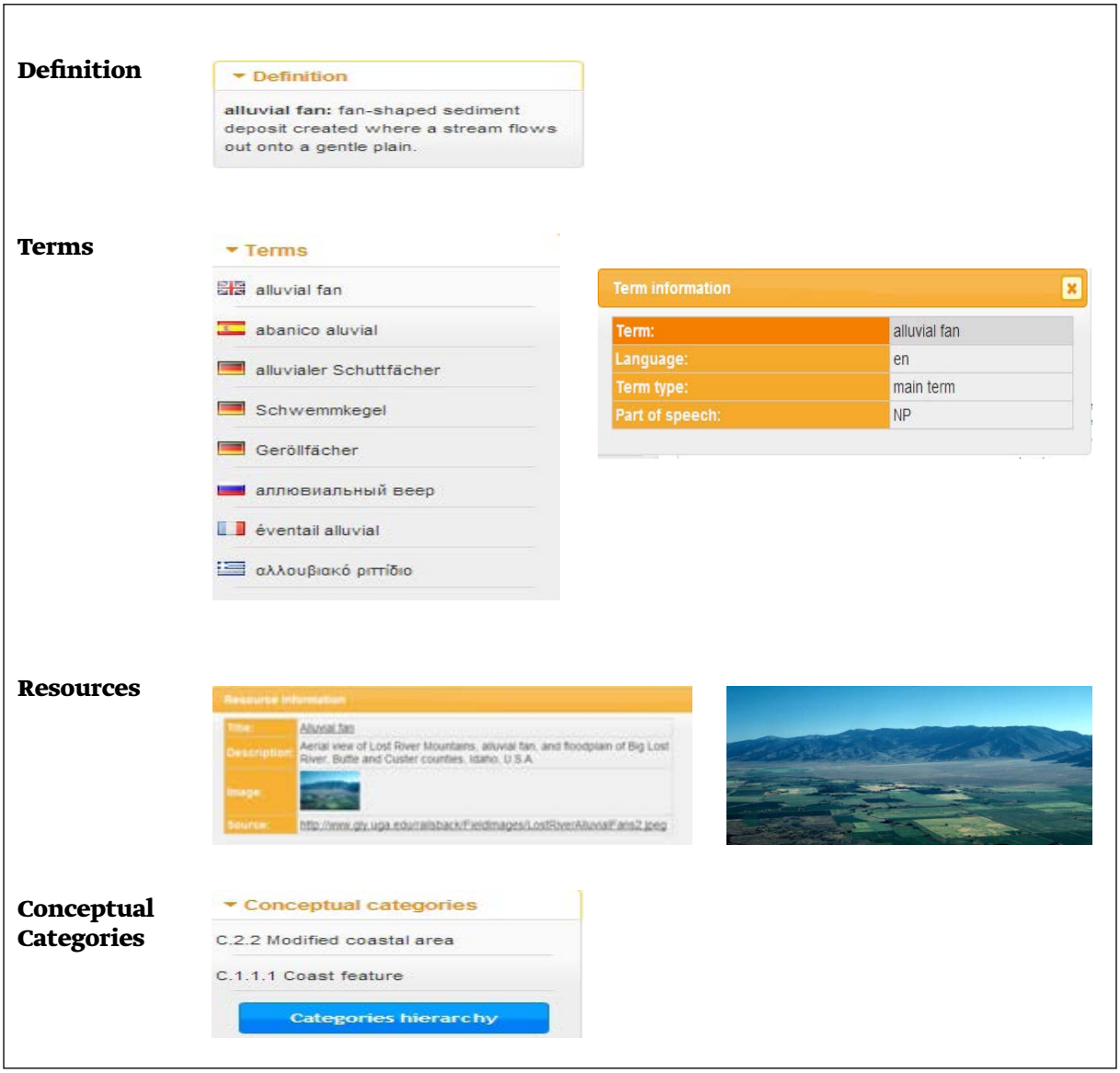


Figure 2: Extract of the side bar associated for each search item in EcoLexicon.

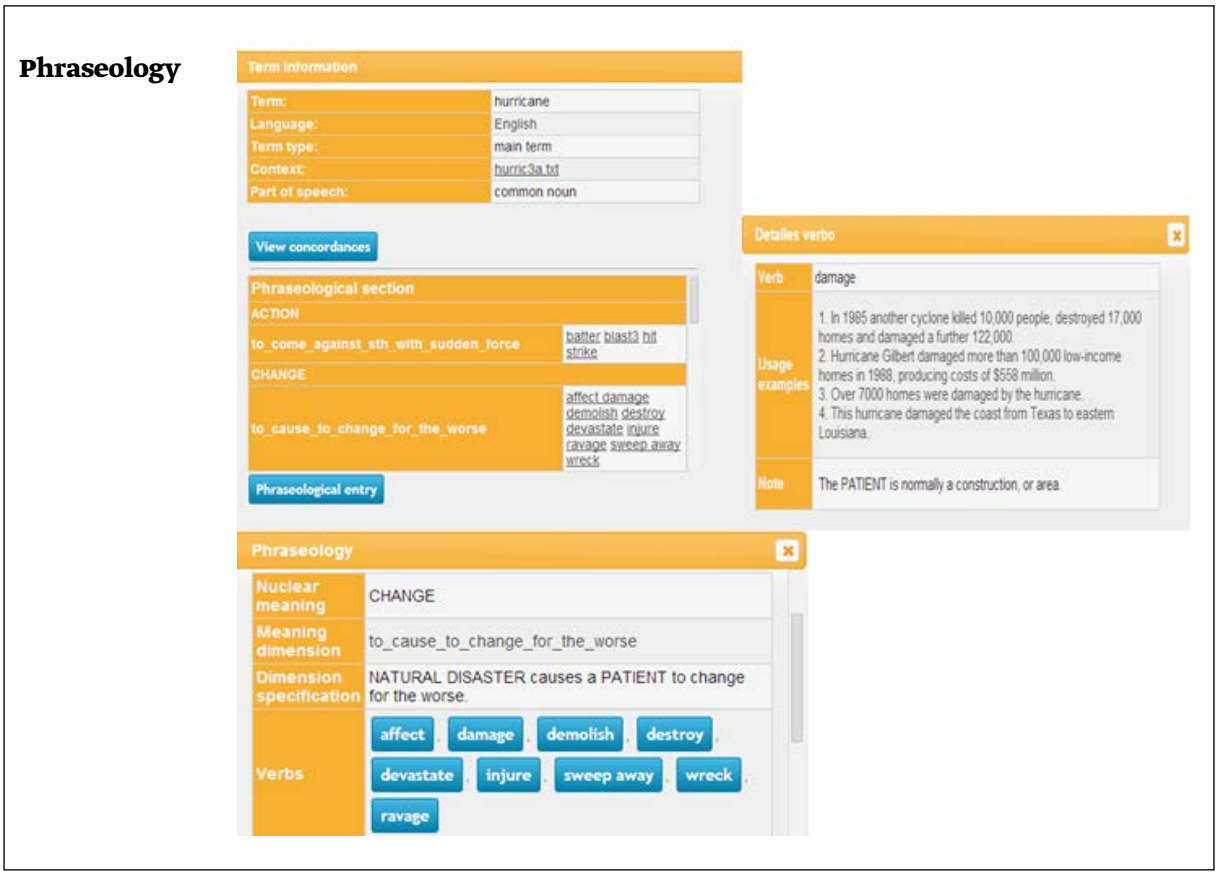


Figure 3: Extract of the phraseology box associated with *hurricane* in EcoLexicon.

- Phraseology. If the central element in the map is a term, this box shows a list of verbs most commonly used with the term. Phraseological information can also be accessed via the term entry in certain concepts. Verb collocations are classified and described according to meaning. For this reason, they were primarily classified in terms of their lexical domain (i.e. the nuclear meaning), and subsequently in terms of the frame activated within each lexical domain (i.e. meaning dimension). Once the lexical domain and frames are stated, the verbs are specified. By clicking on the verbs, the user has access to the usage sentences for the verb in question, as well as a note section with information about meaning restrictions.

3 Conclusion

As shown, Frame-based Terminology provides a full account of the information necessary to describe a specialized knowledge unit in a terminological entry. The practical application of FBT, the environmental knowledge base EcoLexicon, provides conceptual, linguistic, and administrative information for each specialized knowledge unit. To enhance knowledge acquisition,

conceptual information in EcoLexicon is stored and represented in different ways. Specialized environmental knowledge is represented by means of conceptual networks codified in terms of conceptual propositions in the form of a triplet (concept relation concept). The conceptual relations used in EcoLexicon include a set of 17 hierarchical (hypernymic and meronymic) and non-hierarchical relations. Conceptual information is also shown in the form of natural language definitions in English and Spanish, which are based on the conceptual propositions established by the concept to be defined. Additionally, domain-specific knowledge is also presented in the form of images, documents and videos, which complement the previously entered conceptual information. Linguistic information is now being enhanced by the introduction of verb phraseological information regarding each term. In a near future, we hope to be able to provide a complete phraseological description of every term.

4 References

Buendía Castro, M. (2013). Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources. PhD Thesis, University of Granada, Spain.

Buendía Castro, M., & Sánchez Cárdenas, B. (2012). "Linguistic knowledge for specialized text production". In N. Calzolari, K.Choukri, T. Declerck,, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul: ELRA, pp. 622-626.

Faber, Pamela. (2009). The Cognitive Shift in Terminology and Specialized Translation. *MonTI. Monografías De Traducción e Interpretación* 1(1), pp. 107–134.

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

Faber, Pamela (ed.). (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Berlin, Boston: Mouton de Gruyter.

López Rodríguez, C.I., Buendía Castro, M. & García Aragón, A. (2012). User Needs to the Test: Evaluating a Terminological Knowledge Base on the Environment by Trainee Translators. *Jostrans. The Journal of Specialized Translation* (18), pp. 57–76. Accessed at: http://www.jostrans.org/issue18/art_lopez.pdf. [20/12/2012].

Prieto Velasco, J.A. (2009). Traducción e Imagen: La Información Visual En Textos Especializados. Granada: Tragacanto.

Prieto Velasco, J.A. & López Rodríguez, C.I. (2009). Managing graphic information in terminological knowledge bases, *Terminology* 15(2), pp. 179-213.

Sánchez Cárdenas, B., & Buendía Castro, M. (2012). Inclusion of Verbal Syntagmatic Patterns in Specialized Dictionaries: The Case of EcoLexicon. In R. Vatvedt Fjeld, J. Matilde Torjusen (eds.) Proceedings of the 15th EURALEX International Congress. Oslo: EURALEX, pp. 554–562.

Acknowledgements

This research was carried out within the framework of the project RECORD: Representación del Conocimiento en Redes Dinámicas [Knowledge Representation in Dynamic Networks, FFI2011-22397], funded by the Spanish Ministry of Science and Innovation.

Experts and Terminologists: Exchanging Roles in the Elaboration of the Terminological Dictionary of the Brenner Base Tunnel (BBT)

Elena Chiocchetti, Natascia Ralli
European Academy of Bolzano/Bozen (EURAC research)
echiocchetti@eurac.edu, nralli@eurac.edu

Abstract

This paper presents the Italian-German Terminological Dictionary of the Brenner Base Tunnel (BBT) produced and printed in 2011 by the European Academy of Bolzano. In particular, we describe the role played by domain experts and terminologists during the elaboration of the dictionary, which was different from the one that is usually assigned to them in terminological projects. This switching of roles had several consequences on the structure and content of the dictionary, which we will discuss. We also briefly illustrate the challenges faced and how we approached specific problems, e.g. the structure of the definitions, the managing of synonyms and variants, the varied nature of terms selected for the dictionary and their specific treatment with a view to the needs of the target users. The two main target groups consist of experts and semi-experts from various professions who are confronted daily with terminology pertaining to a large array of specialised domains (from environmental to engineering to legal terminology).
Keywords: Terminological dictionary; Domain experts' role; Terminology workflow

1 Introduction

The Brenner Pass is located between Italy and Austria in a favourable position, making it one of the main thoroughfares connecting Central and Northern Europe with the Italian peninsula and the Mediterranean area. Passenger and freight traffic on the motorway and railway crossing on this rather low Alpine pass (1372 m above sea level) is particularly intense all year round and in constant growth (cf. Maino et al. 2011: 10). The ensuing social, economic and environmental consequences call for an improvement of the railway line between Munich and Verona in order to favour a modal shift from road transportation to rail transportation for goods, in order to achieve a more efficient mobility system and relieve the population along the railway line from transit traffic (Maino et al. 2011: 8). The Brenner Base Tunnel (BBT) is the most important element of this improvement project and will be 55 km long, connecting Innsbruck in Austria with Fortezza in Italy. Its completion is currently scheduled for the year 2025.

For the two neighbouring countries of Austria and Italy to be able to cooperate efficiently in the realisation of the project, all relevant documents must be available in two languages, i.e. German and Italian. This calls for a large number of translations, with technical, legal, administrative and commercial content produced over a very long period. For smooth communication, all these texts need to use correct and coherent specialised terminology pertaining to a vast array of domains in both languages. Therefore, the scope of the BBT project, its duration, the number of different professionals involved, the need for overcoming language barriers as well as legal/administrative barriers, together with the complex and varied nature of the terminology needed, represent the background for the creation of a dedicated bilingual dictionary. This dictionary is the Italian-German dictionary *Dizionario Terminologico della Galleria di Base del Brennero/Terminologisches Wörterbuch zum Brenner Basistunnel,* which aims at supporting transnational communication and cooperation for all issues related to the BBT.

The dictionary was planned and created by domain experts and terminologists in close cooperation. However, unlike standard terminology projects in which terminologists perform most of the activities and are ideally supported by domain experts as revisers and consultants, the BBT terminological dictionary was compiled by experts, while the terminologists took over the role of consultants and quality controllers. The knowledge possessed by the domain experts on the special language and conceptual structure of the domains treated was channelled and directed into a precise terminological working method. This exchange of roles with respect to the most common workflow in terminology proved to be challenging but stimulating for both groups. In section 5 we discuss the main challenges faced.

2 The BBT dictionary

The BBT dictionary was promoted and financed by the Brenner Base Tunnel Societas Europae (BBT SE). The BBT Society commissioned the creation of a bilingual dictionary to a group of mainly monolingual Italian or German speaking experts in territorial planning, land management, engineering and mobility, who sought advice from a group of bilingual terminologists. The terminologists supported the dictionary authors in defining two main aspects before starting work, i.e. the target users of the dictionary and the methodology to be adopted during compilation, so as to avoid inconsistencies and the need for later adjustments.

The structure of the definitions and the treatment of variants/synonyms, for example, consistently follow a clearly onomasiological rather than semasiological approach. The onomasiological method is typical of terminology work, as it considers the key concepts of a specialised domain and the relations existing between them as central elements for the selection and definition of the terms to be included in the dictionary. The lemmas in the dictionary being concept-based, all synonyms or variants (e.g. a full form and its corresponding acronym) are listed and defined together, since they all designate the same concept. Unlike dictionaries that follow a semasiological approach, two synonyms are not

explained separately in distinct lemmas: the synonym only has a clear reference to the lemma of the main term, where all necessary information is given, including all equivalents in the other language. The authors and sponsor deemed it necessary to target the dictionary mainly at experts and semi-experts participating in all phases and aspects of the BBT project. The function of the BBT dictionary is therefore to support both groups of users in understanding and producing texts in the foreign language. Experts "will have no reception problems within [their] own field. [They] may have to acquire new knowledge, but [they are] not likely to find this in any lexicographical dictionary" (Bergenholtz & Kaufmann 1997: 102). Thus, they will predominantly use the dictionary for text production and retrieving foreign language terms. Semi-experts are potential dictionary users who come from other related subject fields that are relevant but not strictly specific to railway tunnel projects, e.g. engineers, surveyors, geologists, installers, etc. These professionals work in the public or private sector and daily interact with the world of railway construction and the BBT (cf. Bergenholtz & Kaufmann 1997: 101). They will need to find all synonyms and variants of a term as well as definitions in order to get a better picture of each concept.

The BBT dictionary contains about 2000 terms in German and Italian from diverse domains, ranging from railway construction to tunnel building, economy, energy, geology, telecommunications, transport/mobility, social issues and environmental terminology, thus encompassing – next to strictly technical terminology – also organisational, administrative and legal terminology (cf. Maino et al. 2011: 14).

3 Experts and terminologists: standard forms of cooperation

Keine Terminologie ohne Fachleute – keine Fachleute ohne Terminologie[1] (RaDT 2013: 9)

The quotation above expresses the essence of terminology. Terminologists and experts must work side by side and create synergies to achieve a reliable and high quality terminological product. In principle, terminologists retrieve, reference and record the terms that pertain to a specialised domain in one or more languages (RaDT 2004: 2; Chiocchetti et al. 2013: 41-42), while domain experts are normally consulted for explanations and information (RaDT 2013: 7-8). Often they also revise the final product of the terminologists.

Usually terminologists start by studying and delimiting the domain to be processed terminologically, which is often subdivided into smaller subdomains. To this purpose, terminologists acquire and skim relevant and up-to-date reference material, which may also be collected or selected as indicated by domain experts. The material is collected in all languages that are to be included in the terminologi-

1 "No terminology without domain experts – no domain experts without terminology" (translation by the authors).

cal product, making sure that it is original material written by expert native speakers, so that it reflects the actual language used by the community of experts of a given domain.

This material serves as a basis for understanding the key concepts of the domain under analysis, which are organised in concept systems to illustrate the relation between each concept within the specific domain or sub-domain. The material is also used to retrieve all the terms that designate the concepts of a specialised domain. "A term is a designation consisting of one or more words representing a general concept in a special language in a specific subject field" (ISO 2009:704: 7.2.1). There might be more than one designation for the same concept, i.e. there might be synonyms (ISO 2009:704: 7.2.4). Also term variants, e.g. abbreviated forms like clippings and acronyms, are common in specialised domains (cf. ISO 2009:704: B.2.4).

Based on this preparatory research and analysis, terminologists then compile (or update) fully fledged terminological entries with relevant information, i.e. with definitions, contexts of use, term variants/synonyms, usage notes etc. The most important element of a terminological entry is the definition, as it conveys the meaning of a concept within the specialised domain to which it belongs. All the other pieces of information contribute to explaining how the terms and variants that designate a concept are employed within a specialised domain.

Finally, with the help of all information gathered in the source language and the structure of the relevant concept-system, terminologists retrieve the equivalents in the target language. Target language terms are then processed terminologically in the same way as the terms in the source language are. If no equivalent exists, terminologists may propose new terms (translation proposals) to fill the terminological gaps (see section 5.5).

Written and human sources (i.e. domain experts) may be consulted by terminologists for information and explanations at any time during the entire process. Domain experts are preferably involved in the planning and realisation of terminology projects from the very beginning. Usually their role consists in supporting terminologists (RaDT 2013: 7-8). At the beginning, they can help to plan and organise terminology projects, especially by providing information on relevant reference material and selecting it. Their initial input is also important when delimiting the domains and subdomains to be processed terminologically.

As terminology work proceeds, domain experts may be asked to select the terms that were extracted from the written material by the terminologists, i.e. to choose which terms shall become lemmas of a dictionary or terminological entries in a database. Experts can also check concept systems and verify the correctness of the terms, synonyms and associated definitions. Domain experts usually provide competent advice in case of any doubts. Finally, being part of a scientific community and/or of a practical community, they represent the ideal channel for disseminating the results of the terminology work to their peers.

As we have seen, domain experts normally act as consultants and/or revisers. They are of paramount importance for the success of terminology work and represent a precious source of information, since they assess the quality of source documentation, explain the meaning of concepts belonging to their domain of expertise and/or check whether the terminology collection is correct and complete. However, what they usually do not do is compile terminological entries themselves.

4 Experts and terminologists: exchanging roles for the BBT dictionary

For the creation of the BBT dictionary domain experts and terminologists exchanged roles. This time the terminologists acted as consultants and revisers, while the experts took over most of the work concerning the compilation of the dictionary entries.

The advisory role of terminologists started with helping the experts select the most up-to-date and authoritative sources. As the experts are able to judge personally which pieces of information are correct and precise, they tend to disregard their provenance and to treat all types of reference material in the same way, from highly specialised technical manuals to commercial web pages. By applying strategies for source evaluation and selection learnt from the terminologists, the authors managed to produce a more homogeneous and complete dictionary. Current scientific and technical sources were consistently preferred to general language dictionaries or encyclopaedias. In any case, information was always double-checked and carefully evaluated.

Terminologists provided support also in the initial term selection phase. In fact, they strived to agree with the experts on a coherent set of related terms. For the dictionary this sometimes meant discarding a very specialised – albeit possibly useful – lemma and including a maybe less tricky lemma instead, in order to ensure that most concepts in the concept field of a specific subdomain were represented. This compromise in term selection likewise allowed reaching a higher level of homogeneity in the dictionary.

Great efforts were made by the terminologists in revising definitions. They convinced the experts that the typical structure of terminological definitions could represent a useful definition strategy, especially for semi-experts. Terminological definitions start with stating the superordinate concept and then list all the characteristics that distinguish the concept under analysis from its related concepts nearby (see section 5.6). In this way they provide essential information in a very compact form and help users to quickly understand the position of a concept within its specialised domain. Achieving a more systematic and coherent structure for all definitions also allowed the dictionary to be turned into a more consistent product.

Terminologists took over several other methodological aspects, e.g. by supervising the rigorous treatment of synonyms and abbreviated forms and their cross references to the main entry. They performed the terminological revision, i.e. they checked the consistency of terms and definitions. Most important, they checked whether the terms and definitions in Italian and German actually all referred to the same concept, thus making up for the lack of language competences of the domain experts. In some cases they actually retrieved the equivalents in the target language or advanced translation proposals. Finally, terminologists performed the linguistic revision in both Italian and German.

This swapping of roles was essentially born out of necessity, as the domain experts lacked the necessary competences and methodological basis for dictionary-making as well as (partly) lacking the linguistic competences in both working languages, while the terminologists could not take over their

standard role due to lack of time and financial means. However, despite the challenges faced, the co-operation in this "reversed form" proved stimulating and fruitful for both parts.

5 Challenges faced and compromises reached

Before and during the compilation phase of the dictionary the domain experts were trained by the terminologists to follow the basic principles of terminology and of dictionary-making. In some areas it was particularly difficult to create a common knowledge base and achieve consensus on a methodology, so that both parts had to agree on compromises, as explained in detail in the following paragraphs.

5.1 Hierarchy of sources

Many different types of sources can be used for terminology work (cf. ISO 10241-1:2011: 4.3.5.2). Their respective relevance will vary according to the aim, type and content of terminology work, the domain(s) treated, the languages considered, the end users, etc. (Chiocchetti et al. 2013: 18). Usually terminologists give preference to authoritative sources like legal documents, standards, documents generally recognised by the scientific community (e.g. textbooks) (ISO 10241-1:2011: 4.3.5.2). Often the reference material is classified hierarchically, with the most official and authoritative sources at the top (e.g. laws, standards, etc.) and the less authoritative ones at the bottom (e.g. private webpages, commercial material, etc.). Information retrieved in sources classified at a higher level of the hierarchy will be preferred to information found in documents filed at a lower level. Contrary to this terminological practice, the experts working on the BBT dictionary often used material from general encyclopaedias and from popular websites (e.g. Wikipedia), which are easily retrieved on the Internet but cannot always be considered reliable or specific enough. This is the reason why they are generally avoided by terminologists or at most used for information retrieval, but seldom quoted.
For the BBT dictionary, since the expertise of the authors allowed them to assess the correctness and quality of definitions found in "unconventional" reference material, many definitions were accepted and cited within the dictionary whenever no other source of information was available. This compromise allowed a faster compilation of parts of the dictionary without any notable loss of quality.

5.2 Subdivision into glossaries

According to standard terminological practice, when treating large or very diverse sets of data (see section 2), work is subdivided into thematic glossaries to facilitate compilation and revision. This practice was new to the authors who nevertheless quickly grew accustomed to the method. In the final version of the dictionary all terms could still be listed in alphabetical order, thus ensuring imme-

diate retrievability of each term as well as of all variants or synonyms by the end-users (Maino et al. 2011: 14).

5.3 Term selection

Due to restrictions in space, it was not possible to compile complete concept systems for all glossaries and accommodate all relevant terms in the dictionary. Term selection was therefore guided by the relative relevance for the specific BBT project and not strictly by conceptual diagrams of each domain, as terminologists normally strive to do with the aim of treating all domains equally.

Catering to two different groups of users likewise required a series of compromises in the selection of terms and variants/synonyms to be considered. Experts tend to use highly specialised vocabulary (e.g. also consisting of acronyms, initialisms and formulas) that is monosemic and unambiguous to avoid problems in interpretation (cf. Sobrero 1993; Cortelazzo 1994). Semi-experts, however, call for what is defined as a "variationist approach" in terminology, i.e. an approach where all synonyms and variants used to designate a concept are considered, whether they be "full forms, such as simple, compound or complex terms [and] [...] all their variations" (Bertaccini & Lecci 2009), or abbreviated forms (see Fig. 1). To this aim, the dictionary builds a network of references from all synonyms and variants (e.g. acronyms) to the main entry containing the definition of the concept (see Fig. 2 and 3).

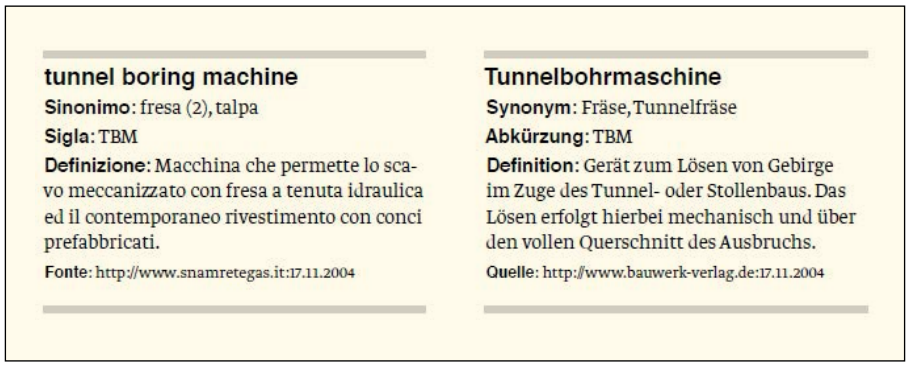


Fig. 1: Entry tunnel boring machine/Tunnelbohrmaschine with synonyms (= sinonimo/Synonym) and initialisms (= sigla/Abkürzung).

5.4 Managing synonyms and variants

The consistent treatment of synonyms and also variants —with the respective references from the synonym/variant to the main defined terms — was taken over and managed by the terminologists. Experts tend to disregard the importance of terminological variation, because they have all the synonyms and variants in mind. But for a semi-expert dictionary user it might, for example, not be so easy to read the full form behind an acronym used in a text. By listing all designations that refer to the

same concept in separate lemmas as well as in alphabetical order in the BBT dictionary, with consistent cross-references to their respective full forms and/or main terms (see Fig. 2 and 3), maximum retrievability of information for all end-users could be ensured.

Fig. 2: Entry *fresa (2)*, cross-reference to the main Italian term *tunnel boring machine*.

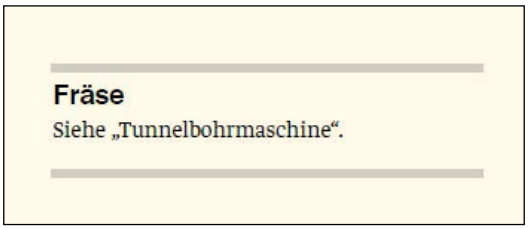


Fig. 3: Entry *Fräse*, cross-reference to the main German term *Tunnelbohrmaschine*.

5.5 Diverging treatment of legal/administrative terminology from purely technical terminology

The nature and origin of the terminology used within the BBT project is quite diverse (see section 2). Legal terminology, which is for example contained in building regulations, poses a particular challenge in such a context, since it is much more difficult to find conceptual equivalence in the legal/administrative than in the technical domain. Legal and administrative terminology is always strongly connected to a specific legal system with its own cognitive and conceptual structures, as well as its written or oral sources (cf. Gambaro & Sacco 1996: 9; Sandrini 1996: 138; Šarčević 1997: 232). For this reason, it might not always be possible to find an equivalent in the target language; the concept designated by the term might be specific to the source legal system and source language and be completely unknown in the target legal system and language. This situation creates a terminological gap.
The presence of a terminological gap causes the need to look for strategies of translation other than equivalence (e.g. paraphrase, neologism, etc.). In the BBT dictionary, terminological gaps in the legal/

administrative terminology were faced by offering translation proposals that conveyed the meaning of the concept in the source legal system and source language into the target legal system and language (see Fig. 4). Due to the intrinsic connection of legal terminology with its legal system, it was not possible to borrow legal terms from other legal systems.

Fig. 4: Translation proposal in German for the Italian administrative term *piano comunale di classificazione acustica.*

Another approach, however, was followed for gaps in the technical terminology. In this case, problems mainly concerned a different level of evolution of the tunnel building techniques in Italy and in Austria, with the ensuing absence of some terms that designate very specific concepts. As the technical terminology used in Switzerland in both Italian and German is very complete and up-to-date, several Swiss sources helped to fill the presumed gaps (see Fig. 5). For similar reasons, some sources of information from Federal Germany could be referred to for the German language terminology.

The terminologists had to explain and discuss this diverging treatment with the dictionary authors. Since none of them was a legal expert, the implication of legal comparison across national borders was not immediately clear and had to be motivated and explained.

Fig. 5: Entry *Bohrkopf* with definition from a Swiss website about tunnel building.

5.6 Definitions

Experts have the tendency to provide longer encyclopaedic definitions rather than much more compact classical terminological definitions. They tend to give more information that the amount strictly necessary for a mere definition of the concept, tending to add explanations on how a defined concept is employed within their domain of expertise (see Fig. 6).

The traditional, most explicit and precise definition in terminology is the intensional definition, stating "the superordinate concept immediately above [the concept that is being defined], followed by the delimiting characteristic(s)" (ISO 704:2009: 6.2) that distinguish it from coordinate or from other related concepts. This type of definition allows full and systematic identification of a concept with respect to all others in the specific domain (Sager 1990: 42). Yet it is very brief and not always sufficient for laypersons or semi-experts to really understand the meaning and usage of the concept defined.

In this case the terminologists gave way to the desires by the dictionary authors for a more in depth explanation of some key concepts. In the dictionary, which has a very compact structure without any notes or comments, definitions were allowed a more flexible structure, sometimes leaving room for the inclusion of necessary additional information and clarifications.

Schwermetall

Definition: Metall mit einer Dichte von über 5 g/cm³, z. B. Zink, Kupfer, Zinn, Chrom, Cadmium, Blei, Quecksilber u. a. Eine Reihe von Schwermetallen sind in Spuren für biologische Vorgänge lebensnotwendig. Hohe Konzentrationen der meisten Schwermetalle sind dagegen äußerst giftig und können über den Abfall, Verbrennungsgase und Abwasser zu erheblichen Umweltproblemen führen, wenn sie über Boden und Pflanzen in die Nahrungskette gelangen.

Quelle: Köppen D., 1998. Definitionen für agrarökologisch relevante Sachverhalte. Fachbereich Agrarökologie, Rostock.

metallo pesante

Definizione: Metallo con densità maggiore di 5 (zinco, cadmio, mercurio, stagno, piombo, cromo, manganese, ferro, cobalto, nichel e rame), nonché i relativi composti che, pur essendo naturali componenti della crosta terrestre, tramite numerose attività umane vengono mobilitati e concentrati a livelli pericolosi per la salute e l'ambiente.

Fonte: Gamba G., Martignetti G., 1995. Dizionario dell' ambiente. Formulazione di responsabilità. ISEDI, Torino.

Fig. 6: Definition of *Schwermetall* with additional information.

6 Conclusions

This role-switching exercise proved very fruitful for both sides. The experts became familiar with the basic principles of terminology work and dictionary-making; particularly the evaluation of source material and the definition-writing skills were considered useful to them beyond the BBT dictionary itself. The terminologists learnt how to find pragmatic solutions to practical problems, as well as how

to reach compromises between standard terminological practice and the limitations of a printed reference work that is aimed primarily at experts and semi-experts, rather than at translators and other language professionals.

The result of this exchange of roles is a dictionary based on many more compromises than other terminological projects. Terminologists usually take over most of the work and the role of the experts is limited to sporadic interventions as advisers and proofreaders. As a consequence, terminologists often have the last word on dictionary structure and content, even though the opinion of the experts is always greatly considered and systematically taken into account. However, for the BBT dictionary, the advice of the terminologists during dictionary compilation was generally limited to more formal aspects, such as revising the structure and wording of the definitions, and to methodological issues, e.g. the consistent treatment of synonyms and short forms.

Great compromises were reached, for example, concerning definitions. As we have seen, in the BBT dictionary definitions still follow the classical terminological structure whenever possible, but their content goes beyond the mere identification of superordinate concept and delimiting characteristics. Definitions thus often include additional (technical) information that the authors considered necessary and useful for either peers or semi-experts. Another compromise was reached for the treatment of linguistic information that is normally given by terminologists in their work, i.e. grammatical information, example sentences, notes distinguishing the contexts of use of different synonyms and term variants, language register, collocations, etc. This type of information is basically absent from the dictionary, due to the fact that the target groups of the dictionary do not primarily include translators and language professionals. While translators might wish for specific linguistic information, this is often unnecessary for experts, so more space was devoted to treating a larger number of lemmas.

The BBT dictionary project has finally proven that, given the different backgrounds and approaches of terminologists and domain experts, it is advisable to provide detailed guidelines on how to handle specific aspects (e.g. the structure of definitions) in order to work along common guidelines and principles from the very beginning (Chiocchetti et al. 2013: 46). It has also shown that it is indeed possible to envisage different forms of cooperation between domain experts and language experts, still ensuring a high level of quality of the final product.

7 References

Bergenholtz, H. & U. Kaufmann (1997). Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. In *Hermes, Journal of Linguistics*, 18/1997, pp. 91-125.

Bertaccini, F. & C. Lecci (2009). Conoscenze e competenze nell'attività terminologica e terminografica. In "Terminologia, ricerca e formazione", *Publifarum*, 9. Accessed at: http://www.publifarum.farum.it/ezine_articles.php?art_id=107 [10/11/2013].

Chiocchetti, E., Heinisch-Obermoser, B., Löckinger, G., Lušicky, V., Ralli, N., Stanizzi, I. & T. Wissik (2013). *Guidelines for Collaborative Legal/Administrative Terminology Work*. Bolzano: EURAC. Accessed at: http://www.eurac.edu/en/research/institutes/multilingualism/Documents/Guidelines_for_collaborative_legal_administrative_terminology_work.pdf [10/11/2013].

Cortelazzo, M.A. (1994). *Lingue speciali. La dimensione verticale*. Unipress: Padova.

Gambaro, A. & R. Sacco (1996). *Sistemi giuridici comparati*. Torino: UTET.

ISO 704:2009. Terminology Work – Principles and Methods.

ISO 10241-1:2011. Terminological Entries in Standards – Part 1: General Requirements and Examples of Presentation.

Maino, F., Cavallaro, F. & M. Wagner (eds.) (2011). Dizionario Terminologico della Galleria di Base del Brennero Italiano-Tedesco/Terminologisches Wörterbuch zum Brenner Basistunnel Deutsch-Italienisch. Bolzano: EURAC/BBT.

RaDT – Rat für Deutschsprachige Terminologie (2004). *Berufsprofil Terminologin/Terminologe*. RaDT: Bern. Accessed at: http://www.iim.fh-koeln.de/radt/Dokumente/RaDT_Berufsprofil.pdf [10/11/2013].

RaDT – Rat für Deutschsprachige Terminologie (2013). *Terminologisches Basiswissen für Fachleute*. RaDT: Köln. Accessed at: http://www.iim.fh-koeln.de/radt/Basiswisssen -RaDT2013-16s_ebook.pdf [10/11/2013].

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Sandrini, P. (1996). Terminologiearbeit im Recht. Deskriptiver begriffsorientierter Ansatz vom Standpunkt des Übersetzers. Vienna: TermNet (IITF-Series no. 8).

Šarčević, S. (1997). *New Approach to Legal Translation*. The Hague: Kluwer Law International.

Sobrero, A. (1993). Lingue speciali. In Sobrero, A. (ed.): *Introduzione all'italiano contemporaneo. La variazione e gli usi*, vol. 2. Bari: Laterza, pp. 237-277.

Cloud Terminology Services Facilitate Specialised Lexicography Work

Tatiana Gornostay, Andrejs Vasiļjevs
Tilde
tatiana.gornostay@tilde.lv, andrejs@tilde.lv

Abstract

In this software demonstration paper we present an innovative cloud-based platform TaaS "Terminology as a Service" developed in an EU-funded project.[1] The TaaS platform provides language workers and language applications (human and machine users, accordingly) with the services to foster the creation, validation, harmonisation, sharing, and application of terminology resources. Under language workers we understand language professionals, for example, technical writers, editors and proof-readers, translators and localisers, terminologists and domain specialists, lexicographers and terminographers, and others. Under language applications (or machine users in other words), in the first place we consider computer-assisted translation (CAT) tools and machine translation (MT) systems (also, knowledge organisation systems in library and information science, search engines, and others). TaaS provides the following terminology services: terminology search in various sources, terminology identification in and extraction from user-uploaded documents, terminology visualisation in user-uploaded documents, translation equivalent lookup in and retrieval from various sources, terminology refinement and approval by users, terminology sharing with other users, collaborative working environment, and terminology reuse in other applications. Among other benefits for language workers, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based specialised terminology work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries.
Keywords: terminology service; terminology work; specialised lexicography

1 Introduction

Lexicography, as the theory and practice of dictionary development, is one of the most labour-intensive human activities. The creation of a new dictionary from the scratch and its delivery to an end user requires many resources in terms of time, labour, and finance. The main drawback of a conventional *paper* dictionary is its static and out-of-date content. In specialised lexicography, it is even more critical since terminology is developing rapidly along with its subject field and science in general.

1 The TaaS Beta was officially launched on November 1, 2013 and is publicly available for open Beta testing at https://demo.taas-project.eu.

To overcome the shortcomings of conventional lexicography, an electronic punch-card machine was first used to create a prototype of a modern electronic dictionary by Roberto Busa in the 20th century. His first work was based on the automatic linguistic analysis of the works of Saint Thomas Aquinas (he lemmatised the texts). Roberto Busa compared the invention of an "electronic book" (instead of a printing book) to the introduction of a printing book by Gutenberg (instead of a manuscript). Since that time automated lexicography has been developing boomingly.

Nowadays, with the evolution of information technologies, the Web, and data (for example, open data, linked data, free language resources etc.), the task of automated specialised lexicographic work is put in the first place. Routine processes have been delegated to a computer. An electronic, or computer-based, specialised dictionary is easy to update and manage, and its main advantage is its flexible, dynamic, and extensible (for example, in terms of new languages) character. Moreover, the new era of information technologies offers the new ways of dictionary representation, for example, on a tablet, mobile, and other devices, and the usage patterns of a dictionary (including a specialised dictionary) are changing with the course of time.

The integration of natural language processing tools within a lexicographer's working environment have made it possible for him/her to linguistically and semantically analyse and tag data and then to extract required pieces of information from it. In specialised lexicography it is possible to identify and extract term candidates automatically for further processing (for example, refinement, approval, sharing and reuse). Thus a lexicographer can consider hundreds thousands of terms in a certain subject field in comparison with that time when only several thousands (usually no more than 2000) could be included in a conventional specialised paper dictionary. This opportunity is critical particularly in emerging domains.

2 TaaS: Terminology as a Service

In this software demonstration paper we present an innovative platform TaaS "Terminology as a Service". The platform provides language workers and language applications (human and machine users, accordingly) with the services to foster the creation, validation, harmonisation, sharing, and application of terminological data. Among others, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-oriented, collaborative, multilingual, interoperable, portable, and cloud-based work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries.

TaaS is being developed within an industry-research collaborative project under the EU Seventh Framework Programme for Research and Technological Development. The main objective of TaaS is to address the need for instant access to the most recent terms and for direct user involvement in the creation, harmonisation, and sharing of terminological data. The Beta version of TaaS was officially launched on November 1, 2013 and is publicly available for open Beta testing.

The concept of innovative cloud terminology services for language workers and language applications is presented in Figure 1 below.

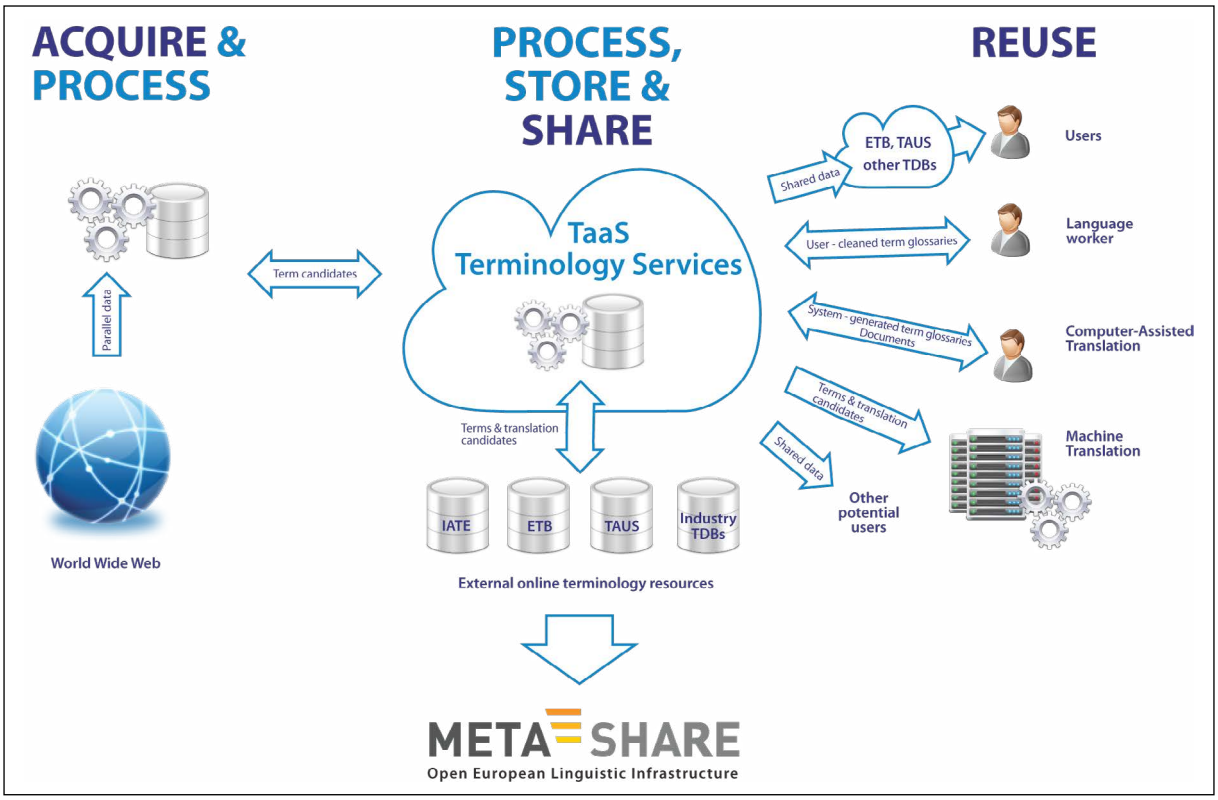


Figure 1: TaaS innovative cloud terminology services for language workers and language applications.

TaaS provides user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based terminology services to perform the following tasks:

- Search terminology in various sources;
- Identify term candidates in user-uploaded documents and extract them automatically applying linguistic and statistical processing;
- Visualise term candidates in user-uploaded documents;
- Look up translation equivalent candidates in various sources (for example, existing external terminology resources EuroTermBank[2], IATE[3], TAUS Data[4], and possible other resources, as well as automatically extracted bilingual terminological data stored in the TaaS Shared Term Repository and used as an additional internal source for target translation lookup);
- Refine term candidates and their translation equivalent candidates;
- Approve refined terminology;

2 www.eurotermbank.com
3 http://iate.europa.eu
4 www.tausdata.org

- Share terminology with other users;
- Collaborate with colleagues in user-friendly working environment;
- Use terminology in other applications via TermBase eXchange ISO-standardised format (TBX), tab-separated value (TSV), and comma-separated value (CSV) export formats and via the TaaS Application Program Interface (API).[5]

To perform most of his/her work in TaaS, the user has to sign up for the services. However, for its non-signed users, TaaS provides the service for terminology search in two sources – TaaS database, which consists of TaaS users' terminology collections made public by its users, and EuroTermBank, which is the largest European online term bank, providing access to more than 2 million standardised terms from more than 100 national terminology resources in 27 languages. For advanced search, the user has to select the source and target language, domain (a.k.a. subject field), and the source to be searched in (see Figure 2).

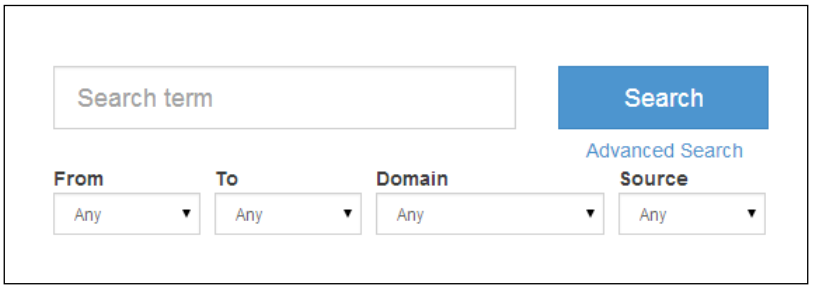


Figure 2: Search form in TaaS.

For signed users the work in TaaS is organised in projects. A signed user gains access to full TaaS services. To start his/her work, the user has to create a new project indicating the source and target language and the domain the user works in (it is relevant to user document(s) domain). More than 10 input format for user documents are supported. The user might also want to specify optional properties, such as product, customer, project description, and the business unit (in case of a corporate user) (see Figure 3).

5 For TaaS API contact TaaS team via e-mail langserv@tilde.com.

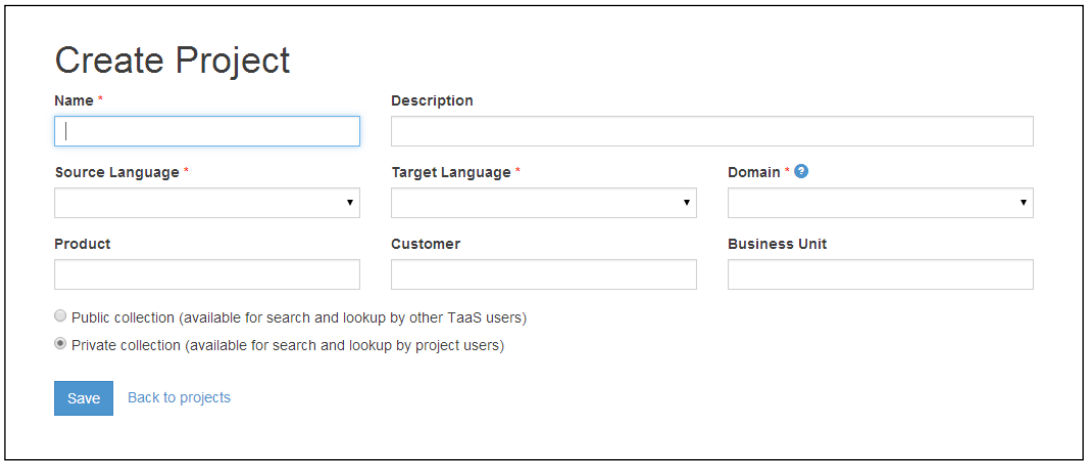


Figure 3: Creation of a new project in TaaS.

TaaS also provides a default project with project properties already set for demonstration purposes. Finally, the user has to set the status of his/her project – private or public. If the status of a project is public, the user's approved terminology will be available for search and lookup by other TaaS users; if the status of a project is private, the user's approved terminology will be available only to project users.

The user can start his/her work with TaaS by using the default project or creating a new project. In both cases, the user is an administrator of his/her project.

TaaS provides facilities for project sharing among users if they work in a team. This functionality that typically involves an interchange of non-confidential, non-competing, and non-differentiating terminology across various actors is highly rated by users. Recent surveys have shown that up to 60% of terminology resource users would share their resources with the community. The concept of sharing, unfortunately, is not present in the current management of major terminology databases and term banks. Instead of providing the opportunity for users to contribute their data, major term banks typically keep to the traditional one-way communication of their high-quality pre-selected terminological data.

To share his/her project with other users, the user has to add their e-mails and assign their roles. There are three available roles to a new user of the shared project: administrator, with full access rights; editor, with limited access to editing rights; and reader, with limited access to reading rights. One project can have more than one administrator; however, the owner of the project (the user, who has created the project) must consider assigning the administrator's role to other users of his/her project as these users will get full access, including the right to delete the project and its terminology collection. The Administrator's role is usually assigned to the project manager in the translation team, who adds documents to the project, and these are later processed by a terminologist, translator(s), editor(s), and other translation team members (see Figure 4).

The main usage scenario for the TaaS services is when the user uploads his/her document(s) under the created project, in order to then execute the terminology processing. TaaS supports user document upload in more than 10 formats including the most widely used MS Word, Excel, and Power Point formats as well as the Portable Document Format (PDF), the XML Localisation Interchange File

Format (XLIFF), and others. The open Beta version has certain limitations in terms of file and project size.

The terminology extraction service performs automatic extraction of monolingual term candidates from user- uploaded documents using generic or language specific terminology extraction techniques.

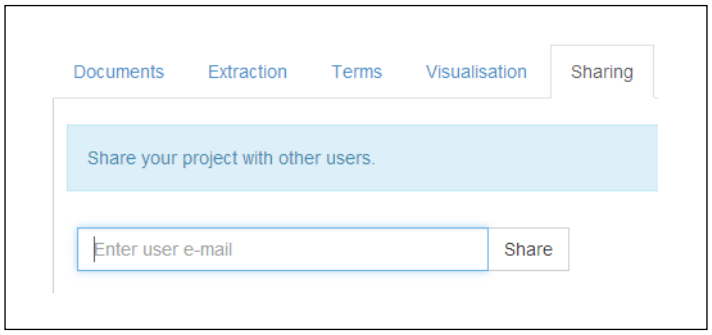


Figure 4: Project sharing in TaaS.

The user can customise the terminology extraction process. He/she can select one or more available (on the platform) terminology extraction tools for term candidate identification in user-uploaded documents. There are two term identification tools integrated into TaaS at the moment. These are the Tilde Wrapper System for CollTerm (TWSC)[6] that includes language specific patterns and morphological analysis and Kilgray Term Extractor that applies generic statistical approach to all supported languages. It is recommended to select the first tool; however, the statistical tools might also be of help in certain cases, for example, when linguistic processing produces insufficient results.

The platform provides the service for automatic retrieval of translation equivalents (for the extracted monolingual term candidates) in user-defined target language from different public and industry terminology databases.

The following terminology resources are available for translation equivalent lookup for term candidates identified in user-uploaded documents:

- TaaS public collections shared by other TaaS users;
- Terminology collections owned by the user;
- EuroTermBank;
- Inter-Active Terminology for Europe (IATE), an inter-institutional terminology database of the European Union[7];
- TAUS Data that stores shared translation memories;

6 See the ACCURAT Toolkit 3.0 at www.accurat-project.eu.
7 http://iate.europa.eu/

TaaS database of raw bilingual terminological data automatically extracted from original and translated texts (a.k.a. comparable and parallel corpora) on the Web.

In the dynamic pace of technological developments and societal changes, new terms are coined every day by industry, translation and/or localisation agencies, collective and individual authors. Although these terms can be found in different online and offline publications, the inclusion of new terms in online public terminology databases and term banks takes months or even years, if it happens at all. As a result, terminology databases and term banks fail to provide users with extensive up-to-date multilingual terminology, especially for terms in under-resourced languages or specific domains that are poorly represented in online public terminology resources.

At the same time many new terms and their translations can be found on the Web – in multilingual websites, online documents, support pages, etc. TaaS provides four bilingual terminology extraction workflows for Web data: one workflow for terminology extraction from parallel data and three workflows from comparable data. The latter three are customised to collect terms from comparable news corpora, from multilingual Wikipedia, and from focused comparable corpora, respectively.

Web data are collected and then automatically processed. As a result, a list of bilingual raw term candidate pairs are extracted and fed into the TaaS terminology repository. During the execution of a terminology project at the translation candidate lookup step, these data are retrieved and proposed to the user for his/her validation. Thus the TaaS aligns the speed of terminology resource acquisition with the speed at which the content is created by mining new terms directly from the Web.

The data collection process is ongoing constantly feeding the TaaS repository with new terms. By April 2014, the TaaS database included more than 8 M bilingual term pairs extracted from the Web data.

TaaS provides facilities for cleaning up raw terminological data extracted automatically that is noisy and needs validation by users. The process of validation can be regarded as a three-step procedure:

- monolingual validation (deletion of “unwanted” and/or unreliable term candidates, definition of termhood, term variant identification, deduplication, deletion of “incorrect” extraction, for example, a part of a longer noun group, synonym identification etc.);
- bilingual validation (bilingual checking of term candidates and their translation candidates, defining the right translation for the source term, deletion of irrelevant and/or incorrect translations, etc.);
- validation in context.

As soon as extraction finishes, the user can see extracted terms from his/her documents and their translation equivalents retrieved by TaaS. The user can hover over terms to get additional information, such as grammar, source, and context (see Figure 5).

Figure 5: Clean-up and validation of raw terminological data in TaaS.

The user can approve terms with a single click and add translations him-/herself, is the right translation from proposed translation candidates is not found.

An extracted term with its translation equivalent(s) forms a terminology entry. For advanced purposes, the user might want to edit a term entry in full entry view using the term entry editor and to add additional information about terms, for example, definitions, notes, grammatical information, and usage properties, such as term type, register, administrative status, temporal qualifier, geographical usage, and frequency. The history of editing is saved and is seen in the full entry view. The user might also want to see term candidates identified by TaaS in his/her documents, and the visualisation functionality is available for this purpose (see Figure 6).

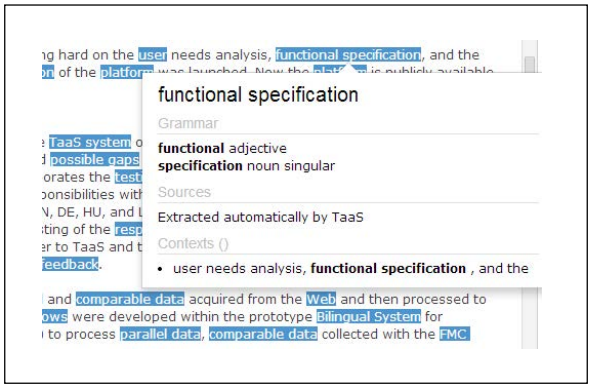


Figure 6: Visualisation of term candidates in the user's document in TaaS.

Validated terminological data can be exported and then reused in other working environments.

During the analysis of user needs and requirements, we also proved our hypothesis that terminology, as a language resource, is central for the second large group of users – language applications (the first user group is represented by language workers). We have already performed first successful experiments on the integration of terminological data acquired within TaaS into the statistical MT system. At the time, the memoQ CAT tool[8] owned and developed by Kilgray, the TaaS project partner, is being integrated with TaaS via the TaaS API developed in the project and available for machine users.

8 See the description at http://kilgray.com/products/memoq.

TaaS demonstrates the efficacy of its terminology services within the following practical usage scenarios:

- For language workers: simplification of processing, storage, sharing, and application of task-specific multilingual terminology.
- For computer-assisted translation (CAT) tools: instant access to term candidates and translation equivalent candidates via the TaaS API.
- For statistical machine translation (SMT) systems: support for domain adaptation by a dynamic integration with TaaS-provided terminological data via the TaaS API.

At EURALEX the usage scenario for language workers with the emphasis on specialised lexicography work will be demonstrated online.[9]

3 Conclusion

In this software demonstration paper we have presented an innovative cloud-based platform TaaS "Terminology as a Service" developed in an EU-funded project. At the present time, TaaS is a unique dynamic cloud-based solution that provides a wide range of terminology services. We foresee the potential of the established platform for a wide range of user groups, both language workers and language applications. Among other benefits for language workers, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based specialised terminology work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries This opportunity is critical particularly in emerging domains. At the EURALEX Congress the platform is demonstrated in real-time during the three days of the event.

Acknowledgements

Research within the TaaS project, leading to these results, has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 296312.

The TaaS platform is a result of fruitful collaborative work of the project partners – coordinator and lead developer Tilde (Latvia), research partners Cologne University of Applied Sciences (Germany) and University of Sheffield (UK), industry partners Kilgray (Hungary) and TAUS (Netherlands).

9 Live demonstration requires Internet access.

Good Contexts for Translators–A First Account of the Cristal Project

Amélie Josselin-Leray*, Cécile Fabre*, Josette Rebeyrolle*, Aurélie Picton**, Emmanuel Planas***
*CLLE-ERSS, University of Toulouse & CNRS, France; **FTI, University of Geneva, Switzerland;
***LINA, University of Nantes, France
josselin@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr, rebeyrol@univ-tlse2.fr,
aurelie.picton@unige.ch, emmanuel.planas@univ-nantes.fr

Abstract

This paper questions the notion of good contexts for translators and describes an experiment which tests the usefulness of two specific kinds of contexts in a translation task, namely (1) contexts that provide conceptual information about a term, and (2) contexts that provide linguistic information about the collocational profile of this term. In the experiment, trainee translators are asked to use several types of resources, including a set of pre-annotated contexts of various types, and to identify the contexts that they consider to be the most relevant for their task. We present the first results of this experiment, which confirm our general assumption about the usefulness of such rich contexts and indicate some differences regarding the use of contexts in the source and target language. This study takes place in the CRISTAL project whose aim is to retrieve from bilingual comparable corpora the contexts that are the most relevant for translation and to provide them to users through a CAT tool.
Keywords: CAT tools; corpus resources for translators; Knowledge-Rich Contexts

1 Introduction

Even though it is widely acknowledged as being essential to the translator, the very idea of context in translation is hard to define (Baker 2006: 321) and it also "lacks a definition that can be applied in the everyday work of a professional translator" as stated by Melby & Foster (2010: 1). Therefore, when one wants to provide translators with tools that better meet their needs–such as improved CAT tools–, one should in the first place wonder about what makes a context relevant for them. In other words, what is a 'good context' for translators? This is one of the questions the CRISTAL project[1] tries to give an answer to. The main aim of the CRISTAL project, an acronym that stands for "Knowledge-Rich Contexts for Terminological Translation" ("Contextes Riches en Connaissances pour la Traduction Termi-

1 CRISTAL is a three-year project (2012-2015) funded by the French National Agency for Research (ANR-12-CORD-0020). It involves four partners: a computing research team at the University of Nantes, France (LINA), a linguistics research team at the University of Toulouse, France (CLLE-ERSS), the Translation Technologies team from the Faculty of Interpreting and Translation at the University of Geneva in Switzerland, and a firm specializing in multilingual text management (Lingua et Machina).

nologique" in French), is to automatically retrieve from bilingual comparable corpora the contexts that are the most relevant for translation and to provide them to users through the CAT tool developed by Lingua & Machina, the Libellex Platform.

The first part of this paper reviews the translators' needs regarding context. It seems necessary to first identify which type of information is essential for translators, to see how this information is recorded in the tools most commonly used by translators, i.e. dictionaries, term banks and corpora, and how satisfied translators are about the way that information is recorded. In order to refine the notion of "good contexts" for translators, in part 3 we investigate what a "good example" in lexicography and what a "Knowledge-Rich Context" in terminology are, and introduce the distinction between conceptually-rich and linguistically-rich concepts. Part 4 then focuses on one aspect of the methodology of the CRISTAL project: an experimentation involving trainee translators in order to refine our idea of a "good context for translators". Finally, part 4 presents the very first results of the experiment.

2 Some Facts about the Needs of Translators Regarding Context

As stated by Rogers & Ahmad (1998: 195), "one of the translator's prime needs is for context-sensitive information". We may wonder what the notion of context-sensitive information encompasses and what sources of information translators can rely on–or not.

2.1 What do Translators need Contextual Information for and Where do they Find it?

2.1.1 Context in Translation: a Preliminary Definition

As thoroughly explained by Melby & Foster (2006), specialists in many fields (e.g. philosophy, psychology, pragmatics, and functional linguistic) have discussed the notion of context, and various definitions have been written. The three facets of context as defined by Halliday (1999), i.e. *context of situation*, *context of culture* and *co-text* are all particularly relevant in translation. However, in this paper, we will only focus on what Halliday calls *co-text*. While both *context of situation* and *context of culture* are outside of language itself, *co-text* specifically pertains to language in use. It can broadly be defined as the surrounding discourse of an utterance. Therefore, our definition of "context" in this paper will be limited to *co-text*, and will rely on the definition provided by Fuchs[2]:

> What is called *context* is the linguistic environment of an element (phonetic unit, word or group of words) within an utterances; i.e. the units that precede and follow it. Thus, in the utterance "Marie est jolie comme un cœur", the element *comme* has as its immediate context "jolie…un cœur" and its

2 http://www.universalis.fr/encyclopedie/concept/. Consulted on April 2, 2014.

wider context "Marie est jolie...un cœur". By extension, the word *context* is also applied to the utterance(s) which precede(s) and follow(s) a given utterance in discourse. [3]

2.1.2 What do Translators Need Context for?

Following Roberts & Bosse-Andrieu (2006: 203), let us remind here that the translation problems translators have to face can be considered as source text-related (for comprehension of the source-text) or target-text related (for transfer into the target text) and classified into three main categories: *encyclopedic*, *linguistic* or *textual*. *Encyclopedic* problems encompass "general subject-related problems as well as more specific problems dealing with proper nouns–that is, a lack of familiarity with the topic of the text or with specific places or people mentioned in the text"; *linguistic problems* are defined as "those attached to specific words and phrases–that is, problems related to the comprehension or translation of a given word or phrase"; finally, *textual* problems are those concerned with text types and the internal organization or reproduction of a given text type.

Bowker (2011, 2012) draws a list of those items of contextual information that can prove "useful" for the translator to solve his source-text and target-text-related problems. They can be summed up as follows: (*i*) information about usage; this of course includes collocations, in particular which general-language words collocate with terms (see also Roberts 1994: 56), (*ii*) information about the frequency of use of a particular word or term, (*iii*) information about lexical and conceptual relations (such as synonymy, meronymy, hyperonymy etc.) (see also Marshman, Gariépy & Harms 2012, Rogers & Ahmad 1998), (iv) pragmatic information about style, register and genre (see also Varantola 1998), (*v*) information about usages to avoid. It thus seems to us that the items that are not situation-linked fall into the following categories: *conceptual* information and *linguistic* information.

To solve those problems and to make decisions, translators need external help, which they typically get by consulting other human experts and conventional resources such as dictionaries and term banks (Rogers & Ahmad 1998: 198).

2.1.3 Where do Translators Find Contextual Information?

As mentioned by Varantola (2006: 216), "the translator's problem-solving techniques have changed dramatically over the past decade or so". In addition to the above-mentioned conventional resources (monolingual and bilingual dictionaries–which have undergone radical changes–; term banks), translators now also partly rely on the information provided by corpora.

- Dictionaries

  It is mostly through examples that dictionaries provide contextual information. The empirical study on scientific and technical words (i.e. terms) in general bilingual and monolingual dictionaries carried out by Josselin-Leray (2005) has shown that up to 80.3% of users turned to dictionaries to find information about how to use the term in a sentence and that bilingual dictionaries always

3 We translated the quotation.

ranked higher in that respect. Among the respondents who chose that answer, it was the « language professionals » user group (which includes translators) that was mostly represented.

- Term Banks

  Term banks typically provide contextual information through the "context" section of the terminological record. The importance given to contextual information in terminological resources by translators is confirmed by the findings of the survey by Duran-Muñoz (2010): examples were considered to be "essential data" by the respondents, and among "desirable data", one found "a greater variety of examples" and "semantic information (semantic relations, frames)".

- Corpora

  Although corpus data is obviously intrinsically made of contextual information, it is a resource which seems still quite scarcely used by translators, as shown by the results of the survey by Duran-Muñoz (2010): only 5.09% of the participants quoted (parallel) corpora as being a terminological resource they "used *more*[4] when translating". However, 41.8% of the respondents to the Mellange Survey[5], which was carried out in 2005-2006 among trainee translators and professional translators, do claim they use corpora in their translation practice (the most frequent type being the corpora in the target language).

Although there seems to be a wide array of resources translators can turn to when they need contextual information, these resources do not necessarily meet the translators' needs.

2.2 The Shortcomings of Existing Resources regarding Context

2.2.1 The Dissatisfaction of Translators regarding Contextual Information in Existing Resources: a Hard Fact

We found it relevant to first look at the findings of various empirical surveys on the use of conventional resources by translators.

- Dictionary Use

  Before starting to compile the *Bilingual Canadian Dictionary*, Roberts (1994: 56) carried out a survey among its potential users in order to clearly identify their needs and reached the following conclusion: "Between one-third and one-half the members in each user group [of sophisticated second language users] appreciated, to varying degrees, the number of examples presented in their present most frequently used dictionaries. But between one quarter and one half of each group felt that improvement was needed in that respect". The study by Josselin-Leray (2005) reached the same conclusion: although users were overall satisfied by the examples provided by their dictionaries (between 41.3% and 67.5% of users said they were satisfied), the level of satisfaction was lower for bilingual dictionaries, and lower among the "language professionals" user group.

4 We added the italics.
5 http://mellange.eila.jussieu.fr/Mellange-Results-1.pdf. Consulted on April 10, 2014.

- Term Banks

  In the survey by Duran-Muñoz (2010), translators also had the opportunity to give their opinions regarding their needs; the second most repeated argument was "include more pragmatic information about usage and tricky translations", and the fifth one was "provide examples taken from real texts". The conclusion of her findings, found in Duran-Muñoz (2012: 82) is straightforward: "we can confirm that most of the terminological resources that are currently available (especially in electronic format) do not fulfill their requirements". Why is that so?

2.2.2 Why are Translators Dissatisfied?

Varantola (e.g. 1994) has written at great length about the *context-free* vs. *context-bound* dilemma faced by translators: dictionaries and term banks only provide context-free examples, i.e. examples that are perceived as prototypical and frequent, while what the translators need to find the suitable equivalent(s) is typically context-bound. Moreover, the examples/contexts provided are not varied enough. This is especially true of term banks, as underlined by Bowker (2011: 214-215) who explains the information found on those records is rather limited and usually consists in definitions and terms presented out of context, or in only a single context. She pinpoints a paradoxical situation in which the advances of research on terminology (especially the work on Knowledge-Rich Contexts, which we will introduce in 3.2) have not been integrated into the tools translators most commonly use, i.e. term banks.

However, Varantola (2006: 217) says the context-free/context-bound dilemma should now be qualified since "context-free definitions of concepts within a particular domain [which] were for a long time the theoretical ideal in terminological theory [...] are now replaced by less rigid, contextually relevant definitions". She ascribes it to the availability of large corpora, whose role is also now central in dictionary-compiling. Some dictionaries now even "provide access to more examples in the form of concordances from the corpus data that lie behind the dictionary". Corpora are no panacea, though. One of arguments against corpora is that they are "tools of shallow intelligence" (Varantola, 2006: 223) when they are raw and non-tagged or POS-tagged, since the user "is left to handle the manipulation, dissection and interpretation of results". In other words, compiling the corpus and analysing the corpus can be too tedious a task for translators who often work under tight constraints[6].

6 Another point worth mentioning is that, in the survey carried out for the Mellange, even though 94.4% of the respondents said they used Google to research terminology, 10.2% found that Google was limited for finding information on language use because the "search results [did] not provide enough context to be useful".

2.3 The Translator's Ideal Workstation?

In 1996, Atkins already suggested (p. 526) that the dictionary of the future should "give its users the opportunity to make their own decisions about equivalences" : the users "should be able to consult as many examples as they need of words used in their various senses, each in a variety of contexts with a variety of collocate partners". More recently, Bowker (2011: 215) suggested: "it would be more helpful for translators to have access not simply to term records that provide a single 'best' term with a solitary context, but rather to information that would allow them to see all possible terms in a range of contexts and thus find the solution that works best in the target text at hand". She insists on the fact that looking at a wide range of contexts should not be considered as a waste of time, and that this has been made easier thanks to corpus-analysis tools that present information in an easy-to-read format. She goes even further by suggesting (Bowker 2012: 391) that translators have access to the whole of the information that lexicographers usually rely on when devising a dictionary entry:

> In order to arrive to that entry, lexicographers have gone through a number of intermediary steps, where they learn about the various characteristics of the words and concepts being described, such as their grammatical and collocational behaviours, the different relationships that hold between words and their underlying concepts, and the characteristics that are necessary and sufficient for distinguishing one concept in an intensional definition.

However relevant that objective might be, it seems rather ambitious and difficult to achieve in the very near future, all the more so as "lexicographers'needs are very different from translators' corpus needs" (Varantola 2006: 217). Narrowing down that objective to providing more corpus-based context data in a way that is more in keeping with the actual working conditions of translators seems more feasible, which is why the main aim of the CRISTAL project is to help design a CAT tool that provides translators with customized contexts automatically retrieved from comparable corpora.
In order to reach that goal–which can also sound ambitious, we first decided to refine the notion of 'good contexts for translators' by doing two things: (*i*) we first looked at the way lexicographers deal with examples in dictionaries and terminographers deal with "Knowledge-Rich Contexts" (part 3), (*ii*) we devised an experiment with trainee translators focusing on what we thought to be "good contexts" (part 4).

3 Dictionary Examples and Knowledge-Rich Contexts

The thoughts of lexicographers and terminologists on "good examples" or "Knowledge Rich Contexts" provide some valuable insight into what a good context for translators might be. After examining those two aspects, we give our own definition of Conceptually Rich Contexts and Linguistically Rich Contexts.

3.1 Good Dictionary Examples

Many studies have underlined the importance of the illustrative component in dictionaries as a means to provide typical contexts about a word's meaning and usage (Atkins & Rundell 2008). In monolingual as well as in bilingual dictionaries, examples are meant to help the dictionary user both in the production and the comprehension process. They have therefore diverse functions (Rey-Debove 2005, Roberts 1994, Siepmann 2005): they can provide syntagmatic information about word patterns and collocations, together with paradigmatic information about words that are semantically-related (synonyms, hyperonyms, etc.). They may also give pragmatic and stylistic indications about registers and specific uses, or be used as a more concrete and accessible complement to definitions, with an epilinguistic dimension.
Authentic examples that meet at least some of these requirements are very difficult to extract from corpora:

> Finding good examples in a mass of corpus data is labour-intensive. For all sorts of reasons, a majority of corpus sentences will not be suitable as they stand, so the lexicographer must either search out the best ones or modify corpus sentences which are promising but in some way flawed (Rundell & Kilgarriff 2011).

Kilgarriff et al. (2008) have developed a method to automatically collect sentences that are good candidate dictionary examples, using two criteria: readability (judged from sentence length and average word length; it penalizes sentences with infrequent words, more than one or two non-a-z characters, or anaphora) and informativeness (judged from the density of collocates in the sentence).

3.2 Knowledge Rich Contexts (KRCs) in Terminology

Knowledge Rich Contexts play a very important role in identifying terms in specialized texts because they show conceptual relationships between terms. It is within that framework that Ingrid Meyer defined Knowledge Rich Contexts as "a context indicating at least one item of domain knowledge that could be useful for conceptual analysis." (Meyer 2001: 281). These contexts are used in order to develop knowledge extraction tools for text-based terminology and ontology building (Condamines & Rebeyrolle 2001). Rich contexts for terminologists typically contain terms that are specific to the domain together with linguistic patterns that signal the conceptual relations between these terms as illustrated below in Meyer (2001):
(1) Compost is an organic material deliberately assembled for fast decomposition.
(2) Compost contains nutrients, nitrogen, potassium and phosphorus.

This type of information helps building networks of terms, generally focusing on hyperonyms (example 1) and meronyms (example 2).

3.3 KRCs for Translators: Conceptually vs. Linguistically Rich Contexts

Based on the type of information needed by translators as described in 2.1.2, and the type of information provided by dictionary examples and Knowledge-Rich Contexts as detailed in 3.1 and 3.2, we decided to extend the notion of KRC in our experiment, considering under this category two types of contexts: those that provide 'conceptual' information about a given term-called "Conceptually Rich Contexts" (CRCs) in our study, and those that provide 'linguistic' information about that term "Linguistically Rich Contexts" (LRC). In our experiment, the contexts which are neither conceptual nor linguistic are considered as *poor* (cf. Reimerink *et al.* 2010: 1934).

4 An Experiment Centered on "Good Contexts" within the CRISTAL Project

The main aims of the experiment were (*i*) to check that rich contexts extracted from corpora are useful to translators, (*ii*) to identify which types of rich contexts (CRCs or LRCs) are the most useful to them.

A pilot study was carried out in December 2013 at the University of Geneva (Switzerland). This allowed us to test the protocol and to make the necessary adjustments for the two experiments we conducted in March 2014: one at the Université Catholique de l'Ouest (Angers, France), and one at the University of Toulouse le Mirail (Toulouse, France). We will now describe the main aspects of the protocol designed for the experimentation.

4.1 Protocol

4.1.1 Participants

For both the pilot study and the two experiments, participants were all trainee translators[7]. 7 students from the Faculty of Translation and Interpretation of the University of Geneva took part in the pilot study. There were 4 Masters' students and 3 PhD students. As for the experiments in Angers and Toulouse, the participants (42 in total) were students in their final year of a Master's Translation program.

4.1.2 Translation Task

The participants were asked to translate a text from English into French (i.e. from L2 into L1 for most students). The text is around 150 words long, it is a popular-science text on volcanology entitled "Cinder Cones"[8]. It was chosen because it is well-structured (the two phases of the building of a cinder

7 This is the case in many empirical studies on translation: see for example Bowker 1998, Künzli 2001, and Varantola 1998.

8 It was taken from *What's so hot about volcanoes?* by Wendell A. Duffield (2011), Mountain Press.

cone are described), because it contains a certain number of terms whose translation might be complex for a translator, even if they are not highly specialized (*basalt cinder cone, fountaining stage*...), and a number of syntactic patterns or collocations that are particularly tricky to translate (e.g. *bubble off*; transitive use of the verb *erupt*). Only one group out of the two was already a little familiar with the field of volcanology. The participants were allocated around 2 hours to translate the text, choose the relevant contexts and fill out an online questionnaire about the main translation difficulties and the use and usefulness of conventional resources and KRCs. Then several group interviews and a couple of one-to-one interviews were conducted.

4.1.3 Resources

Since we wanted the conditions of the experiment to be as close to a real-life context as possible for translators,[9] the participants had at their disposal the same type of resources as the ones they usually have when they translate in professional environment, i.e. various dictionaries and term banks. What made the experiment specific is that we added an extra resource, i.e. shortlisted contexts.

4.1.4 The Argos Interface

The participants used a customized interface, Argos, with four different windows (see figure 1): (*i*) one for the source-text, (*ii*) one for the target-text, (*iii*) one with several icons allowing access to term banks (*Termium, le Grand Dictionnaire Terminologique*), a specialized bilingual dictionary of earth sciences, a general bilingual dictionary, (*iv*) one with a list of shortlisted contexts.

9 This is what Ehrensberger-Dow & Massey (2008) call "ecological validity".

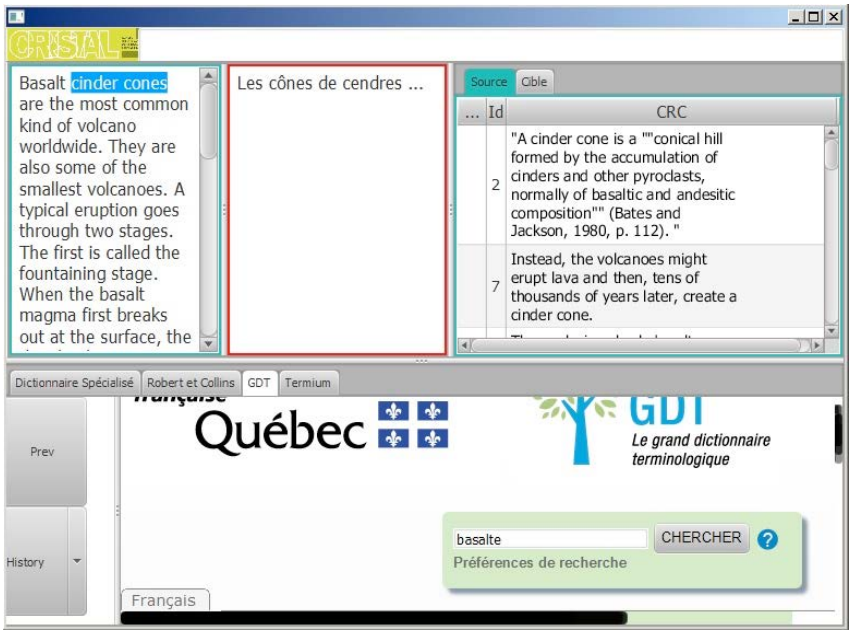


Figure 1: the Argos interface

4.1.5 The Contexts

Participants were provided with contexts in the source language, and contexts in the target language, which were presented in random order. These contexts had been carefully chosen beforehand according to a classification devised by the team of linguists: details about the selection of contexts and the type of contexts provided will be discussed in the next section (4.2). During the translation process, the participants had to choose the contexts that had been most useful to them when translating by clicking on them. Once the translator had typed in a word (in the source language or the target language) in the search window for contexts, the target text window was blocked in order to ensure the translator chose at least one context, or explicitly chose that none was useful.

4.1.6 Extra Data Compiled

In addition to the final translations themselves and the answers to the online questionnaires, the data compiled comprise the following:

- screen-recordings performed through specific software[10]
- logs: all keyboard activity, as well as change of windows shifts, was recorded
- audio recordings of the two types of interviews[11] during which the participants were asked to give more detail about the usefulness of some given contexts.

10 BBFlashback Express.

11 In the one-to-one interview, the participant viewed part of the screen recording (the extracts corresponding to his translation of two terms for which contexts were provided: *cinder* and *fountaining stage*) and was asked to verbalize what he was doing, following the methodology used by Ehrenbersger-Dow and Massey (2008).

4.2 The List of Contexts

The list of contexts provided to the participants was created with two principles in mind: we wanted to compile a large enough set of contexts, in order to limit the chance that the translator would look for contextual information on one particular word and get no results. At the same time, considering that the selection of contexts is a labor-intensive task, we wanted to limit ourselves to a reasonable number of words, and to keep only words for which our definition of Rich and Poor Contexts applies (see 3.3 above): in particular, the notion of Conceptually Rich Contexts (CRCs) is irrelevant for very familiar words that are not characteristic of the field of volcanology from which the text is drawn. This compromise was difficult to reach, so we took advantage of the pilot study to adjust and complement the list of terms that we had initially created.

4.2.1 Term Selection

For the pilot study in Geneva, we compiled a first set of contexts illustrating the use of:

- 7 words (noun, verbs and adjectives) in the source language, namely some lemmas from the text to be translated that we regarded as terminological units, or at least as words related to the field of volcanology (*basalt magma*, *blobs*, *cinder*, *cinder cone*, *fountaining*, *scoria*, *vesicles*)
- 11 words in the target language, selected among the possible equivalents of the corresponding source words.

We considered both simple and multiword units.

For example, contexts are provided for the word in bold type in the following sentence:

(3) As the **cinders** fall back to the Earth, they form layers that pile up into a cone-shaped hill.

One outcome of the pilot study was that the initial list of words proved to be very insufficient, especially in the target language: the logs compiled thanks to Argos (cf. 4.1.6) provided us with a much larger list corresponding to words that had actually been typed in the search window by the participants in order to get contexts. We thus decided to complement the first list with words that have been searched for by at least 2 participants. The final list for the experiments in Angers and Toulouse contains contexts for 22 English words and 41 French words.

In the same sentence as the one mentioned above, contexts were provided for four words (in bold type) instead of just one (i.e. 2 nouns, 1 verb, 1 adjective):

(4) As the **cinders** fall back to the Earth, they form **layers** that **pile up** into a **cone-shaped** hill.

4.2.2 Context Selection

The contexts were selected from several sources:

- we preferentially used a 800,000 word corpus of volcanology composed of specialized and popular science texts, which is used for the CRISTAL project as a whole,
- this source was complemented by a variety of web sites, giving priority where possible to texts dedicated to the presentation of volcanology to a wide audience.

We chose not to test the readability dimension of contexts (3.1.), so we only selected contexts that meet the criteria of readability (well-formed, not too long, with no anaphora elements, etc.). Contexts are one or two sentences long.

As explained before, our aim is to test whether the opposition between rich and poor contexts as defined in section 3 is relevant for the translation task. As a consequence, we annotated contexts according to this dimension. In figure 1, we give 3 examples illustrating (1) a linguistically rich context for the word *basalt* (with the presence of the term *basalt lava*), (2) a conceptually rich context providing a definition of the term, (3) a poor context. Note that linguistic and conceptual richness can combine in some contexts, which is not the case here. When possible, the conceptually rich contexts were classified into the following subcategories: definition, meronymy, hyponymy and co-hyponymy.

	Term	Rich or Poor	CRC	LRC	Type of CRC	Context
1	basalt	rich	no	yes	n/a	Shield volcanoes are made of thousands of thin basalt lava flows.
2	basalt	rich	yes	yes	def	Basalt is dark volcanic rock made up of small crystals and glass.
3	basalt	poor	no	no	n/a	When basalt enters water passively, it forms pillow basalt.

Figure 5: Examples of contexts for the word *basalt.*

We intended to balance the number of poor and rich contexts for each word. This proved impossible to achieve in many cases, since the great majority of contexts exhibit at least one relevant collocate. We collected 10 contexts per word, including no less than 2 poor contexts, totalling 222 contexts in English and 441 contexts in French.

4.3 First Results

We report here some preliminary observations about the results.

For the source language, 48% of the available contexts were chosen by at least one participant (108 contexts) versus 36% for the target language (152 contexts). This is an indication that the contexts are perceived as helpful, but the data are very dispersed, since about 40% of the selected contexts were chosen by only one participant in either language. Some terms are found several times in this list (fountain, cinder cone, basalt magma, and their French counterpart), showing specific translation problems.

To get a first picture of the results, we have chosen to focus on the 20 contexts that were selected most in either language. Each context was chosen between 4 and 14 times. The following examples show two of the most frequent ones.

(5) Hawaiian Eruptions are types of volcanoes and types of eruptions wherein basaltic lava is normally thrown up the air in jets. This process is called **fountaining**.

(6) During an eruption of gas-rich magma, small **blobs** of magma are ejected.

Example 5 is a conceptually rich context, more specifically a definition. Example 6, which contains several collocations (*blobs of magma*, *blobs ejected*), is a linguistically-rich context.

	Rich contexts	**CRCs (definitions)**	**LRCs**
Source language			
All available contexts	69%	25% (11%)	52%
20 most selected contexts	90%	65% (60%)	25%
Target language			
All available contexts	70%	29% (14%)	49.5%
20 most selected contexts	90%	55% (40%)	35%

Table 2: Distribution of the contexts.

Table 2 makes a comparison between this subset of contexts and all the contexts that were made available. First, this shows that the great majority of contexts that are considered as helpful by the participants are rich contexts (90%). Second, participants show a strong preference for conceptually-rich contexts, mainly definitions, as opposed to linguistically-rich contexts. Yet we can observe that if this overall pattern applies to both languages, there are some differences: the distribution between LRCs and CRCs is different when the users are exploring source (English) and target (French) contexts. They seem to give a higher priority to CRCs and definitions in the source language. This is consistent with the idea that the CRCs should provide help for the comprehension of terms and LRCs should be more useful when checking the usage of the words in the target language.

These are encouraging results: they confirm our assumption that rich contexts are seen as helpful by the participants, and they suggest differences in the way the translator uses these contexts in the source and target language. However, these first observations must be confirmed and complemented by the analysis of the entire set of data and the analysis of the replies to the questionnaires.

5 Conclusion

The notion of context is where lexicography, terminology and translation meet. Even though the specific needs of translators regarding their resources now seem quite clearly identified, addressing them still seems quite challenging. We hope the findings of the CRISTAL project will help tailor the tools according to the translator's profiles in one aspect, that of contextual information.

To fulfil that objective, we plan to explore in detail the considerable amount of data we have collected (around 60 hours of screen recordings, and just as much structured translation logs). The evidence gathered will enable us to answer the following questions:

- apart from definitions, do some sub-categories of KRCs play a specific role in the translation process?
- in which precise situations do translators give preference to KRCs over conventional resources such as monolingual or bilingual dictionaries?
- what is the impact of the use of KRCs and the other resources on the translation quality of the 49 final translations

The main challenge the CRISTAL project plans to address in the near future is to devise a method to automatically retrieve the 'good contexts' whose main features will have then been identified.

6 References

Atkins, Beryl T. Sue. (1996). "Bilingual Dictionaries: Past, Present and Future." In: Euralex'96 Proceedings, Martin Gellerstam et al., ed. Gothenburg: Gothenburg University. 515-590.

Baker, M. (2006). Contextualization in translator- and interpreter-mediated events. *Journal of Pragmatics*, 38(3), pp. 323-337.

Bowker, L., (1998). « Using Specialized Monolingual Native-Language Corpora as a Translation Resource: a Pilot Study ». *Meta,* 43(4), pp.631-651.

Bowker L., (2011). "Off the record and on the fly," *Corpus-based Translation Studies: Research and Applications* (Eds. A. Kruger, K. Wallmach and J. Munday). London/New York: Continuum, pp. 211-236.

Bowker L., (2012). "Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities", in S. Granger, M. Paquot (eds) *Electronic Lexicography*, Oxford University Press, pp. 379-387.

Condamines, A., Rebeyrolle, J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results. In D. Bourigault, C. Jacquemin & M.-C. L'Homme *Recent Advances in Computational Terminology*, John Benjamins, pp.127-148.

Durán Muñoz, I. (2010). Specialised lexicographical Resources: a Survey of Translators' needs. In S. Granger & M. Paquot (eds.), *eLexicography in the 21st century: New challenges, new applications*, *Proceedings of Elex2009. Cahiers du Cental*, vol. 7. Louvain-la-Neuve. Presses Universitaires de Louvain, pp. 55-66.

Durán Muñoz, I. (2012). Meeting translators' needs: translation-oriented terminological management and applications. *Journal of Specialised Translation* (18), pp. 77-92.

Ehrensberger-Dow, M., Massey, G. (2008): Exploring Translation Competence by Triangulating Empirical Data. In *Studies in Translation*. 16, 1–20.

Halliday, M.A.K. (1999). "The notion of Context in Langue education). In M. Ghadessy (ed.) *Text and Context in Functional Linguistics*. Philadelphia: John Benjamins Publishing Company.

Josselin-Leray, A. (2005). Place et rôle des terminologies des terminologies dans les dictionnaires unilingues et bilingues. Etude d'un domaine de spécialité : volcanologie. Unpublished PhD thesis, Université Lyon 2, France.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P., (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *proceedings of XIII EURALEX Congress*, E. Bernal & J. DeCesaris (Eds), Barcelona, Universitat Pompeu Fabra, pp.425-431.

Künzli, A. (2001). Experts vs. novices. L'utilisation de sources d'information pendant le processus de traduction. In *Meta*, 46, pp. 507-523.

Marshman, E., J., Gariépy & Harms, C. (2012). Helping translators manage terminological relations: Storing and using occurrences of terminological relations and lexical relation markers. *JoSTrans: Journal of Specialised Translation* 18, 30-56.

Melby, A.K. (2010). Context in translation: Definition, access and teamwork. *Translation & Interpreting*, 2 (2), pp. 1-15.

Meyer, I., (2001). Extracting Knowledge-Rich Contexts for terminography: A conceptual and methodological framework. In D. Bourigault, C. Jacquemin & M.-C. L'Homme *Recent Advances in Computational Terminology*, John Benjamins, pp. 279-302.

Reimerink, A., Garcia de Quesada, M. & Montero-Martinez, S. (2010). Contextual information in terminological knowledge bases: a multimodal approach, *Journal of Pragmatics*, 42(7), pp. 1928-1950.

Rey-Debove, J., (2005). Statut et fonction de l'exemple dans l'économie du dictionnaire. In M. Heinz (eds), *L'exemple lexicographique dans les dictionnaires français contemporains*. In Proceedings of the « Premières journées allemandes des dictionnaires » (Klingenberg am Main, 25-27 juin 2004). Max Niemeyer Verlag, Tübingen, pp.15-20

Roberts, R. P. (1994). Bilingual Dictionaries Prepared in Terms of Translators' Needs, *Proceedings of CTIC 3rd Conference*, *Translation in the Global Village*, Banff, CTIC, pp. 51-65.

Roberts, R.P. & Bosse-Andrieu J. (2006). Corpora and Translation. In Bowker L. (ed.), *Lexicography, Terminology, and Translation – Text-based Studies in Honour of Ingrid Meyer.* Ottawa : Presses de l'Université d'Ottawa, pp. 201-214.

Rogers M. & Ahmad K. (1998). The Translator and the Dictionary: Beyond Words? In: B.T. Sue Atkins (ed.) *Using Dictionaries. Studies of Dictionary Use by Language Learners and Translators.* Tübingen: Niemeyer (Lexicographica. Series Maior), pp.193-204.

Siepmann, D. (2005). Discourse markers across languages: A contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography. Routledge.

Varantola, K. (1994). The Dictionary User as Decision Maker. Sixth Euralex conference proceedings, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands, pp. 606-611.

Varantola, K. (1998). Translators and their Use of Dictionaries. In B.T.S. Atkins (ed), User Needs and User Habits. Studies *of Dictionary Use by Language learners and Translators*, Tübingen: Niemeyer (Lexicographica. Series Maior), pp. 179–192.

Varantola, K. (2006). The Contextual Turn in Learning to Translate. In Bowker L. (ed.), *Lexicography, Terminology, and Translation – Text-based Studies in Honour of Ingrid Meyer.* Ottawa : Presses de l'Université d'Ottawa, pp. 215-226.

Kontextbasierte lexikalische Kontrolle von Anforderungsdokumenten

Jennifer Krisch
Daimler AG
jennifer.krisch@daimler.com

Abstract

Das Verwalten und Prüfen von Anforderungsdokumenten ist ein sehr wichtiger Bereich in den frühen Phasen von Industrieprojekten. Je früher unpräzise und unvollständige Anforderungen identifiziert werden, desto geringer sind die Folgekosten und die Überarbeitungszeit. Unpräzise und unvollständige Anforderungen resultieren häufig aus der Verwendung von Weak-Words, d.h. von unscharfen Wörtern oder Konstruktionen, die in bestimmten Kontexten Mehrdeutigkeiten auslösen. Um Fehlinterpretationen aufgrund von Weak-Words entgegenzuwirken, müssen diese Wörter oder Konstruktionen identifiziert werden. Eine reine Wortsuche reicht hierbei nicht aus, weil Weak-Words nur in bestimmten Satzkontexten Mehrdeutigkeiten auslösen. Der Beitrag beschreibt eine als Prototyp entwickelte automatisierte Weak-Word-Analysemethodik, durch welche ausgewählte Weak-Words identifiziert werden und bei der aufgrund des Satzkontexts entschieden wird, ob eine Warnung bezüglich eines Mangels an Präzision und Eindeutigkeit des Texts an den Autor zurückgegeben werden soll. Bei der Weak-Word-Analyse kommt ein Lexikon zum Einsatz, in welchem die Weak-Words selbst und deren Kontexte abgelegt sind. Die Evaluation des entwickelten Werkzeugs hat ergeben, dass die Anzahl der Warnungen, die an den Benutzer zurückgegeben werden, durch eine Kontextanalyse im Verhältnis zur wortbasierten Suche erheblich reduziert werden kann, ohne dass problematische Kontexte übersehen werden. Dies stellt eine Verbesserung gegenüber den bisher bekannten Systemen dar.
Keywords: Weak-Words; Kontext-Analyse; Linguistische Annotation

1 Einführung

Das Verfassen von Anforderungen ist in der Industrie der erste Schritt, um von einer Idee zu einem Produkt zu kommen. Die Ideen der Projektbeteiligten müssen angemessen dokumentiert werden, damit jeder weiß was im Projekt gefordert wird. Anforderungstexte sind also „Aussagen über Eigenschaften oder Leistungen eines Produktes, eines Prozesses oder der am Prozess beteiligten Personen“ (Rupp 2007).

Die Qualität einer Anforderung wird durch *Qualitätskriterien* festgelegt. Je mehr Qualitätskriterien von einer Anforderung erfüllt werden, desto höher ist deren Qualität. Anforderungen werden meist in

natürlicher Sprache verfasst, die aber das Risiko von Mehrdeutigkeit, mangelnder Präzision oder Unklarheiten über die rechtliche Verbindlichkeit birgt; daher ist es wichtig, Anforderungen vor allem auf die Qualitätskriterien *Präzision* und *Eindeutigkeit* zu prüfen. Eine Anforderung erfüllt das Qualitätskriterium *Präzision*, „wenn diese keine ungenauen Angaben enthält, wenn wo immer möglich und sinnvoll quantitative Angaben gemacht werden und wenn alle Angaben genauso präzise sind, wie es für die Problemstellung erforderlich ist" (Daimler AG 2013). Das Qualitätskriterium *Präzision* hängt eng mit dem Qualitätskriterium *Eindeutigkeit* zusammen:

> Eine eindeutige Anforderung [sollte] nur auf eine Art und Weise verstanden werden können [...]. Es darf nicht möglich sein, andere Sachverhalte hineinzuinterpretieren. Alle Leser einer Anforderung sollten zu einer einzigen, konsequenten Interpretation der Anforderung gelangen. (Rupp 2007)

Eines der zentralen Qualitätsprobleme von Präzision und Eindeutigkeit ist (fach-) lexikographischer Natur: *Weak-Words*, auch *unscharfe Wörter* genannt, verstoßen in gewissen (Satz-)Kontexten gegen die aufgeführten Qualitätskriterien. Weak-Words sind „Wörter oder Phrasen, deren Benutzung in einem Freitext darauf schließen lässt, dass der Freitext mit hoher Wahrscheinlichkeit unpräzise ist" (Melchisedech 2000). Ein Beispiel für ein Weak-Word ist das Adverb *lang* in der Anforderung, die in Beispiel (1) dargestellt ist.

(1) *Die Taste muss lang gedrückt werden.*

Das Wort *lang* eröffnet hier einen großen Interpretationsspielraum. Ein Projektbeteiligter wird das Wort *lang* als eine Zeitspanne von drei Sekunden interpretieren, ein anderer möglicherweise als eine Zeitspanne von 500 Millisekunden. Dieser Interpretationsspielraum kann im Produktentwicklungsprozess zu großen Problemen führen, weil ein Fehler in der Anforderungsformulierung sich durch viele Teile eines Projekts ziehen kann und möglicherweise erst zu einem sehr späten Zeitpunkt aufgedeckt wird. Aus diesem Grund ist es wichtig, Anforderungen auf das Vorkommen von Weak-Words zu prüfen. Praktisches Ziel der hier beschriebenen Arbeiten ist es, einen Teil dieser Prüfung zu automatisieren.

Die Annahme liegt nahe, dass jedes Auftreten eines Weak-Words die Qualität einer Anforderung mindert, man deshalb Weak-Words identifizieren und Anforderungsautoren bitten sollte, ihre Formulierungen nochmals zu überarbeiten. Es ist aber zu beachten, dass Weak-Words nicht immer die Qualität einer Anforderung mindern. Dies ist im hohen Maße vom Kontext abhängig, wie man in Beispiel (2) sehen kann.

(2) *Die Taste muss 2 Sekunden lang gedrückt werden.*

Hier eröffnet das Weak-Word *lang* keinen Interpretationsspielraum, sondern trägt ganz im Gegenteil dazu bei, die Anforderung zu präzisieren. Bei einer Weak-Word-Analyse muss folglich der Kontext der Weak-Words miteinbezogen werden, um entscheiden zu können, ob diese in einer Anforderung Interpretationsspielraum eröffnen oder nicht (Krisch 2013).

2 Stand der Technik

Es existieren bereits Prüfwerkzeuge, die eine automatisierte Weak-Word-Analyse durchführen. Das sind u.a. *ReQualize* (Heidenreich 2010), *DESIRe*® (Stöckel et al. 2009) und *QuARS* (Lami 2005). Bei allen drei Werkzeugen kann gezielt nach Wörtern gesucht werden, welche in einer Liste abgelegt sind. Wird ein solches Wort in einem Lastenheft gefunden, werden eine Warnung und ein Hinweis an den Benutzer zurückgegeben, dass er die entsprechende Anforderung nochmals überarbeiten sollte. Die Kontexte der Wörter werden bei diesen Werkzeugen nicht in die Analyse miteinbezogen. Bei einer großen Textmenge hat dies die Konsequenz, dass sehr viele Fehlermeldungen produziert werden und der Benutzer sehr oft unnötig gewarnt wird; das führt wiederum dazu, dass das System vom Benutzer nicht akzeptiert wird.
Diese unnötigen Warnungen sollen durch den hier beschriebenen Ansatz zur Erstellung kontextabhängiger Prüf-Regeln für bestimmte Weak-Words reduziert werden.

3 Kontextanalyse und Entwicklung eines lexikalischen Prüfwerkzeugs

3.1 Korpuserstellung

Um die Kontexte ausgewählter Weak-Words identifizieren zu können, musste zunächst ausreichend Textmaterial bereitgestellt werden. Hierfür wurden Lastenhefte aus einer Datenbank der potentiellen Anwender exportiert, durch ein eigens entwickeltes Werkzeug tokenisiert (Identifikation von Wort- und Satzgrenzen) und linguistisch annotiert (mit den *mate*-Werkzeugen, (Bohnet 2009)). Die tokenisierten Lastenhefte wurden zunächst lemmatisiert, anschließend mit Wortart- und Morphologie-Annotationen versehen und zuletzt noch auf der Grundlage von Dependenzsyntax analysiert (geparst). Abbildung 1 zeigt ein Beispiel eines Anforderungssatzes nach der computerlinguistischen Annotation. In schwarz ist die jeweilige Satz- bzw. Wortnummer dargestellt und in dunkelblau die Wortform (Token). Orange markiert sind die Lemmata, grün die Wortart-Tags, pink die Morphologie-Tags, rot die Abhängigkeiten zwischen den einzelnen Tokens und hellblau die syntaktischen Informationen. Das Wort *Wiedereinschalten* beispielsweise ist das zweite Token im ersten Satz mit der ID *SPM-8014*. Die Grundform, also das Lemma, ist *Wiedereinschalten*. Außerdem ist es ein Nomen, hat den Kasus Nominativ, steht im Singular und ist neutrum. *Wiedereinschalten* bezieht sich auf Token 3, also auf *erfolgt*, und ist dessen Subjekt.
Die linguistische Annotation der Anforderungen dient vor allem zur Generalisierung der in Abschnitt 3.2 entwickelten Regeln.

```
1_1_SPM-8014    Das     der     ART     case=nom|number=sg|gender=neut   2       NK
1_2_SPM-8014    Wiedereinschalten       Wiedereinschalten       NN      case=nom|num-
ber=sg|gender=neut 3    SB
1_3_SPM-8014    erfolgt erfolgen        VVFIN   number=sg|person=3|tense=pres|-
mood=ind 0      --
1_4_SPM-8014    immer   immer   ADV     _       3       MO
1_5_SPM-8014    durch   durch   APPR    _       3       MO
1_6_SPM-8014    eine    ein     ART     case=acc|number=sg|gender=fem           8
NK
1_7_SPM-8014    erneute erneut  ADJA    case=acc|number=sg|gender=fem|de-
gree=pos 8      NK
1_8_SPM-8014    Anforderung     Anforderung     NN      case=acc|number=sg|gender=fem
5       NK
1_9_SPM-8014    vom     von     APPRART         case=dat|number=sg|gender=neut  8
MNR
1_10_SPM-8014   Logikträger      Logikträger     NN      case=dat|number=sg|gender=neut
9       NK
1_11_SPM-8014   .       --      $.      _       10      --
```

Abbildung 1: Darstellung eines Anforderungs-Satzes mit linguistischen Annotationen.

Insgesamt wurde mit diesem Vorgehen ein Korpus erzeugt, das 88 166 Sätze und 990 011 Tokens enthält. Auf Grundlage dieses Korpus wurden Experimente zur Identifikation präziser und unpräziser Kontextmuster ausgewählter Weak-Words durchgeführt. Für die Evaluation wurde ein Korpus mit neuen Lastenheften erzeugt (vgl. Kapitel 4).

3.2 Ermittlung der Kontexte ausgewählter Weak-Words

Die verwendete Arbeitsmethodik ist von der korpusbasierten Lexikographie inspiriert: Zur Entwicklung lexemspezifischer Regeln für die Identifikation von Weak-Words im Kontext wurde das Anforderungs-Korpus interaktiv mit dem Suchwerkzeug CQP (Evert & Hardie 2011) durchsucht. In Tabelle 1 sind Beispiele für „gute“ und „problematische“ Kontexte des potentiellen Weak-Words *mal* dargestellt.

Zusammensetzung	Beispiel
CARD mal	3 mal
erst.+ mal	das erste Mal
XXX_X... mal	<Parameter> mal
tbdX (default CARD) mal	tbd1 (default 10) mal
[0-9]+ten mal	nach dem 3ten Mal
nächste.* mal	das nächste Mal
letzt.+ mal	beim letzten Mal
[a-z] mal	x mal
jedes mal wenn	jedes mal wenn
auch mal	auch mal

Tabelle 1: Kontextmuster des Weak-Words *mal/Mal.*

Die grün markierten Spalten sind Kontextmuster von *mal*, in welchen die Verwendung des Wortes nicht dazu führt, dass die Anforderung unpräzise wird. Angaben wie *3 mal* oder *nach dem 3ten Mal* sind präzise und schaden der Qualität einer Anforderung nicht. Auch Kontextmuster, bei denen ein Parameter vor dem Wort *mal* steht, sind in Ordnung. Genauso verhält es sich mit Konstruktionen, die eine flektierte Form von *erst* und ein nachfolgendes *Mal* enthalten und mit Konstruktionen, die die Form *tbdX (default 3) mal*[1] aufweisen. Die rot markierten Muster hingegen sollten in Anforderungen vermieden werden, da sie die Verständlichkeit der betreffenden Anforderungen mindern. Ein Beispiel hierfür ist in (3) gegeben.

(3) *Die Taste kann auch mal klemmen.*

In (3) wird nicht eindeutig spezifiziert, wann dieser Fall, dass die Taste klemmen kann, genau auftreten kann. Das Weak-Word *mal* eröffnet einen großen Interpretationsspielraum und aus diesem Grund müssen Anforderungen wie diese zur nochmaligen Überarbeitung an den Benutzer zurückgeliefert werden.

Die orange markierten Muster sind Fälle, bei denen aufgrund des Satzkontexts alleine nicht sicher ist, ob sie präzise genug sind. In manchen Fällen wird es klar sein, wann *das nächste Mal* oder *das letzte Mal* ist, in anderen Fällen wiederum nicht. Hier sollte zur Sicherheit eine Warnung an den Benutzer ausgegeben werden.

Der Grundgedanke der Weak-Word-Analyse ist, dass nur die wirklich präzisen Konstruktionen mit potentiellen Weak-Words dem Benutzer nicht zurückgeliefert werden sollen. Alle anderen Konstruktionen (rot und orange markiert) sollten an den Benutzer zurückgegeben werden, damit dieser sich die betreffende Anforderung nochmals anschauen und sie überarbeiten kann. Dies kann zwar dazu

1 Konstruktionen wie diese sind Standardkonstruktionen in Lastenheften und müssen deshalb nicht als Warnung an den Benutzer zurückgegeben werden.

führen, dass weiterhin unnötige Warnungen an den Benutzer geliefert werden, dafür wird aber auch keine unpräzise Anforderung übergangen.

In gleicher Weise wurden auch die Kontextmuster für das Weak-Word *lang* identifiziert. In Tabelle 2 sind ausschließlich die präzisen Kontextmuster von *lang* angegeben. Insgesamt wurden acht unpräzise Kontextmuster und drei Kontextmuster ermittelt, bei denen nicht klar entschieden werden konnte, ob sie immer zu präzisen Anforderungen führen oder nicht. Konstruktionen wie *länger als 3 Sekunden* und *langes Drücken von 1 Sekunde* sind präzise und lösen keinen Interpretationsspielraum aus. Ein ebenfalls präzises Beispiel ist *länger als für die Nachlaufzeit*. Hier weist der bestimmte Artikel darauf hin, dass der Begriff *Nachlaufzeit* in einer vorangehenden Anforderung schon verwendet wurde und somit spezifiziert sein sollte. An dieser Stelle wird zwar keine Warnung an den Autor ausgegeben, am Schluss einer Analyse werden aber dem Autor alle Nomina zurückgegeben, denen ein bestimmter Artikel vorangeht und die mit dem Wort *lang* zusammenhängen, mit der Bitte um Prüfung, ob die Wörter eindeutig beschrieben sind. Tritt hingegen eine Konstruktion mit unbestimmtem Artikel auf wie in *länger als eine Erholungszeit*, wird in jedem Fall eine Warnung ausgegeben, da der unbestimmte Artikel darauf hinweist, dass das Wort *Erholungszeit* an dieser Stelle neu eingeführt wird und noch nicht vollständig beschrieben ist.

Für die Weak-Words *kurz* und *schnell* wurden dieselben Regeln verwendet, die auch bei *lang* zum Tragen kommen. Ziel war es zu testen, ob sich Regeln eines Weak-Words auf semantisch ähnliche Weak-Words übertragen lassen, d.h. ob die definierten Regeln generalisierbar sind. Die Weak-Words und die korrespondierenden Regeln wurden in einem Lexikon abgelegt, in einen ersten Prototypen implementiert und anschließend evaluiert.

Zusammensetzung	**Beispiel**
lang als ca.	länger als ca. 3 Sekunden
CARD NN/NE lang	15 Sekunden lang
lang als CARD	länger als 3 Sekunden
lang als für def.ART	länger als für die Nachlaufzeit
wie lang	wie lange
lang als def.ART	länger als die Verzögerungszeit
XXX_X_... lang	<Parameter> lang
lang VERB als	nicht länger betrieben als
lang als XXX_X_...	länger als <Parameter>
lang CARD NN/NE	länger 500 Millisekunden
lang NN/NE (</> ...)	langes Drücken (> 1 Sekunde)
lang NN/NE von	langes Drücken von 1 Sekunde
lang def.ART NN/NE	länger der Nachlaufzeit

Tabelle 2: Präzise Kontextmuster des Weak-Words *lang*.

Für jedes Weak-Word wurde ein Lexikoneintrag mit den entsprechenden Regeln erstellt. Die Kontextregeln sind als Suchmuster in der Skriptsprache Python, auf der Basis der in Abschnitt 3.1 illustrierten Annotationen, implementiert. Das Lexikon enthält auf dem Stand von Frühjahr 2014 die Weak-Words *mal*, *lang*, *kurz* und *schnell*. Neue Wörter und zugehörige Regeln können analog hinzugefügt werden. Im Lexikon stehen nur die präzisen Kontextmuster der Weak-Words. Dies gewährleistet, dass alle präzisen Kontexte vom System erkannt werden und alle anderen Kontexte, die nicht im Lexikon vermerkt sind, an den Benutzer zurückgegeben werden. Dadurch werden einerseits Fehlermeldungen, die bei anderen Tools unnötigerweise ausgegeben werden, vermieden, und andererseits keine unpräzisen Anforderungen übergangen.

4 Evaluation

Die Regeln wurden anhand eines Korpus von insgesamt 99 955 Sätzen und 1 160 409 Tokens evaluiert. In diesem Korpus kam das Weak-Word *mal* 44 mal vor, *lang* trat 147 mal, *kurz* 106 mal und *schnell* 207 mal auf. Es wurden alle Wortformen der vier Weak-Words analysiert, nicht nur die Grundformen.
Tabelle 3 zeigt die Auswertung der Ergebnisse für die 44 Erscheinungsstellen von *mal*. 26 Anforderungen wurden automatisch korrekt als präzise Anforderung klassifiziert. 15 unpräzise Anforderungen wurden ebenfalls richtig als unpräzise Anforderungen klassifiziert. Drei präzise Anforderungen hingegen wurden fälschlicherweise als unpräzise interpretiert, d.h. es wurden insgesamt drei überflüssige Warnungen an den Benutzer zurückgegeben. Interessant ist vor allem, dass keine unpräzise Anforderung übersehen wurde.
Das Ziel der Weak-Word-Analyse ist es, die unnötigen Warnungen an den Benutzer zu reduzieren, es soll dabei aber keine unpräzise Anforderung übergangen werden. Durch die Formulierung von Kontextregeln des Weak-Words *mal* für unproblematische Fälle konnten gegenüber einer wortbasierten Strategie 26 Warnungen entfallen, was fast 60 % der Gesamtmenge an Warnungen des wortbasierten Systems entspricht.

automatische Analyse →

manuelle Analyse ↓

	Unpräzise Anforderung	**Präzise Anforderung**
Unpräzise Anforderung	15	0
Präzise Anforderung	3	26

Tabelle 3: Evaluation der Ergebnisse für das Weak-Word *mal*.

Ein Beispiel für eine falsch klassifizierte Anforderung, die das Wort *mal* enthält, ist in (4) gegeben.

(4) *Die Funktion darf nicht ein viertes Mal gestartet werden.*

Die beiden anderen falsch klassifizierten Anforderungen sind von derselben Art wie (4). Der POS-Tagger interpretiert das Zahlwort *viertes* nicht als Zahl (Tag = *CARD*), sondern als Artikel. Eine Konstruktion wie diese wird bisher nicht mit den Regeln aus Tabelle 1 abgefangen. Eine entsprechende Regel sollte in zukünftigen Arbeiten entworfen und in das Lexikon integriert werden.
In Tabelle 4 sind die entsprechenden Precision- und Recall-Werte und die daraus resultierenden F1-Measure-Werte für die Klassen *Unpräzise Anforderung* und *Präzise Anforderung* zu sehen. Der Precision-Wert gibt den Anteil der Suchergebnisse an, die tatsächlich relevant sind in Bezug auf die Gesamtgröße des Ergebnisses (Dörre, Gerstl, & Seiffert 2004). Der Precision-Wert der Klasse *Unpräzise Anforderung* liegt beispielsweise bei 83,33 %. Der Grund hierfür ist, dass zwar alle relevanten Anforderungen gefunden wurden, aber noch überflüssige Warnungen an den Benutzer geliefert werden. Dies ist bei der Klasse *Präzise Anforderung* nicht der Fall, und deshalb beträgt hier der Precision-Wert 100 %.
Der Recall-Wert bezeichnet den Anteil der relevanten Suchergebnisse in Bezug auf die Menge aller für diese Suche relevanten Ergebnisse (Dörre, Gerstl, & Seiffert 2004). Für die Klasse *Präzise Anforderung* werden nicht alle relevanten Instanzen gefunden, für die Klasse *Unpräzise Anforderung* allerdings schon, was die Werte von 89,66 % und 100 % erklärt.

	Klasse: Präzise Anforderung	Klasse: Unpräzise Anforderung
Precision	100 %	83,33 %
Recall	89,66 %	100 %
F_1-Measure	94,55 %	90,91 %

Tabelle 4: Precision, Recall und F_1-Measure: *mal*.

Das F1-Maß ist eine Kombination aus Precision und Recall (geometrisches Mittel) und gibt einen Gesamtüberblick über die Performanz des entwickelten Werkzeugs in Bezug auf die einzelnen Klassen.
Tabelle 5 zeigt die Auswertung der Ergebnisse für das Weak-Word *lang*. 129 Kontexte von *lang* konnten korrekt klassifiziert werden. Es gab nur 18 Fehlklassifizierungen. Insgesamt konnten 69 unnötige Warnungen vermieden werden, was in der Gesamtmenge knapp 50 % entspricht.

automatische Analyse →

manuelle Analyse ↓

	Unpräzise Anforderung	**Präzise Anforderung**
Unpräzise Anforderung	60	0
Präzise Anforderung	18	69

Tabelle 5: Evaluation der Ergebnisse für das Weak-Word *lang*.

Die Fehlanalysen beruhen vor allem auf Klammer- und Adjunkt-Konstruktionen, die vom System bisher nicht behandelt werden können. Ein Beispiel für eine Klammerkonstruktion, die falsch klassifiziert wurde, ist in (5) zu sehen.
(5) *Die Eingriffe sollen nicht innerhalb einer applizierbaren langen Vergangenheit (5 Sekunden) liegen.*

Für Konstruktionen wie diese wurden bisher keine Regeln formuliert. Inwiefern hier eine Regel entworfen werden kann, soll in weiteren Arbeiten untersucht werden.

Weak-Word	*lang*		*kurz*		*schnell*	
Klasse	*Präz. Anf.*	*Unpräz. Anf.*	*Präz. Anf.*	*Unpräz. Anf.*	*Präz. Anf.*	*Unpräz. Anf.*
Precision	100 %	76,92 %	100 %	82,83 %	100 %	98,47 %
Recall	79,31 %	100 %	29,12 %	100 %	78,57 %	100 %
F_1-Measure	88,46 %	86,95 %	45,11 %	90,61 %	88 %	99,23 %

Tabelle 6: Precision, Recall und F_1-Measure: *lang*, *kurz*, *schnell*.

Eine Adjunkt-Konstruktion, die ebenfalls nicht korrekt klassifiziert werden konnte, ist in Beispiel (6) gegeben.
(6) *War die Strecke bis zu Position A zu lang, d.h. es wurden mehr als drei Hallimpulse gezählt, wird das Glasdach geschlossen.*

Bei Beispiel (6) handelt es sich um eine linguistisch sehr komplexe Anforderung. Regeln für eine Konstruktion wie diese zu definieren, ist maschinell nur schwer umsetzbar. Man sollte sich hier auch die Frage stellen, ob sich der Aufwand eine Regel zu finden lohnt, oder ob es nicht doch sinnvoll ist, den Autor auf eine so komplizierte Formulierung wie in (6) hinzuweisen, damit er die Anforderung nochmals in eine verständlichere Konstruktion umwandelt.
Tabelle 6 zeigt die Auswertung der Ergebnisse der Weak-Words *lang*, *kurz* und *schnell*. Für die Analyse von *kurz* und *schnell* wurden dieselben Regeln verwendet, die auch bei *lang* zum Tragen kommen. Die Ergebnisse von *lang* und *schnell* sind solide. Bei *kurz* kann man einen deutlichen Einbruch erkennen. Zwar generalisieren die Kontexttypen von *lang* gut auf *schnell* und ordentlich auf *kurz*, aber die drei Wörter werden durchaus unterschiedlich verwendet. Regeln, die für jedes einzelne Wort separat festgelegt werden, könnten also noch bessere Ergebnisse liefern. Auch hier resultieren die Fehlanalysen des Systems vor allem aus Klammer- und Adjunkt-Konstruktionen.

5 Zusammenfassung und Ausblick

Es wurde ein sehr stark auf Kontext-Typen aus Anforderungstexten zugeschnittenes Lexikon von potentiell unspezifischen Wörtern erstellt, welches automatisch anwendbar ist. Das Lexikon enthält Wörter und deren Kontextmuster; die Arbeitsmethodik ist von der korpusbasierten Lexikographie inspiriert, zielt aber auf sehr präzise, automatisch überprüfbare Muster von Kontexttypen. Die entwickelte Weak-Word-Analyse, die dieses Lexikon als Quelle heranzieht, liefert solide und über semantisch verwandte Wörter generalisierbare Ergebnisse.

In weiteren Arbeiten sollen weitere Weak-Words in die Analyse integriert und neue Lexikoneinträge mit zugehörigen Regeln erzeugt werden. Für zukünftige Arbeiten wäre es auch interessant zu untersuchen, welche anderen linguistischen Probleme, neben den Weak-Words, gegen die Qualitätskriterien verstoßen und ob auch diese automatisiert behandelt werden können.

6 References

Bohnet, B. (2009). Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009,* Boulder, Colorado, US.

Daimler AG (2013). Qualitätskriterien, Daimler AG, Böblingen, DE.

Dörre, J., Gerstl, P., & Seiffert, R. (2004). Volltextsuche und Text Mining. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, & H. Langer, Computerlinguistik und Sprachtechnologie: Eine Einführung. München: Spektrum, pp. 479-495.

Evert, S., Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham, UK.

Heidenreich, M. (2010). Metriken und Werkzeugunterstützung zur Überprüfung von Anforderungen. OBJEKTspektrum, Ausgabe RE/2010.

Krisch, J. (2013). Identifikation kritischer Weak-Words aufgrund ihres Satzkontextes in Anforderungsdokumenten. Diplomarbeit. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, DE.

Lami, G. (2005). QuARS: A Tool for Analyzing Requirements. Carnegie-Mellon Software Engineering Institute.

Melchisedech, R. (2000). Verwaltung und Prüfung natürlichsprachlicher Spezifikationen. Dissertation. Universität Stuttgart, Stuttgart, DE.

Rupp, C. (2007). Requirements-Engineering und -Management: professionelle, iterative Anforderungsanalyse für die Praxis. München: Hanser Verlag.

Stöckel, F., Stolz, P., Uddin, I., & Endriss, L. (2009). DESIRe® Dynamic Expert System for Improving Requirements. HOOD Group.

From Term Dynamics to Concept Dynamics: Term Variation and Multidimensionality in the Psychiatric Domain

Pilar León-Araúz, Arianne Reimerink
Department of Translation and Interpreting, University of Granada
pleon@ugr.es, arianne@ugr.es

Abstract

Medical terminology is one of the most dynamic terminological domains, and the choice of one term instead of the other is not random, but the result of different perspectives towards reality. VariMed is a research project on medical term variants and its overall objective is to generate a multifunctional resource on the medical domain for linguistic research, translation and technical writing. In this paper, we propose a systematic way of extracting term variants from large corpora within the subdomain of Psychiatry and how to represent them according to cognitive and communicative parameters. Our aim is to discover if different conceptualizations, or different conceptually motivated term variants, of the same concept are preferred in expert or semi-specialized communication. A corpus on Psychiatry was compiled and classified according to user types. A grammar was designed in NooJ (Silberztein, 2003) in order to extract term variants based on the usual lexico-syntactic patterns accompanying synonyms (*also known as*; *commonly referred to as*, etc.). Corpus analysis results indicate that, from a cognitive perspective, term variants reflect the prototypical dimensions in which psychiatric disorders may be classified. From a communicative perspective, terms and dimensions can also be associated with user-based parameters.

Keywords: term variation; cognitive motivation; communicative motivation

1 Introduction

The medical domain has over 25 centuries of history and involves numerous disciplines which affect all human beings to some extent. Therefore, medical terminology is one of the most dynamic terminological domains, and the use of one term instead of the other implies perceiving and conceptualizing aspects of reality from different perspectives (Prieto Velasco et al 2013: 168). VariMed is a research project on medical term variants and its overall objective is to generate a multifunctional resource on the medical domain for linguistic research, translation and technical writing. In this paper, we propose a systematic way of extracting term variants from large corpora within the subdomain of Psychiatry and how to represent them according to cognitive and communicative parameters. The relationship of specialized communications between a terminological resource and its user implies a

prototypical discursive positioning (Harré and Langenhove 1999), which is reflected in a specialized text sender and receivers with a different background knowledge level. Terminological resources should provide adequate terminological units and an adequate knowledge load (Tarp 2005: 8-9), always according to their potential users' continuum of general-specific language and knowledge purposes (León-Araúz et al 2013: 33).

This is in consonance with the Functional Theory of Lexicography (FTL; Bergenholtz and Tarp 1995, 2003). According to the FTL, there are two main types of lexicographic functions that cover *use situations* and different *user needs* (Wiegand 1989). These functions are cognition and communication-oriented (Bergenholtz and Tarp 2003; Bergenholtz and Nielsen 2006). In cognition-oriented situations, users seek additional information to widen their knowledge about the conceptual structure of a particular subject-field (psychiatry, neurology, oncology, etc.). Bergenholtz and Nielsen (2006: 286) explain that in these situations, the only communicative act taking place is between the terminographer and the users of the resource. The users want knowledge and the lexicographers provide it at a cognitive level, nothing more. The most difficult task is then, for the terminographer to decide *how much* information is to be included and how to represent its underlying structure to make the dictionary suitable to meet users' needs. On the other hand, in communication-oriented situations, two or more persons are engaged in producing or receiving a piece of language. This is the case of a translator who receives and must subsequently produce a text, as well as scientific writers, proofreaders, etc. Here the terminographer acts as a kind of mediator who helps to solve communication problems. We believe any terminological resource should satisfy both (León-Araúz et al 2013: 34).

In section 2, we give a brief overview on term variation. In section 3, we present how term variants are extracted from a specialized corpus on Psychiatry with a pattern-based grammar in NooJ, an NLP application (Silberztein, 2003). In section 4, a selection of the extracted variants is classified according to dimensional features and the results are compared across three subcorpora in order to see if certain dimensions are preferred in one discourse or the other. Finally, section 5 covers the conclusions and further research.

2 Term Variation

Although specialized language initially aspired to having one linguistic designation for each concept for greater precision, it is true that the same concept can often have many different types of linguistic designations. In the same way as in general language, there is terminological variation based on user-based parameters of geographic, temporal or social variation or usage-based parameters of tenor, field, and mode (Gregory and Carroll 1978). However, terminological variation also occurs for reasons that are often considerably more complex and difficult to explain. Freixa (2006: 52) classifies the causes for terminological variation in the following categories: (1) dialectal, caused by different origins of the authors; (2) functional, caused by different communicative registers; (3) discursive, caused by dif-

ferent stylistic and expressive needs of the authors; (4) interlinguistic, caused by contact between languages; (5) cognitive, caused by different conceptualizations and motivations. According to Freixa (2002), certain term variants are not only formally different, but also semantically diverse, as they give a particular vision of the concept. In this sense, Fernández-Silva et al (2011) describe this phenomenon as the linguistic reflection of conceptual *multidimensionality*. Multidimensionality has been defined by many authors (Bowker 1997, Kageura 1997, Wright 1997, Rogers 2004) as the phenomenon in which certain concepts can be classified according to different points of view or conceptual facets. This has important consequences in how domains are categorized and modelled. According to Picht and Draskau (1985, 48 *apud* Rogers 2004, 219), multidimensionality depends on who is the classifier as well as the different knowledge sources that may reflect different criteria when organizing the same domain or knowledge node. For example, botanists would classify roses different from rose growers. However, multidimensionality has also an impact on term variation, since concepts can be designated in more than one way based on the different characteristics that it possesses (Fernández-Silva et al 2011). Thus term variation should not be regarded as a linguistic phenomenon isolated from conceptual representations, since it is one of the manifestations of the dynamicity of categorization and expression of specialized knowledge (Fernández Silva et al in press).

Fernández-Silva (2010: 60-71) classifies the cognitive factors involved in term variation, based on numerous authors, according to two criteria. Firstly, the first category division depends on whether the cognitive factor refers to the conceptual organization or to its usage. Secondly, within the usage category, the categorization of the factor depends on how reality is conceptualized by certain groups of people or individuals or how reality is conceptualized according to the specific context in which the concept is used (Table 1).

Conceptual organization	
	Conceptual system
	Conceptual class
	Multidimensionality of the conceptual system
	Flexibility of the concept
	Linguistic system/culture
Usage	
Different conceptualizations	Knowledge evolution
	Dialects/cultures
	Thematic areas in interdisciplinary contexts
	Schools of knowledge/Ideologies
	Socio-professional groups
	Individual/individual point of view
Adaptation to specific context use	Adaptation to level of expertise of the receiver
	Intention/Aim/Point of view

Table 1: Cognitive factors of term variation (adapted from Fernández-Silva 2010: 61).

As can be inferred by Table 1, all Freixa's causes for term variation can be approached from a cognitive perspective according to Fernández-Silva. This study combines Freixa's second (functional) and fifth (cognitive) causes for term variation with Fernández-Silva's perspective, since it analyses the multidimensionality of the conceptual system and how the different conceptual dimensions correlate with the adaptation to the level of expertise of the receiver. Our aim is to discover if different conceptualizations, or different conceptually motivated term variants, of the same concept, are preferred in expert communication or semi-specialized communication.

3 Extracting Term Variants

A specialized corpus was compiled on the Psychiatry domain, which has more than 10 million tokens, and it was divided according to user and genre types: expert, semi-specialized and encyclopaedic. The expert corpus contains specialized books and journal papers written by experts for experts, such as the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013). The semi-specialized corpus consists of web pages and brochures written by experts from Medline Plus and the National Institute of Mental Health, which combines basic and clinical research with information for patients, or their relatives, suffering from any kind of mental disorder. Finally, the encyclopaedic corpus consists of a Wikipedia dump which was automatically collected through categories such as *Psychiatry*, *Syndromes*, *Disorders*, etc. We considered that Wikipedia should belong to this corpus because, being an encyclopaedic resource, it usually contains metalinguistic information on synonyms and variants that could be useful in our research.

Once the corpus was compiled and classified, a NooJ local grammar was designed in order to extract term variants (Figure 1). The grammar is based on the usual lexico-syntactic patterns accompanying synonyms (*also known as*; *commonly referred to as*, etc.) combined with specialized terms, namely syndromes, disorders and diseases.

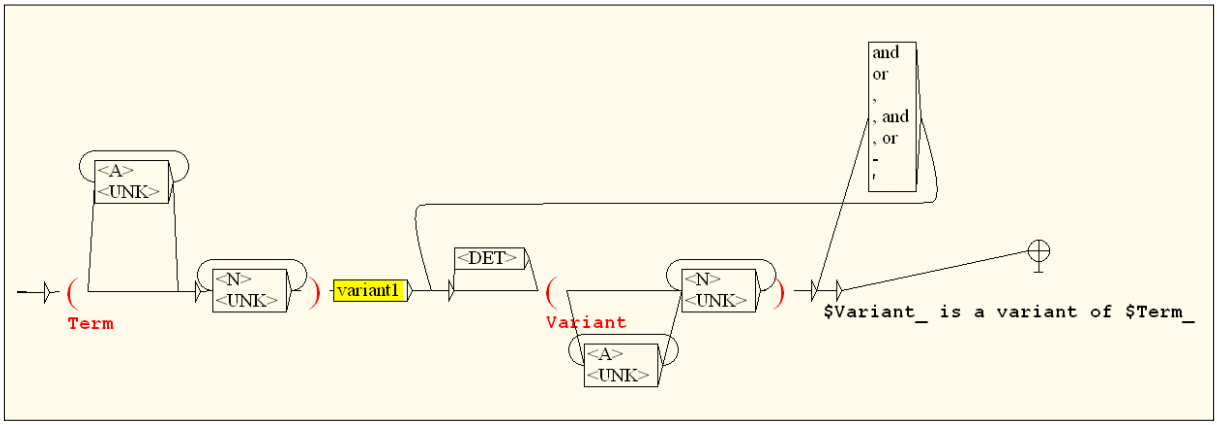


Figure 1: Pattern-based grammar for term variants' extraction.

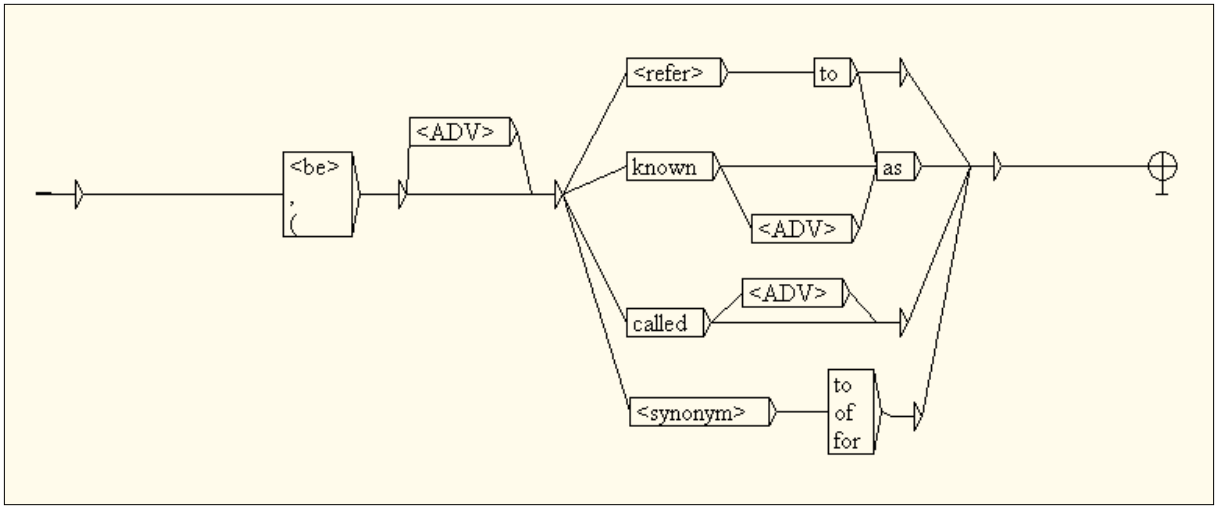


Figure 2: Sub-graph variant1.

Local grammars in NooJ work in conjunction with dictionary-like resources that act like a parser. In this case we used the default general language dictionary in NooJ as a POS tagger. However, since the dictionary does not include highly specialized terms, we had to include the <UNK> code in order to locate the terms that are *unknown* to the system. Thus, the grammar in Figure 1 identifies different sequences where a noun (<N>) or an unknown word (<UNK>), optionally preceded by an adjective (<A>) or another unknown word, are followed by the patterns in the *variant1* sub-graph together with another similar structure. This helps us locate specialized terms on both sides of term variance structures and store them in variables from which we can generate the following output: *x is a variant of y* (Figure 3). This output let us build our own dictionary, where variants are different entries but are linked to the same concept.

Before	Seq.	After
edicineTopic = 939 }}	Delusional parasitosis, also known as Ekbom's syndrome/Ekbom's syndrome is a variant of delusion of parasitosis	, is a type of chronic
ility disorders sidebar}}	Depressive personality disorder (also known as melancholic personality disorder/melancholic personality disorder is a variant of depressive personality disorder	.<ref name="
what is clear is that	depressive pseudodementia, synonymously referred to as depressive dementia/depressive dementia is a variant of depressive pseudodementia	) is a controversial p
bytes="10312">Social-	Emotional Agnosia, also known as emotional agnosia or expressive agnosia/expressive agnosia is a variant of emotional agnosia	(3) or major depressi
ot;. * Individuals with	expressive aphasia (also called Broca's aphasia/Broca's aphasia is a variant of expressive aphasia	, is the inability to pe
aphasias=== AOS and	expressive aphasia (also known as Broca's aphasia/Broca's aphasia is a variant of expressive aphasia	) were once thought
) = D009494	}}	General paresis, also known as general paralysis/general paralysis is a variant of general paresis	) are commonly miste
Social anxiety disorder|	generalized social phobia (also known as social anxiety disorder/social anxiety disorder is a variant of generalized social phobia	of the insane or para
D = D020820	}}	Hemiballismus, sometimes called ballism/ballism is a variant of hemiballismus	or simply social anxi
few, or all of them.	Hypersensitivity is also known as sensory defensiveness/sensory defensiveness is a variant of hypersensitivity	, is a very rare move
}} Hypochondriasis or	hypochondria (sometimes referred to as health phobia or health anxiety/health anxiety is a variant of hypochondria	. Examples of hypers
ccess. A lesser form of	hypomania is called hyperthymia/hyperthymia is a variant of hypomania	) refers to excessive
ID = D014202 }}	Intention tremor, also known as cerebellar tremor/cerebellar tremor is a variant of intention tremor	. ==Causes== While
age to the cerebellum,	intention tremors are often referred to as cerebellar tremors/cerebellar tremor is a variant of intention tremor	, is a dyskinesia|dysk
is being looked into is	Internet addiction (also known as pathological Internet use/pathological internet use is a variant of internet addiction	.<ref name=seeber
		).<ref name=Cao&q

Figure 3: Term variant extraction output.

Furthermore, this grammar has a recursive path in order to identify the cases where different variants are enumerated, as in *...dermatillomania (also known as neurotic excoriation, pathologic skin picking or compulsive picking)*. In this case, three different outputs are produced (*neurotic excoriation* is a variant of *dermatillomania*; *pathologic skin picking* is a variant of *dermatillomania*; *compulsive picking* is a variant of *dermatillomania*). By using the statistical module in NooJ (based on the standard score), we can conclu-

de that, not surprisingly, term variance structures are most often lexicalized in the Wikipedia corpus, then in the semi-specialized corpus, and finally in the expert corpus (Figure 4).

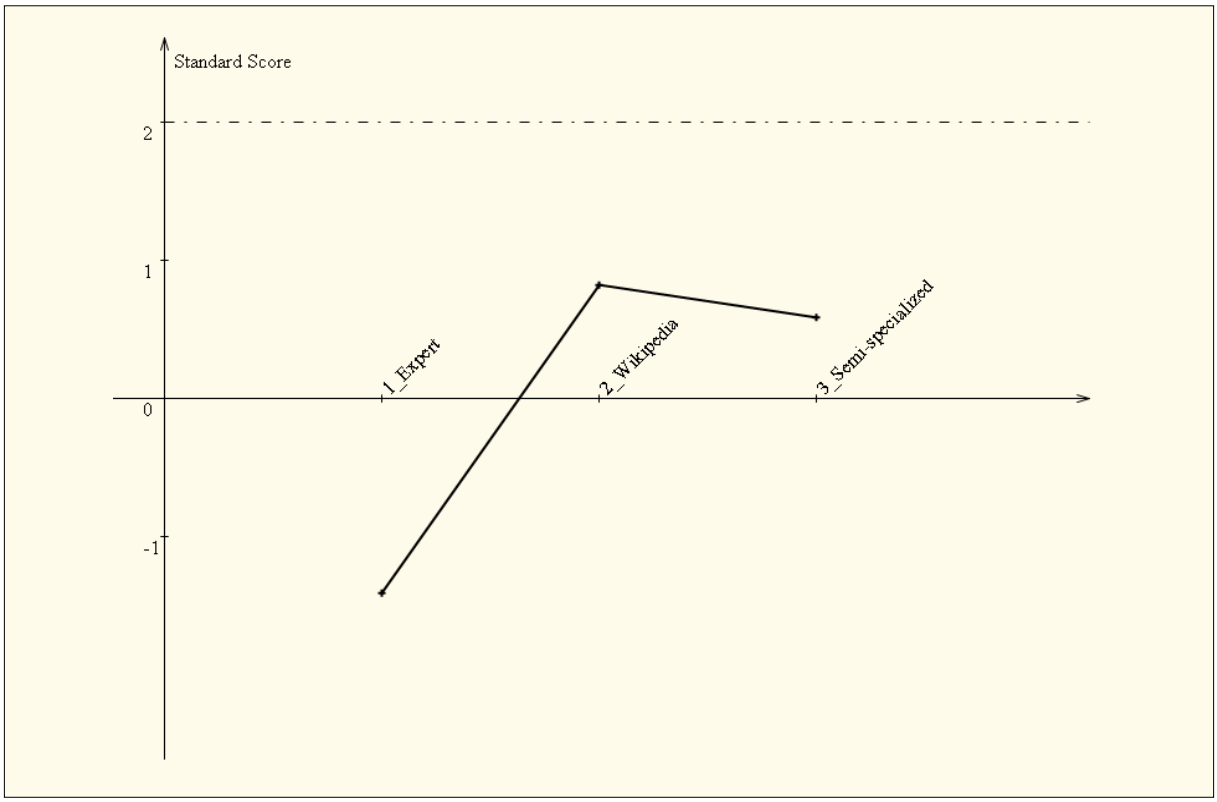


Figure 4: Term variance structures in the three corpora.

4 Representing Term Variants

In terminological resources, users are often confronted with a vast array of variants with no other information on how term variation arises and how their use may be constrained. However, they need to know when to use each of the variants and the conceptual connotations they imply, since this will affect the receiver's interpretation of the message.

In our study, we have found many different types of term variants for the same concept. Some of them were just acronyms, dialectal, orthographic or morphological variants, which, of course, need to be stored in any terminological resource, but their impact on communication is obvious, and their use does not usually need any further explanation. In this paper, however, we focus on dimensional variant types, which need more in-depth study, since they affect both cognitive and communicative situations. Dimensional variants show different conceptualizations of the same concept according to different facets and are usually conveyed by multi-word terms. For instance, *Ganser syndrome*, *nonsense syndrome* and *prison psychosis* are all variants of the same concept, but the first one highlights a DISCO-

VERER dimension (Sigbert Ganser was the first to describe the syndrome), the second one focuses on the SYMPTOM dimension (saying nonsense is one of them) and the third one on the LOCATION dimension (it often takes place in prisons, since it affects inmates). We collected from the corpus all the concepts that showed more than one variant type and classified their corresponding variants according to the dimension conveyed. In Table 2, we show the dimensions we found with an illustrative example.

Dimension	Term variant
+Discoverer	Korsakoff's psychosis
+Symptom	burning-mouth syndrome
+Cause	alcohol-induced amnestic disorder
+Body_part	Broca's aphasia
+Patient	boxer's dementia
+Result	bedwetting
+Intensity	mild cognitive impairment
+Time	short-term insomnia
+Location	prison psychosis

Table 2: Dimensional variant types found in the corpus.

Of course, there are variants that may show different dimensions at the same time, such as *chronic* [+Time] *traumatic* [+Cause] *brain* [+Body_part] *injury associated with boxing* [+Cause] or *alcohol-induced* [+Cause] *amnestic* [+Symptom] *disorder*; and different variants for the same concept that highlight the same dimension, such as *Alice in Wonderland* [+Symptom] *syndrome* and *Lilliputian hallucination* [+Symptom], which refer to the same symptom of the disorder, or *nonsense* [+Symptom] *syndrome* and *.syndrome of approximate answers* [+Symptom], which convey the same dimension but refer two different symptoms.

Terminological resources should add the conceptual dimension conveyed by each variant so that users can make a cognitively sound choice, but as previously stated, term choice also depends on communicative situations, namely, the expert-lay continuum. Therefore, term entries should also add use-related information. When querying the corpus, we see that the variant *dementia pugilistica* is much more often used in the expert corpus as compared to *punchdrunk syndrome* or *boxer's dementia*. In this case, the latin origin of the term points to the usual preference in expert settings. However, this preference is not always as straight forward. In such cases, users should have this information at hand. For instance, in Figure 5 three different variants for the same concept are represented according to usage-based preferences: *postnatal depression* (red), *postpartum depression* (blue), and *baby blues* (black). For this analysis, the three corpora were merged into a single file in order to observe these preferences as a continuum. When a single file is loaded in NooJ, its statistical module automatically splits it into 20 parts. Thus, in Figure 5, the first third of the graph represents the expert corpus, the second the Wi-

kipedia corpus and the third the semi-specialized corpus. Not surprisingly, *baby blues* is the preferred term in semi-specialized communication, although the other two variants are also commonly found. At first, *postnatal* or *postpartum* would seem interchangeable choices, but corpus analysis indicates that the use of one term or the other imposes a strong constraint on the communicative situation. *Postpartum depression* seems to be a more neutral term most often found in the Wikipedia corpus and *postnatal depression* is the preferred term in expert texts.

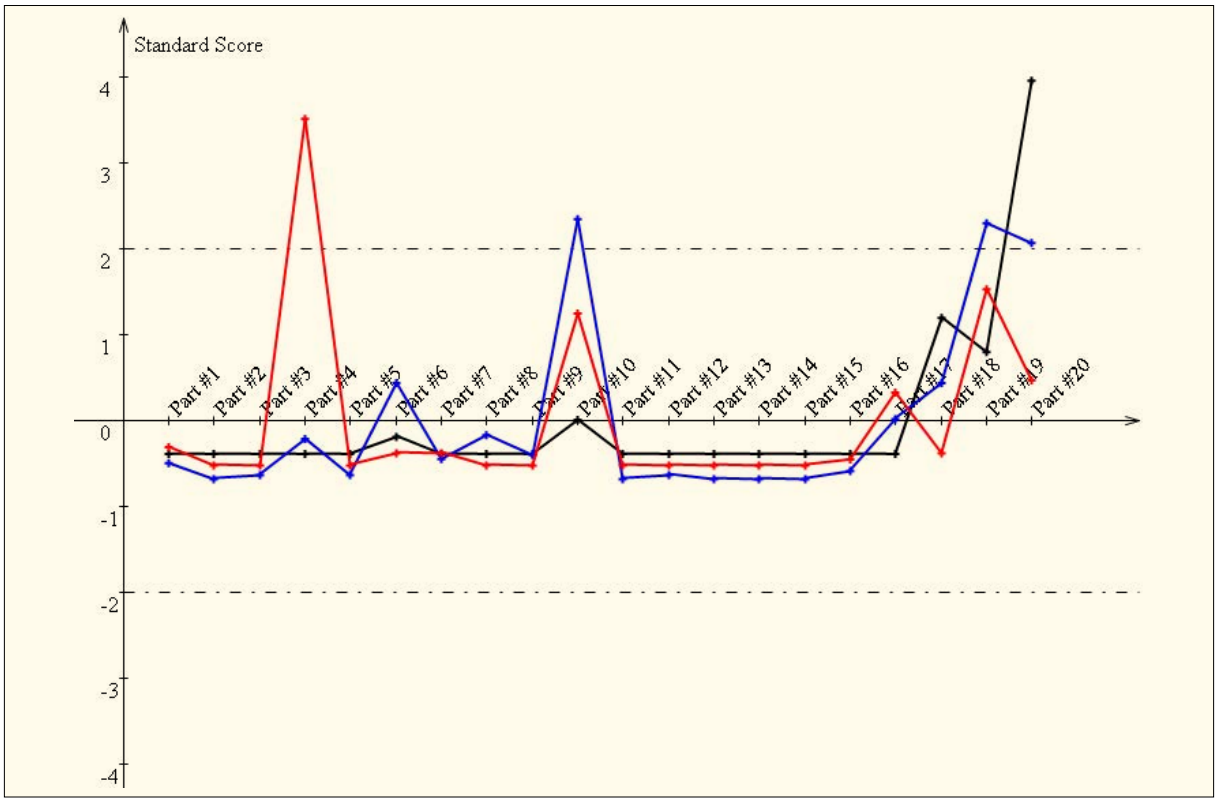


Figure 5: Usage-based preferences for *baby blues* (black); *postpartum depression* (blue); *postnatal depression* (red).

Nevertheless, apart from showing usage information related to individual terms, we believe that more generalized patterns can be found in the comparison of the three corpora with regards to the prevalence of dimensions. For this reason, as a further step, we built a new dictionary with all the dimensional variants related to the same entry in order to automatically annotate them as such in the corpus. So far, the dictionary has 73 concept entries associated with 326 dimensional variants. Figure 6 shows and example of a NooJ dictionary entry, where N means noun, FLX=TABLE refers to the inflectional paradigm, and UNAMB is a special code that lets us deal with multi-word terms as a single unit. Apart from NooJ codes, we can also add any other semantic feature to each entry (e.g. Term, Symptom).

```
ganser syndrome,N+FLX=TABLE+Term+Discoverer+UNAMB
ganser's synsdrome,ganser syndrome,N+FLX=TABL+Term+Discoverer+UNAMB
nonsense syndrome,ganser syndrome,N+FLX=TABLE+Term+Symptom+UNAMB
balderdash syndrome,ganser syndrome,N+FLX=TABLE+Term+Symptom+UNAMB
syndrome of approximate answers,ganser syndrome,N+FLX=TERMOFTERM+Term+Symptom+UNAMB
pseudodementia,ganser syndrome,N+FLX=TABLE+Term+Intensity+Old+UNAMB
prison psychosis,ganser syndrome,N+Term+Location+UNAMB
```

Figure 6: NooJ dictionary entry for *Ganser syndrome* and its variants.

Once we had our own dictionary, we applied it to the main corpus in order to build three new subcorpora, which would only contain sentences where our previously extracted term variants occurred. This was done for a better performance of the system and especially in order to eliminate any bias towards lexical diversity or distribution differences in the three corpora. After that, we performed new queries based on four of the most relevant dimensions found in this domain (Figure 7): +Discoverer (black), +Cause (red), +Symptom (blue), +Body_part (green).

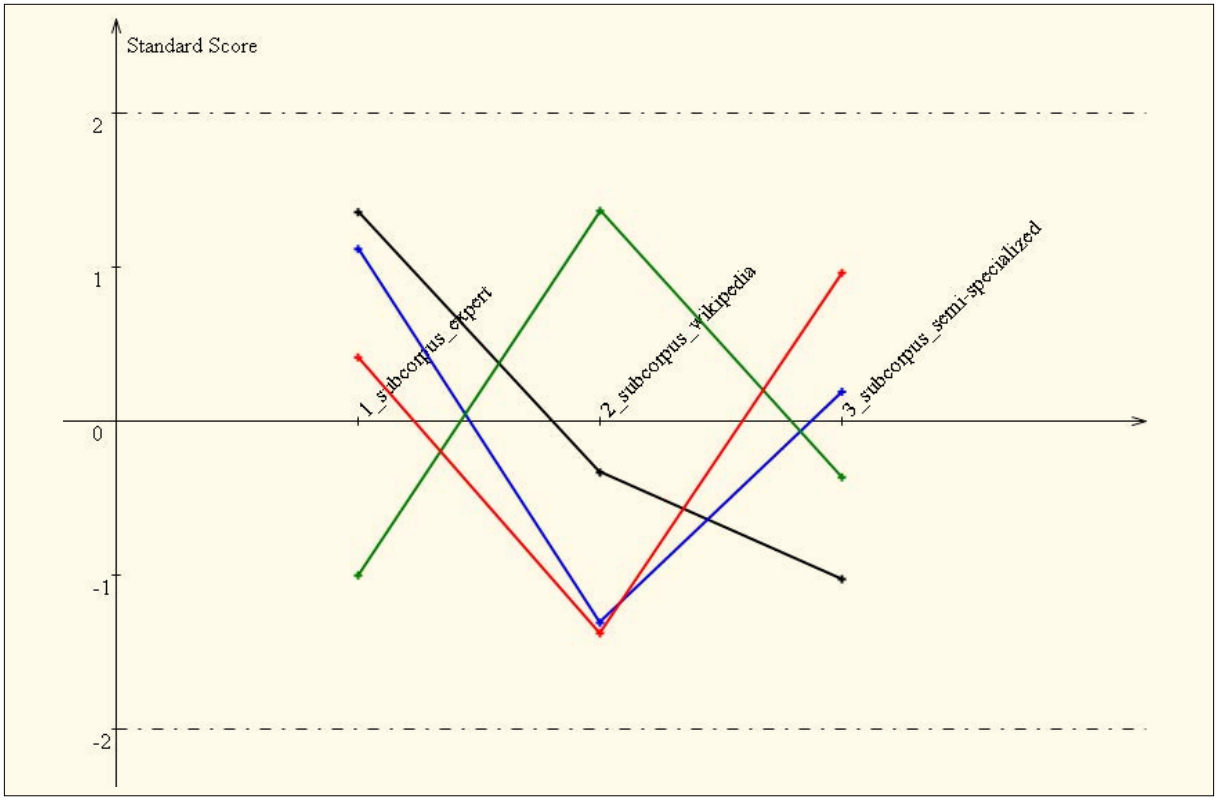


Figure 7: Correlation of dimensional variants with the three corpora.

According to a patient-oriented approach, one might think at first that the +Symptom and +Body_part dimensions could be more semi-specialized than expert-related. However, surprisingly enough, it seems that the +Body_part dimension is only significant in the Wikipedia subcorpus, whereas the +Symptom dimension is most often found in the expert subcorpus. In the expert subcorpus +Discove-

rer is the most prevalent dimension, but +Cause is also significantly represented. As for the semi-specialized subcorpus, +Cause and +Symptom dimensions are most prototypical, but in an inverted way as compared to the expert subcorpus.

This is only a first approach to the study of cognitive and communicative correlation that should be further extended to a higher number of variants and other medical domains, since results can change dramatically when comparing domain-based differences. In this sense, Tsuji and Kageura (1998) observed in a medical corpus that person or virus names were more dominant than other variant types.

5 Conclusion

In this paper we have shown how term variants in the psychiatric domain are cognitive and communicatively motivated. From a cognitive perspective, variants reflect the prototypical dimensions in which psychiatric disorders may be classified. From a communicative perspective, terms and dimensions can also be associated with user-based parameters. However, further studies need to be done with regards to the correlation between the cognitive and communicative factors underlying term variation, especially from a cross-linguistic perspective, since not all cultures conceptualize specialized domains in the same way and nor do they address their audience in the same manner.

6 References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bergenholtz, H. and S. Nielsen. 2006. Subject-Field Components as Integrated Parts of LSP Dictionaries. *Terminology* 12 (2): 281-303.

Bergenholtz, H. and S. Tarp. 2003. Two Opposing Theories: On H. E. Wiegand's Recent Discovery of Lexicographic Functions. *Hermes: Journal of Linguistics* 31: 171-196.

Bergenholtz, H. and S. Tarp (eds.) 1995. Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries. Amsterdam: John Benjamins.

Bowker, L. 1997. Multidimensional Classification of Concepts and Terms. In Wright, S. E. and G. Budin (eds.) *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam: John Benjamins, 133-143.

Cruse, D. A. 2002. Hyponymy and Its Varieties. In Green, R., C. A. Bean and S. H. Myang (eds.) *The Semantics of Relationships*. Dordrecht: Kluwer, 3-21.

Cruse, D. A. 1995. Polysemy and Related Phenomena from a Cognitive Linguistic Viewpoint. In Dizier, P. St. and E. Viegas (eds.) *Computational Lexical Semantics*. Cambridge: Cambridge University Press, 33-39.

Fernández-Silva, S. 2010. Variación Terminológica y Cognición. Factores Cognitivos en la Denominación del Concepto Especializado. PhD Thesis. Universitat Pompeu Fabra.

Fernández-Silva, S., J. Freixa and M. T. Cabré. 2011. A Proposed Method for Analysing the Dynamics of Cognition through Term Variation. *Terminology* 17 (1): 49-73.

Fernández-Silva, S., Freixa, J., Cabré, M.T. (in press). "A proposed method for analysing the dynamics of naming from a intralingual and interlingual perspective". In: Temmerman, R. and Van Campenhoudt, M. (eds.) *The dynamics of culture-bound terminology in monolingual and multilingual communication*. Amsterdam/Philadelphia: John Benjamins.

Freixa, J. 2006. Causes of Denominative Variation in Terminology: A Typology Proposal. *Terminology* 12 (1): 51-77.

Freixa, J. 2002. Reflexiones acerca de las causas de la variación denominativa en terminología. In Guerrero, G. and M. F. Pérez (coords.) *Panorama actual de la terminología*. Granada: Editorial Comares-Interlingua, 107-115.

Gregory, M. and S. Carroll. 1978. *Language and situation: Language varieties and their social contexts.* London: Routledge and Kegan Paul.

Harré, R. and L. van Langenhove (eds.) 1999. *Positioning Theory: Moral Contexts of Intentional Action*. Malden: Blackwell.

Kageura, K. 1997. Multifaceted/Multidimensional Concept Systems. In Wright, S. E. and G. Budin (eds.) *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam: John Benjamins, 119-132.

León Araúz, Pilar, Arianne Reimerink, and Alejandro García Aragón. 2013. Dynamism and Context in Specialized Knowledge. *Terminology* 19 (1): 31-61. doi:10.1075/term.19.1.02leo.

Picht, H. and J. Draskau. 1985. *Terminology: An Introduction*. Guildford: University of Surrey.

Prieto Velasco, Juan Antonio, Maribel Tercedor, and Clara Inés López Rodríguez. 2013. La multidimensionalidad conceptual en la traducción médica. *Skopos. Revista Internacional de Traducción e Interpretación* 2: 167-183.

Rogers, M. 2004. Multidimensionality in Concepts Systems: A Bilingual Textual Perspective. *Terminology* 10 (2): 215-240.

Tarp, S. 2005. The Pedagogical Dimension of the Well-Conceived Specialised Dictionary. *Iberica* 10: 7-21.

Tsuji, K, Kageura, K. 1998. An analysis of medical synonyms: the word-structure of preferred terms. Terminology, 5:2, 229-249

Wiegand, H. E. (ed.) 1989. *Wörterbücher in der Diskussion* [I] (Lexicographica. Sieres Maior 27). Tübingen: M. Niemeyer.

Wright, S. E. 1997. Representation of Concept Systems. *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam: John Benjamins, 89-97.

Acknowledgements

This study was carried out in the framework of the R&D project VARIMED (FFI2011-23120) financed by the Spanish Ministery of Education, Culture and Sports.

Bon usage vs. Fachliches: Fachsprache in der Geschichte der französischen Sprachpflege und Lexikographie

Martina Mayer
Universität Innsbruck
contact@martina-mayer.com

Abstract

Frankreich ist seit dem 16. Jahrhundert dafür bekannt, sich besonders intensiv der Pflege seiner Nationalsprache zu widmen. Seit dem 17. Jahrhundert ist die französische Sprachpflege und Sprachnormierung institutionalisiert und wird durch entsprechende lexikographische Aktivitäten gestützt. Allerdings zielten all diese Bemühungen bis zur Mitte des 20. Jahrhunderts in erster Linie auf die Gemeinsprache ab – ein Umstand, der vergessen lässt, dass auch die Pflege der Fachsprachen in Frankreich eine nähere Betrachtung verdient. Dies bezieht sich nicht nur auf die modernen sprachpflegerischen Bemühungen an den Fachsprachen, sondern hat durchaus ab dem 16. bzw. 17. Jahrhundert Gültigkeit. Der vorliegende Beitrag wird sich dem Aspekt der Pflege der Fachsprachen und ihrer Berücksichtigung in der französischen Lexikographie daher aus einer historischen Perspektive widmen. Der Schwerpunkt liegt dabei vor allem auf entsprechenden Aktivitäten im *Grand Siècle*, das mit seinen Sprachpflegern, dem *bon usage* des *honnête homme*, der Gründung der *Académie française* und seinen lexikographischen Referenzwerken bis in die Gegenwart maßgeblich auf die französische Sprache einwirkte.

Keywords: Fachsprachen; Sprachpolitik; Sprachpflege; Lexikographie; Frankreich;

1 Einleitendes zur Sprachpflege und Sprachnormierung in Frankreich[1]

Die beiden Schlagwörter *Sprachpflege* und *Sprachnormierung* begleiten Europa seit der Renaissance, in der viele europäische Sprechergemeinschaften mit einer intensiven Auseinandersetzung mit ihren noch jungen Nationalsprachen begannen. Das Ziel dieser von humanistischem Gedankengut beeinflussten Aktivität: die Nationalsprachen zunächst in möglichst vielen Domänen gemein- und fachsprachlicher Natur zu effizient einsetzbaren Kommunikationswerkzeugen auszubauen, um ihnen im Weiteren eindeutige Regeln zu verleihen. Das traf auch auf Frankreich zu: Nach dem von der Transiti-

1 Der vorliegende Beitrag beruht zum Teil auf der 2013 veröffentlichten Diplomarbeit *Sprachpflege und Sprachnormierung in Frankreich am Beispiel der Fachsprachen vom 16. Jahrhundert bis in die Gegenwart* (Mayer 2013), die in chronologischer Reihenfolge entsprechende Schlüsselmomente beleuchtet.

on vom Lateinischen zum Französischen sowie entsprechenden Elaborierungsbestrebungen geprägten 16. Jahrhundert begannen die französischen Sprachnormierer und Lexikographen des *Grand Siècle*, ihr sprachpflegerisches Augenmerk vor allem auf die präskriptive Definition des *bon usage* zu richten. Damit wurde im 17. Jahrhundert ein sprachpuristischer Kurs eingeschlagen, mit dem eine strikte Separation von Gemein- und Fachsprache einherging. Es sollten unbedingt Regelhaftigkeit und Reinheit in der Gemeinsprache gewährleistet werden, um einerseits die Schönheit sowie das Prestige des Französischen zu steigern und seine Tauglichkeit als das Konversationsmedium des Hofes sicherzustellen: Der Sprachnutzer bei Hof und dessen Sprachgebrauch standen im Mittelpunkt dieser Bemühungen; der fachlich gebildete *pédant* hingegen, mit seinen spezialisierten Kenntnissen und der entsprechenden Art zu kommunizieren, repräsentierte einen als geistlos geltenden Antityp, der bei Adel und Dichtern kaum Ansehen genoss. Andererseits sollte über den Umweg der sprachlichen Vereinheitlichung bzw. der Ausgrenzung unschicklicher Wörter sowie fachlichen Wortschatzes letztlich auch die politische Stabilität des Landes gefördert werden. Zugleich setzten intensive lexikographische Aktivitäten ein, die im Sinne der sprachpuristischen Standards der offiziellen französischen Sprachpolitik zumeist auf eine Beschneidung des Französischen ausgerichtet waren. Als maßgebliche normierende Instanz sei hier die *Académie française* genannt, der bis in die Gegenwart der Ruf einer altmodischen bzw. konservativen Einrichtung anhaftet. Diese sprachnormativen Bemühungen halten in Frankreich schon seit dem 17. Jahrhundert an. Damit kann die seit über vierhundert Jahren betriebene staatlich koordinierte Sprachpflege mit ihren lexikographischen Bestrebungen als ein französisches Phänomen bezeichnet werden, das eine gewisse Faszination auf die Sprachwissenschaft vor allem anderer Länder auszuüben scheint.

Der Forschungsbereich der Sprachpflege und Sprachnormierung in Frankreich wurde vor allem in Bezug auf die Gemeinsprache und die entsprechenden lexikographischen Referenzwerke umfassend analysiert. Weniger häufig konzentrierte sich die Forschung in diesem Zusammenhang bisher jedoch auf die fachsprachliche Komponente: Auf den ersten Blick mag es gar scheinen, als hätte in Frankreich vor dem 20. Jahrhundert kaum eine strukturierte Fachsprachenreflexion stattgefunden, als seien die Fachsprachen immer nur verdrängt bzw. ignoriert anstatt bewusst ins sprachpflegerische Denken einbezogen worden. Die aktuellen Versuche eines Ausbaus der originär französischen Fachlexik im Kontext einer *crise du français* durch die ministeriellen Terminologiekommissionen beispielsweise könnten in Bezug auf Fachsprache nahezu als einzige Manifestation sprachpflegerischen und –normativen Wirkens in der Geschichte Frankreichs wahrgenommen werden, was allerdings der Realität nicht gerecht würde. Dieser Beitrag wird daher nicht der ohnehin häufig besprochenen Gegenwart, sondern dem historischen Aspekt besondere Aufmerksamkeit widmen: Im Mittelpunkt steht das noch für die heutige französische Sprachpolitik und Lexikographie prägende 17. Jahrhundert mit seinen Sprachnormierern, dem Sprachideal des *honnête homme*, seinen lexikographischen Referenzwerken und der Haltung der *Académie française*. Damit verbunden ist natürlich auch die Frage, wie es überhaupt dazu kommen konnte, dass die offizielle Sprachpolitik im *Grand Siècle* kontinuierlich die

Fachsprachen zurückgedrängt hat. Die Ursache dieses Phänomens ist chronologisch früher anzusetzen.

2 Französische Fachsprachenreflexion vor dem 17. Jahrhundert

Reflektiert man die Entstehung von Fachsprache, so ist von einer Herauskristallisierung auf Basis eines spezialisierten Kommunikationsbedarfs durch menschliche Arbeitsteilung auszugehen. Eine solche bringt unabhängig von der Domäne menschlicher Tätigkeit notwendigerweise Fachkommunikation hervor, sei es in Form einer Werkstatt- oder einer Wissenschaftssprache, sei es auf mündlichem oder schriftlichem Wege. Mündliche Fachkommunikation prägte Europa bis in die frühe Neuzeit; dann stieg die Anzahl schriftlicher Zeugnisse von Fachkommunikation aus den verschiedensten Bereichen. Entsprechende Textbelege weisen sowohl theoretisch-wissenschaftlich orientierte (z.B. Chemie, Medizin, Philosophie, Recht) als auch praktisch-handwerklich ausgerichtete Fachsprachen auf, wobei das kommunikative Ziel jeweils ein fachinterner, also auf das jeweilige Kollektiv begrenzter Austausch bzw. eine entsprechende Konservierung von Inhalten war (Fluck 1996: 27). Interessant ist hierbei die sprachliche Aufteilung: Während sich wissenschaftliche Fachkommunikation nach Tradition der *septem artes liberales*, der Sieben Freien Künste, damals natürlich noch in erster Linie in lateinischer Sprache vollzog, lässt sich für die Fachkommunikation der *artes mechanicae*, also der Handwerklichen Künste, schon von Anfang an eine Konzentration auf die Volkssprachen feststellen, was auch verständlich ist, mussten die fachsprachlichen Inhalte doch der richtigen Zielgruppe zugänglich sein – in Bezug auf Werkstattsprachen nun eben dem „einfachen Volk", sofern es überhaupt lesen konnte (Berschin/Felixberger/Goebl 2008 198; Roelcke 2005: 162ff).

Mit dem 16. Jahrhundert erfuhr nun Frankreich im Bereich der Sprachpolitik einen ersten entscheidenden Umbruch bezüglich dieser Zielgruppenorientierung bzw. Vulgarisierung: Mit der *Ordonnance de Villers-Cotterêts* vom 25. August 1539 wurde der *„langage maternel françois"* durch François Ier als offizielle Gerichts-, Urkunden- und damit Amtssprache festgesetzt, wodurch der breiten Bevölkerung ganz neue Inhalte zugänglich gemacht wurden.[2] Damit war der sich in den verschiedensten fachlichen Domänen fortan zügig vollziehende Ersatz des Lateinischen durch das Französische besiegelt (Müller 1975: 36). Pécheur (2001: 51) spricht hier bildhaft von einem „[...] embryon de politique linguistique [...]", dessen Ziel eine Festschreibung des verpflichtenden Gebrauchs der französischen Sprache im Sinne einer erstmaligen Normierung als Standard war. Im Folgenden wurde Joachim du Bellay (1522-1560) zum ersten Verfechter der französischen Sprachpflege. Der Titel seiner 1549 erschienenen *Deffen-*

2 Die *Ordonnance de Villers-Cotterêts* umfasst 192 Artikel, von denen vor allem die Artikel 110 und 111 maßgeblichen Einfluss auf die Entwicklung des Französischen hatten. Nachzulesen sind sie im vollständigen Text der Verordnung auf der Website der *Assemblée Nationale* (http://www.assemblee-nationale.fr/histoire/villers-cotterets.asp, Stand: 5.4.2014).

ce et Illustration de la Langue Francoyse ist programmatisch: Das Französische sei dem Lateinischen gegenüber von der Basis her gleichwertig und müsse gegen den Vorwurf eines mangelnden Kommunikationspotenzials verteidigt werden – eine entsprechende Pflege im Sinne eines umfassenden Sprachausbaus sowohl der Literatur- als auch der Fachsprache sei allerdings notwendig und käme den Sprachnutzern selber zu (du Bellay 1969: 47f). Zu du Bellays Zukunftszielen zählte außerdem eine kontinuierliche Etablierung des Französischen in weiteren, über die Fachsprache des Rechts hinausgehenden spezialisierten Kontexten der Wissenschaften und des Handwerks, sodass über kurz oder lang alle Inhalte allen Sprachnutzern gegenüber demokratisiert werden sollten (du Bellay 1969:133ff). Du Bellays Vision einer üppigen französischen Lexik der vielseitigen kommunikativen Möglichkeiten wurde Realität; außerdem hatte das 16. Jahrhundert bereits eine ganze Reihe von Glossaren und frühen Wörterbüchern zu bieten, wie Lindemanns (1994) umfassende Bestandsaufnahme *Die französischen Wörterbücher von den Anfängen bis 1600* eindrucksvoll zeigt. Allerdings hatte es die französische Sprechergemeinschaft durch den sehr freizügigen Sprachausbau der Renaissance selbst trotz der bereits bestehenden lexikographischen Aktivitäten in Folge mit einem fast überbordenden und uneinheitlich verwendeten Wortschatz zu tun.

3 Der Sprachpurismus des *Grand Siècle*

Dieser nicht unproblematischen sprachlichen Situation gegenüber fanden sich nun die Sprachnutzer wie auch die Sprachnormierer des *Grand Siècle* wieder. Vor allem Letztere reagierten darauf mit einer puristischen Grundhaltung sowie stark normativen Bestrebungen, was der Stellung der Fachsprachen innerhalb der französischen Sprachpflege und Sprachpolitik nicht unbedingt zuträglich war, zumindest was die offizielle Lexikographie betraf. Dies erklärt sich allerdings schlichtweg aus den sprachpolitischen Bedürfnissen eines zunehmend zentralistisch organisierten Staates, der eine einheitliche, universell einsetzbare und starke Sprache benötigte, um sich von innen heraus zu stabilisieren.

3.1 François de Malherbe

Die wohl erste und eine der wichtigsten unter diesen normierend eingreifenden Persönlichkeiten war der Hofdichter François de Malherbe (1555-1628) mit seiner Suche nach sprachlicher Reinheit und seiner Kategorisierung der Wörter in drei Gruppen: Er klassifizierte zunächst die schicklichen *mots nobles*, weiters die aus diversen Gründen (z.B. aufgrund eines Konnexes zur Welt der Arbeit) zumindest für den Adel unverwendbaren *mots bas* und schließlich die tabuisierten, da auf Körperfunktionen oder den Bereich der Sexualität bezogenen *mots sales*. Damit schloss er im Grunde sämtliche fachsprachliche Begriffe, die großteils natürlich als *mots bas* zu betrachten waren, von einem literarischen Gebrauch aus, erklärte sie als nicht salonfähig und verbannte sie damit kollektiv aus dem Sprachgebrauch der führenden Schicht wie auch der Autoren, worin Brunot (1930: 3-9) ganz eindeutig einen Rückschritt für die französische Sprache

erkannte: Schließlich gingen dadurch in dieser Zeit, die auch noch nicht von umfassenden lexikographischen Bemühungen durchdrungen war, zahlreiche Wörter verloren, die nirgends verzeichnet waren und somit tatsächlich aus dem Französischen verschwanden. Eines ist Malherbe diesbezüglich allerdings nicht vorzuwerfen: Er versuchte zwar, die Literatursprache und die Kommunikation in nicht-fachsprachlichen Kontexten von Terminologie und anderen „Störfaktoren" der Lexik zu reinigen, verfolgte aber nicht das Ziel, normativ in den Bereich der Fachkommunikation selber vorzudringen.

Die von Malherbes Bestrebungen direkt angesprochenen Gruppen, die *noblesse* und das literarische Milieu, akzeptierten jedenfalls seine linguistischen Entscheidungen großteils – bzw. hätte es anders betrachtet auch gar keine Möglichkeit gegeben, in irgendeiner Weise dagegen vorzugehen: Malherbes normatives Wirken war schließlich durch den König gedeckt und war somit stillschweigend zu akzeptieren. Inwiefern? Während es dem Landadel ohnehin an Mitteln und an Einfluss fehlte, so war der Hochadel durch den absolutistischen Willen des Königs an ein starres Gerüst aus Regeln gebunden, welches das Leben bei Hofe dominierte und als obligatorisch zu betrachten war; bei Missachtung drohte laut Elias (1983: 353f) eine „Einschränkung oder Verminderung der königlichen Gunst" und damit „eine schwer zu ertragende Gefahr", die ganz präzise in einem Verlust an sozialem Prestige und folglich einer Verschlechterung der Zukunftschancen, der Lebensverhältnisse usw. bestand. Das Konzept des höfischen Menschen sah es nicht vor, dass ein *honnête homme* jemals Unabhängigkeit vom Hof erlangen sollte: Er musste dem König unterworfen bleiben, damit dessen absolute Macht keinesfalls angetastet werden konnte. Diese Unterwerfung des Adels wiederum konnte nur solange funktionieren, als dieser keine Möglichkeit hatte, sein Leben dank der Anwendung etwaiger nicht ausschließlich höfischer Kompetenzen oder Fachkenntnisse in finanzieller Hinsicht eigenständig zu bestreiten. Während also die fachliche und damit auch finanzielle Unselbständigkeit des *honnête homme* als grundlegende Bedingung des Erhalts der königlichen Macht bezeichnet werden kann, war für den Adel umgekehrt der Erhalt der königlichen Gunst durch eine demonstrative Anpassung an das Regelkorsett des Hofes die grundlegende Bedingung für den Fortbestand seines Lebensstils. Ein guter Grund für jeden *honnête homme*, zum eigenen Schutz nicht einmal verbal auf das Vorhandensein etwaiger spezialisierter Kenntnisse hinzuweisen und sich präventiv dem präskriptiven Sprachgebrauch des *bon usage* zu beugen, der vollkommen auf die mündliche Konversation am Hof bzw. in den Pariser Salons ausgerichtet war, das gesellschaftliche und zugleich schöngeistige Leben durchdrang und alles Fachliche ausgrenzte (Grimm 2005: 145; Hassler 1998: 323f). Da sich die Malherbesche Beschneidung der Sprache in ausgerechnet diesem historischen Kontext situierte, der stark von sozialen Gegebenheiten und Machtverhältnissen determiniert und vom Einsetzen der französischen Klassik durchdrungen war, fiel sie auf fruchtbaren Boden und fand eine entsprechende Nachfolge.

3.2 Claude Favre de Vaugelas

Claude Favre, baron de Pérouges et seigneur de Vaugelas (1585-1650), der sich nicht nur als Mitglied der *Académie française*, sondern auch individuell sprachnormativ betätigte, stellte sein Wirken eben-

falls in das Zeichen des normativen Purismus, wenngleich unter anderen Vorzeichen als Malherbe. Im Rahmen seiner *Remarques sur la langue française, utiles à ceux qui veulent bien parler et bien écrire* aus dem Jahr 1647, die man unter Rückgriff auf die Kategorien der heutigen französischen Lexikographie am ehesten mit einem *dictionnaire des difficultés* vergleichen kann, prägte Vaugelas (1880: 12f) die Begriffe des *bon usage* sowie des *mauvais usage* erstmals in einer schriftlichen Definition, die keinen Zweifel darüber ließ, dass *bon* und *mauvais* für ihn von einem elitären Sprachverständnis und sozialer Exklusion ausgehend zu deuten waren. Wie bei Malherbe stand auch für Vaugelas eine Politik der Ab- und Ausgrenzung im Mittelpunkt, allerdings war diese anders ausgerichtet: Während Malherbe die Reinheit und Klarheit der Sprache in ganz bestimmten Bereichen sprachlicher Äußerung gewährleisten wollte, so ging es Vaugelas darum, den sprachlich-modellhaften Charakter einer Elite von Sprachnutzern als Optimum darzustellen. Vaugelas konnte sich darin natürlich auf das Resultat von Malherbes Vorarbeit stützen – den bereits reinen und schönen Sprachgebrauch des Hofes, dem von nun an unumstrittene Vorbildhaftigkeit zugeschrieben wurde und der bereits ohne fachsprachliche Elemente auskam, allerdings bis auf eine große Ausnahme: den höfisch-fachlichen Wortschatz, der sich auf den typischen Zeitvertreib des *honnête homme* und das Leben bei Hofe bezog.

Dementsprechend finden sich in den *Remarques* nur vereinzelt fachsprachliche Elemente. Im Allgemeinen zögerte Vaugelas kaum, sich in Hinblick auf den Gebrauch einzelner Wörter – auch in Hinblick auf Fachsprache – sprachnormativ ganz eindeutig zu positionieren, um Vorgaben durchzusetzen, die er vom Sprachgebrauch der *„plus saine partie de la Cour"* (Vaugelas 1880:12f) ableitete: beispielsweise dann, wenn Vaugelas (1880: 144) dem auch heute noch verwendeten Begriff *naviguer* aus dem Technolekt der Seeleute zu Gunsten der vom Hof präferierten phonetisch-orthographischen Variante *naviger* ohne objektive Begründung seine Richtigkeit abspricht; wenn er sich ganz offensichtlich über sämtliche etymologische Grundlagen hinwegsetzt, indem er aus *tempe*, der Schläfe, ohne entsprechende Argumentation *temple* macht (Vaugelas 1880: 266) oder seiner sonstigen Haltung gegenüber völlig konträr den fachsprachlichen Neologismus (für Vaugelas' übliche Maßstäbe eigentlich zwei Ablehnungskriterien in einem) *conjoncture* in seine *Remarques* aufnimmt und sich diesem Begriff gegenüber noch dazu höchst positiv gestimmt zeigt (Vaugelas 1880: 345). Vaugelas' Arbeit diente wiederum als Grundlage weiterer normativer Ansätze.

3.3 Dominique Bouhours

Auch Dominique Bouhours (1628-1702) war sprachnormativ tätig, entfernte sich zunächst in seiner schriftlichen Befassung mit dem Thema allerdings von der durch Vaugelas im weitesten Sinne vorgegebenen Form und bediente sich in *Entretiens d'Ariste et Eugène* aus dem Jahr 1671 des Lehrdialogs, um zwei sprachpflegerische Grundhaltungen einander gegenüber zu stellen: einerseits das Bewusstsein, dass der Bedarf einer Emanzipation des Französischen anderen Sprachen gegenüber (Spanisch, Italienisch und Latein) bestünde, die durch Perfektion erreichbar sei, und andererseits den Gedanken, dass bei aller Hinwendung zu ebendieser Perfektion aber nicht die kommunikativen Bedürfnisse der Spra-

che ignoriert werden dürfen; fachsprachlicher Wortschatz sei demnach keinesfalls zu vernachlässigen – schon allein deshalb, weil sogar die für das Leben bei Hof wichtigen Sprachen der Jagd, der Falknerei oder der Kunst auch nicht ohne Fachlichkeit auskommen konnten (Bouhours 1962: 47ff). Bouhours bewies damit also durchaus ein realistisches Gespür für die Notwendigkeiten einer lebenden Sprache, was sich auch an seinen sprachnormativen Folgewerken feststellen lässt: Die *Remarques nouvelles sur la Langue françoise* aus dem Jahr 1675 wie auch die *Suite des Remarques nouvelles sur la Langue françoise* von 1687 folgten sowohl dem Titel nach als auch von der Struktur her betrachtet Vaugelas' direktem Vorbild, beriefen sich auch auf den *bon usage*, konzentrierten sich aber unter anderem mit Rückgriff auf Autorenzitate vermehrt auf die Sprache der schöngeistigen Literatur und hatten nicht den Anspruch, eine elitäre Sprachauffassung unter Abwertung bestimmter sozialer Gruppen bzw. unter absolutem Ausschluss der von diesen Gruppen verwendeten Technolekte zu propagieren.

3.4 Die Académie française

Der 1635 gegründeten *Académie française* hingegen wurde und wird immer wieder unterstellt, sie verfolge ein derartiges Ziel, habe sich in ihrem sprachnormativen und lexikographischen Wirken der systematischen Bekämpfung allen fachlichen Wortschatzes verschrieben und vertrete damit eine Grundhaltung, die einer modernen Sprache gegenüber nicht angemessen sei. Nun mag es zwar stimmen, dass die *Académie française* seit der Definition des guten Sprachgebrauchs des *honnête homme* durch eine entsprechende Gestaltung ihrer Wörterbücher fachsprachliche Elemente konsequent aus der französischen Sprache zurückdrängt; jedoch wäre es wünschenswert, hier zu einem etwas differenzierteren Bild zu finden. Einen ersten Ansatzpunkt dafür bietet ein Blick in den Artikel XXIV der am 22. Februar 1635 beschlossenen Gründungsstatuten der Akademie:

> La principale fonction de l'Académie sera de travailler avec tout le soin et toute la diligence possibles à donner des règles certaines à notre langue et à la rendre pure, éloquente et capable de traiter les arts et les sciences[1].
>
> [1] Article essentiel qui formule la raison d'être de l'Académie, lui prescrit sa mission et fonde son autorité.
>
> (Académie française s.a.: sites/academie-francaise.fr/files/statuts_af.pdf, Hervorhebung des Absatzes weggelassen, MM)

Obgleich eine intensive Pflege der Fachsprachen aus den im Kapitel 3.1 bereits dargelegten machtpolitischen Gründen nicht wirklich mit dem höfischen Sprachideal zu vereinbaren und die *Académie française* diesbezüglich ganz sicher Kompromissen unterworfen war, stellen also zumindest die Gründungsdokumente der Akademie eine gewisse Sensibilität für die kommunikativen Anforderungen unter Beweis, die sich einer in Entwicklung befindlichen Sprache und Sprechergemeinschaft stellen. Eine völlige Abkoppelung fachsprachlicher Inhalte war nie von Grund auf intendiert – ganz im Ge-

genteil, was die Statuten auch beweisen –, ergab sich aber wohl in Folge aus dem von Vaugelas postulierten Sprachideal. Das 1694 erstmalig erschienene Akademiewörterbuch, der *Dictionnaire de l'Académie française*, erwies sich dementsprechend auch als rein gemeinsprachlich ausgerichtetes lexikographisches Werk, was der Akademie allerdings kaum vorzuwerfen ist: Schließlich steht für ein Wörterbuch wie für jedes andere auf ein gewisses Ziel ausgerichtete Werk der Rezipient, oder in diesem Falle präziser der Nutzer, im Fokus – und die Zielgruppe bestand in diesem Fall nun einmal aus dem Adel und den Dichtern jener Zeit, für die ein klar prädefinierter Normgebrauch galt, dem der *Dictionnaire* wiederum gerecht werden musste. Mit dieser erstmaligen, zumindest aus nachvollziehbaren Gründen den Fachsprachen gegenüber restriktiven Positionierung ihres lexikographischen Wirkens stellte nun die *Académie française* zwar die Weichen für ihre gesamte weitere Redaktion von Wörterbüchern, blieb allerdings im Lauf der Zeit in Bezug auf ihre Konzeption der Sprachpflege bei den Fachsprachen auch nicht so vollkommen statisch wie es ihr regelmäßig vorgeworfen wird. Dies lässt sich anhand einer diachronischen Analyse der Paratexte der Akademiewörterbücher nachvollziehen.
Das Vorwort der ersten Ausgabe des Akademiewörterbuchs von 1694 zeugt ganz klar von der Zielgruppenorientierung des Werkes (wenn im folgenden Zitat vom *Discours* die Rede ist, so bezieht sich dies natürlich exklusiv auf jenen der *honnestes gens*) und außerdem vom Anspruch, als ein Medium des kollektiven linguistischen Gedächtnisses die französische Sprache in ihrem Istzustand zu illustrieren:

> [...] C'est dans cet estat où la Langue Françoise se trouve aujourd'huy qu'a esté composé ce Dictionnaire; & pour la representer dans ce mesme estat, l'Académie a jugé qu'elle ne devoit pas y mettre les vieux mots qui sont entierement hors d'usage, ni les termes des Arts & des Sciences qui entrent rarement dans le Discours; Elle s'est retranchée à la Langue commune, telle qu'elle est dans le commerce ordinaire des honnestes gens, & telle que les Orateurs & les Poëtes l'employent; Ce qui comprend tout ce qui peut servir à la Noblesse & à l'Elegance du discours. Elle a donné la Definition de tous les mots communs de la Langue dont les Idées sont fort simples; & cela est beaucoup plus mal-aisé que de definir les mots des Arts & des Sciences dont les Idées sont fort composées; Car il est bien plus aisé, par exemple, de definir le mot de Telescope, qui est une Lunette à voir de loin, que de definir le mot de voir; Et l'on esprouve mesme en definissant ces termes des Arts & des Sciences, que la Definition est tousjours plus claire que la chose definie; au lieu qu'en definissant les termes communs, la chose definie est tousjours plus claire que la Definition. [...] (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 1re préface)

Vor dem Hintergrund, dass die Gemeinsprache der Fachsprache als ein Subsystem ihrer selbst die notwendigen grammatischen und lexikalischen Rahmenbedingungen vorgibt, ist auch der Verweis auf definitorische Schwierigkeiten bei gemeinsprachlichen Lemmata interessant, deren lexikographische Erfassung laut dem obigen Auszug eine größere Leistung darstellt als jene von Terminologie: Fachsprache kann ohne Gemeinsprache nicht existieren; Gemeinsprache hingegen wird durch Fach-

sprache maximal durch die Übernahme neuen Wortgutes aus der fachlichen Domäne in die gemeinsprachliche beeinflusst. Dies hat auch die *Académie française* indirekt berücksichtigt:

> L'Académie en bannissant de son Dictionnaire les termes des Arts & des Sciences, n'a pas creu devoir estendre cette exclusion jusques sur ceux qui sont devenus fort communs, ou qui ayant passé dans le discours ordinaire, ont formé des façons de parler figurées; comme celles-cy, Je luy ay porté une botte franche. Ce jeune homme a pris l'Essor, qui sont façons de parler tirées, l'une de l'Art de l'Escrime, l'autre de la Fauconnerie. On en a usé de mesme à l'esgard des autres Arts & de quelques expressions tant du style Dogmatique, que de la Pratique du Palais ou des Finances, parce qu'elles entrent quelquefois dans la conversation. (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 1re préface)

Das Vorwort der zweiten Auflage des Wörterbuchs von 1718 zeigte ein ähnliches Bild (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 2e préface); mit der dritten Auflage ab 1740 hingegen begann die *Académie française* auf eine zusätzliche Öffnung den Fachsprachen gegenüber hinzuweisen, die mit dem nun schon sehr weit fortgeschrittenen Ersatz des Lateinischen durch das Französische in Zusammenhang stand: Die französische Sprache hatte zahlreiche neue Fachbereiche erobert, und entsprechendes Vokabular war von diesen Fachsprachen aus in die Gemeinsprache übergegangen – ein Phänomen, dem auch die ausdrücklich gemeinsprachlich ausgerichtete Lexikographie gerecht werden musste (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 3e préface). Die vierte Auflage von 1762 bekräftigte dies und räumte einen weiteren Aspekt ein: Da die schöngeistige Literatur inzwischen auch wissenschaftliche und fachliche Themen behandle, sei der *honnête homme* in seinem Zeitvertreib zunehmend mit entsprechender Lexik konfrontiert. Im Sinne einer Ausrichtung des Werkzeugs Wörterbuch auf die Bedürfnisse des Nutzers schien es damit wiederum geboten, zusätzliche fachsprachliche Elemente in den *Dictionnaire* zu übernehmen. Diese Aussage zeigt doch, dass die *Académie française* dem tatsächlichen Sprachgebrauch gegenüber nicht vollkommen unsensibel war; weiters zeugt sie von einem bedarfsbedingten Konnex zwischen lexikographischen Bestrebungen und schöngeistiger Literatur (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 4e préface). Die Préface der fünften Auflage von 1789 brachte diesbezüglich wieder keine Neuerungen ein (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 5e préface); die sechste Auflage hingegen bewies mit ihrem Vorwort, dass die Akademie durchaus auch zu Selbstreflexion in der Lage war (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 6e préface): 1835 stellte die *Académie française* fest, dass ihr *Dictionnaire* nur über ein sehr begrenztes fachsprachliches Repertoire verfügte und dass dies ein Versäumnis sei – allerdings ein zu entschuldigendes, seien doch Nomenklaturen ohnehin schnell veraltet und habe nur die Literatursprache wirklich Bestand. Demnach wäre es sinnlos gewesen, Energie auf die Integration fachlicher Lemmata zu verschwenden. Zugleich sei die den Fachsprachen gegenüber restriktive Haltung aber doch zu stark beibehalten worden, außer in Bereichen wie Wappenkunde und Jagdwesen, die dem *Dictionnaire* einzelne Termini wie auch stehende Wendungen beschert hätten. Mit der siebten Auflage des Wörterbuchs von 1878 zeigte die

Akademie sich schließlich den Fachsprachen gegenüber noch offener: 2200 neue Lemmata seien zum Wörterbuch hinzugekommen, wobei auch den Fachsprachen ein gewisser Raum zugestanden worden sei. Vor allem Felder wie die Philosophie, die Archäologie, die Philologie, die Politikwissenschaften, die Industrie und die Landwirtschaft seien nun berücksichtigt worden, wobei diese lexikographische Leistung aber nicht ohne Vorbehalte der eigenen Haltungsänderung gegenüber zustande gekommen sei:

> Naturellement la part des sciences et des inventions nouvelles a été grande dans les deux mille mots ajoutés. Les chemins de fer, la navigation à vapeur, le télégraphe électrique ont fait irruption dans notre bon vieux français, avec leurs dénomination [sic] d'une forme souvent bizarre ou étrangère; force a été d'admettre: *un télégramme*, *un steamer*, *un tunnel*, *des tramways*: l'ombre de nos prédécesseurs a dû plus d'une fois en frémir. (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 7e préface).

Das Vorwort der achten Auflage von 1935 unterstrich dies, indem es zwar über die Aufnahme zahlreicher spezialisierter Lemmata berichtete, sich allerdings negativ zu den Fachsprachen äußerte, die sich kontinuierlich und maßlos ausbreiten würden. Die neu aufgenommenen Termini seien daher sehr bewusst ausgewählt worden, als Kriterium habe die vermutete Dauer ihres Bestehens gegolten. Außerdem sei es natürlich nach wie vor Aufgabe der Akademie, ein Wörterbuch des *bon usage* zu redigieren (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 8e préface). Die neunte Auflage des Akademiewörterbuches ist noch im Entstehen begriffen, liegt derzeit bis zum Eintrag *réglage* vor und umfasst im Vergleich zur vorherigen Auflage etwa 10 000 zusätzliche Lemmata (Académie française s.a.: Le Dictionnaire – La 9e édition). Die Aufnahmekriterien sind, laut eigenen Angaben, wieder höchst pragmatisch angelegt; sie orientieren sich erneut daran, welche fachsprachlichen Elemente bereits in die Gemeinsprache übergegangen sind:

> Nous ne donnons entrée, parmi les termes techniques, qu'à ceux qui, du langage du spécialiste, sont passés par nécessité dans le langage courant, et peuvent donc être tenus pour réellement usuels. (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 9e préface)

In Summe zeigt sich also, dass die Akademie zwar einerseits an einem konservativen Konzept von Sprachpflege festhält, dass sie sich andererseits aber auch nicht vollständig der Notwendigkeit einer Aktualisierung ihres "allgemeingültigen" Wortschatzes verschließt. Eine Institution wie die *Académie française* kann sich letztlich nicht über die Jahrhunderte hinweg völlig wirklichkeitsfremd zeigen, indem sie einen tatsächlichen Kommunikationsbedarf ignoriert.

3.5 Die Fachwörterbücher des 17. Jahrhunderts

Dennoch fand Fachlexik damals zunächst in lexikographischen Werken Beachtung, die nicht auf Initiative der staatlichen Sprachpolitik zurückgingen. Zu nennen ist beispielsweise César-Pierre Richelets *Dictionnaire François, contenant les Mots et les Choses* von 1680 und außerdem Antoine Furetières *Dic-*

tionnaire universel, contenant généralement tous les Mots François von 1690. Vor allem das lexikographische Werk des Letzteren wurde von der *Académie française* als Konkurrenz wahrgenommen und löste die sogenannte *Bataille des dictionnaires* aus, die schließlich indirekt in der Redaktion eines dritten großen fachsprachlichen Wörterbuches des 17. Jahrhunderts gipfelte, des *Dictionnaire des Arts et des Sciences* von Thomas Corneille.

Während die Akademie sich darauf konzentrierte, 1694 in ihrem ersten Wörterbuch den tatsächlichen Sprachgebrauch ihrer primären Zielgruppe abzubilden, veröffentlichte César-Pierre Richelet schon 14 Jahre vorher ein zweibändiges Fachwörterbuch mit dem Titel *Dictionnaire François, contenant les Mots et les Choses : Plusieurs nouvelles Remarques sur la Langue Françoise : Ses Expressions Propres, Figurées & Burlesques, la Prononciation des Mots les plus difficiles, le Genre des Noms, le Regime des Verbes : Avec les Termes les plus connus des Arts & des Sciences. Le tout tiré de l'Usage et des bons Auteurs de la Langue françoise.* Dieses lexikographische Werk war vom *bon usage* losgelöst, obgleich es derselben Zielgruppe gewidmet war wie die Akademiewörterbücher. Richelet verzeichnete darin auch *mots sales* bzw. *mots bas* und legte ein besonderes Augenmerk auf die Fachlexik, dies laut eigener Auskunft im Vorwort, um der Zielgruppe ein noch nützlicheres Hilfsmittel an die Hand zu geben (Richelet 1680: Avertissement, ohne Paginierung). So bot Richelet ein im Vergleich zu den Akademiewörterbüchern vollständigeres Bild der damaligen französischen Sprache und erlebte in Frankreich auch einen gewissen Erfolg; vor allem konnte er aber von der *Académie française* unbehelligt arbeiten.

Antoine Furetière erregte mit seinem 1684 teilweise, 1690 postum vollständig veröffentlichten *Dictionnaire universel, contenant généralement tous les Mots François, tant vieux que modernes, & les Termes de toutes les Sciences et des Arts* hingegen den Unmut der Akademie: All ihren Grundhaltungen entgegen hatte das Akademiemitglied Furetière die Absicht, die französische Sprache mit ca. 40 000 Lemmata im Sinne der Vermittlung von enzyklopädischem Wissen so vollständig wie möglich zu illustrieren, was natürlich auch Fachsprache einschloss. Furetière erhielt für sein Projekt 1684 tatsächlich eine königliche Druckerlaubnis und publizierte daraufhin eine Vorschau auf sein Wörterbuch. Damit überholte er sozusagen die *Académie française*, die ihrerseits seit 1674 ein königliches Druckprivileg für die Erstellung eines französischen Wörterbuches innehatte, aber für die Redaktion ihres Werkes sehr viel Zeit benötigte. Die Akademie erhob Furetière gegenüber nun Plagiatsvorwürfe, schloss ihn aus ihren Reihen aus und erreichte auch die Aufhebung seines Druckprivilegs. Der *Dictionnaire universel* erschien trotzdem, allerdings in Den Haag und erst nach dem Tod seines Autors (Chassagne et al. 1994:63).

Diese Geschehnisse brachten allerdings die *Académie française* in Zugzwang: Sie musste den Fachsprachen daraufhin etwas mehr Raum geben und beauftragte Thomas Corneille 1690 mit der Redaktion eines separaten *Dictionnaire des Arts et des Sciences*, das als Ergänzung zum damals noch in Ausarbeitung befindlichen fachsprachenrestriktiven Akademiewörterbuch gedacht war und ebenfalls 1694 erschien. Bemerkenswert ist die extrem kurze Redaktionszeit dieses weniger enzyklopädisch, sondern hauptsächlich definitorisch angelegten Wörterbuches – ein Umstand, der nun wiederum Corneille den Vorwurf eintrug, Richelets Wörterbuch plagiiert zu haben (Chassagne et al. 1994: 58). Das Vorwort des *Dictionnaire des Arts et des Sciences* befasste sich wiederum in Form einer harschen Kritik intensiv

mit dem Wörterbuch von Furetière und einer ganzen Reihe an Defekten, die selbiges nach Meinung der Akademie aufwies – ganz anders als das *Dictionnaire des Arts et des Sciences* selber, das laut Aussage seines Autors mit höchster Sorgfalt erstellt worden war (Corneille 1694: Préface, ohne Paginierung).

4 Fachsprachen im Frankreich des 18. Jahrhunderts und danach

Wie bereits unter 3.4 festgestellt, passte die *Académie française* ihre lexikographische Ausrichtung in den Folgejahren aber jeweils nur so geringfügig wie möglich an die sich ändernde Situation an und nahm im 18. Jahrhundert auch nicht aktiv am Umbruch hin zum *âge encyclopédique* teil: Im Zuge einer generellen Hinwendung der Öffentlichkeit zu Fachlichem und Fachsprachlichem bildete sich damals das Bewusstsein heraus, dass Fachsprache und Gemeinsprache nicht strikt voneinander getrennt werden können. Der bis dahin vorherrschende, fachsprachenfreie *style noble* wich nun einem neuen Sprecherverhalten, das sich durch die zunehmende Verwendung von Fachsprache in nicht-fachlichen Kommunikationssituationen der gehobenen Schichten auszeichnete. Damit ging wiederum ein Aufschwung der Fachwörterbücher einher: Das für das 18. Jahrhundert wichtigste fachsprachliche lexikographische Werk war zweifelsohne Diderots und d'Alemberts vielimitierte *Encyclopédie*, deren Ziel es war, sämtliches Wissen ihrer Zeit zu erfassen.

Etwas später, während der Französischen Revolution, entstand dann aus einer politischen Motivation heraus ein starkes wissenschaftliches Interesse für die sprachlichen Gegebenheiten in Frankreich, das in der großen Sprachumfrage des Abbé Henri Grégoire seine Verkörperung fand: Im Hinblick auf die Fachsprachen versuchte er vor allem herauszufinden, in welchen Bereichen die französische Standardvarietät ein zu schwaches Kommunikationspotenzial aufwies – es waren schließlich wie bereits dargestellt auch aufgrund der sprachpuristischen Bestrebungen des 17. Jahrhunderts sprachliche Lücken entstanden. Nach einer vier Jahre andauernden Datenerhebung zeigte sein *Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française*, den er am 4. Juni 1794 der Nationalversammlung vorlegte, nicht unerhebliche Problemstellen im Bereich der Fachsprachen: Die Sprechergemeinschaft musste zur Gewährleistung einer effizienten Kommunikation beispielsweise in der Handwerker- und Bauernsprache bzw. in Fächern, die Berührungspunkte mit dem Alltag der einfachen Bevölkerung aufwiesen, häufig auf Begriffe aus Regiolekten zurückgreifen, da es schlichtweg keine standardfranzösischen Ausdrucksmittel gab. Diese Tendenz machte Lösungsstrategien zugunsten einer einheitlichen und starken Sprache im Sinne einer geeinten und starken Nation notwendig. Der Wandel weg von regional geprägten Termini, hin zu einer panfranzösisch einheitlichen Fachsprachverwendung, sollte mittels einer kostenlosen Grundschulausbildung für alle Kinder in französischer Sprache, der Übersetzung fachsprachlichen Wortgutes aus den Patois ins Normfranzösische und durch die Verbreitung entsprechender fachsprachlicher Texte mit Relevanz für das all-

tägliche Leben erreicht werden (Bochmann 1993: 64ff; ib.: 88; ib.: 139). Die offizielle französische Lexikographie trug allerdings nur wenig dazu bei.

Nicht nur für das 18. Jahrhundert, sondern auch noch bis in die Gegenwart lässt sich feststellen, dass mehr als vier Jahrhunderte fachsprachenrestriktiver Sprachpolitik in Frankreich Spuren in Form von sprachlichen Lücken hinterlassen haben. Obgleich sich heute neue Werke der französischen Lexikographie den Fachsprachen gegenüber aufgeschlossener zeigen und auch die institutionalisierte Sprachpflege sehr bemüht ist, vorhandenen Problemen entgegenzuwirken, strömen etwa seit Ende des 19. Jahrhunderts zahlreiche Fremdwörter – vornehmlich aus dem Englischen – ins Französische, um dort vor allem fachsprachliche Defizite auszugleichen. Von dort aus finden sie anschließend häufig Eingang in die Gemeinsprache, die sie immer stärker durchsetzen. Dies hat zu einem Aufruf zur „Verteidigung der Sprache" geführt. Dieser Kampf hat bis heute nichts von seiner Vehemenz verloren: Eine moderne Sprachgesetzgebung sowie eine beachtliche Anzahl staatlicher und privater Sprachpflegeorganisationen stehen im Dienst des Schutzes des Französischen, was unter anderem einen koordinierten Ausbau der französischen Fachsprachen und ein entsprechendes lexikographisches Wirken impliziert.

5 Resümee

Zusammenfassend lässt sich feststellen, dass die französische Sprache im europäischen Vergleich durchaus eine Art Sonderstatus innehat. In kaum einem anderen Land wird die Tradition der Sprachpflege so großgeschrieben und so konsequent betrieben; in kaum einem anderen Land bringen die Sprachnutzer ihrer Nationalsprache so viel Aufmerksamkeit entgegen. Zugleich war die französische Sprachpflege allerdings lange Zeit über stark auf die Gemeinsprache fokussiert und vernachlässigte die Fachsprachen. Deshalb wird ihr in ihrer institutionalisierten Form, der *Académie française*, auch immer wieder vorgeworfen, sie würde den Anforderungen einer modernen Sprache im Wandel nicht gerecht; sie praktiziere eine Art der lexikalischen Zensur, die der Sprache ernst- und dauerhaft Schaden zufügen könne. In der Tat schlug sich das sprachpolitische Eingreifen der Akademie seit dem 17. Jahrhundert mit immer noch anhaltender Wirkung im tatsächlichen Sprachgebrauch nieder: Die französische Sprache ist eine vereinheitlichte Sprache, eine wahre Standardvarietät, die allerdings gerade ob der zur Vereinheitlichung und ästhetischen Perfektionierung gesetzten restriktiven Maßnahmen teils nach wie vor unter Defiziten in der Lexik leidet. Für das Heute gilt, dass sich der für das Französische typische, lebhafte sprachpolitische Diskurs im Lauf der Zeit inhaltlich gewandelt hat, aber immer noch von der immensen Bedeutung zeugt, die der Pflege und Wertschätzung der Nationalsprache in Frankreich zugesprochen wird.

6 Primär- und Sekundärliteratur

Académie française > Le Dictionnaire > Les neuf préfaces > 1re préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/preface-de-la-premiere-edition-1694 [06/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 2e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/preface-de-la-deuxieme-edition-1718 [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 3e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-troisieme-edition-1740 [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 4e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-quatrieme-edition-1762 [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 5e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-cinquieme-edition-1798 [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 6e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/sixieme-preface [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 7e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-septieme-edition-1877 [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 8e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-huitieme-edition-1932-1935 [08/04/2014].

Académie française > Le Dictionnaire > La neuvième édition. Zugriff unter: http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/preface-la-neuvieme-edition [08/04/2014].

Académie française > Le Dictionnaire > Les neuf préfaces > 9e préface. Zugriff unter: http://academie-francaise.fr/le-dictionnaire/la-9e-edition [08/04/2014].

Académie française > L'institution > Les statuts. Zugriff unter: http://www.academie-francaise.fr/sites/academie-francaise.fr/files/statuts_af.pdf [06/04/2014].

Assemblée nationale > Accueil > Histoire et Patrimoine > Ordonnance de Villers-Cotterêts. Zugriff unter: http://www.assemblee-nationale.fr/histoire/villers-cotterets.asp [05/04/2014].

Berschin, H./Felixberger, J./Goebl, H. ([2]2008) *Französische Sprachgeschichte. 2., überarbeitete und ergänzte Auflage*. Hildesheim: Georg Olms Verlag.

Bochmann, K. (1993) Sprachpolitik in der Romania. Zur Geschichte sprachpolitischen Denkens und Handelns von der Französischen Revolution bis zur Gegenwart. Berlin/New York: Walter de Gruyter.

Bouhours, D. (1962) Les Entretiens d'Ariste et d'Eugène. Présentation de Ferdinand Brunot. Paris: Armand Colin.

Bouhours, D. (1973) Remarques nouvelles sur la Langue françoise (1675). Suite des Remarques nouvelles sur la Langue françoise (1687). Genf: Slatkine Reprints.

Brunot, F. ([2]1930) Histoire de la Langue française des Origines à 1900, Tome III: La Formation de la Langue classique. Première Partie, Édition revue et corrigée. Paris: Librairie Armand Colin.

Chassagne, A./Gasnault, P./Pastoureau, M./Service du Dictionnaire de l'Académie française (1994) *Le Dictionnaire de l'Académie française: 1694-1994 – sa naissance et son actualité*. Paris: Institut de France.

Corneille, T. (1694) Dictionnaire des Arts et des Sciences. Par M. D. C. de l'Académie françoise. Tome Premier. A-L. Paris: Chez la Veuve de Jean Baptiste Coignard et chez Jean Baptiste Coignard. Zugriff unter: http://gallica.bnf.fr/ark:/12148/bpt6k50507s [09/04/2014].

Du Bellay, Joachim (1969) La Deffence et Illustration de la Langue francoyse. Edition critique par Henri Chamard. Genf: Slatkine Reprints.

Ehlich, K./Osnner, J./Stammerjohann, H. (eds.) *Hochsprachen in Europa. Entstehung, Geltung, Zukunft.* Freiburg im Breisgau: Fillibach Verlag.

Elias, N. (1983) Die höfische Gesellschaft. Untersuchungen zur Soziologie des Königtums und der höfischen Aristokratie. Frankfurt/Main: Suhrkamp (Taschenbuch Wissenschaft 423).

Fluck, H. ([5]1996) Fachsprachen. Einführung und Bibliographie. Fünfte, überarbeitete und erweiterte Auflage. Basel/Tübingen: A. Francke Verlag (UTB 483).

Grimm, Jürgen (2005) *Französische Klassik. Lehrbuch Romanistik*. Stuttgart/Weimar: Verlag J. B. Metzler.

Hassler, G. (1998) Anfänge der europäischen Fachsprachenforschung im 17. und 18. Jahrhundert. In: Hoffmann, L./Kalverkämper, H./Wiegand, H. (eds.) *Fachsprachen. Languages for Special Purposes. 1. Halbband*. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1), 322-326.

Hoffmann, L./Kalverkämper, H./Wiegand, H. (eds.) *Fachsprachen. Languages for Special Purposes. 1. Halbband*. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1).

Lindemann, Margarete (1994) Die französischen Wörterbücher von den Anfängen bis 1600. Entstehung und typologische Beschreibung. Tübingen: Max Niemeyer Verlag (Lexicographica Series Maior 54).

Mayer, M. (2013) *Sprachpflege und Sprachnormierung in Frankreich am Beispiel der Fachsprachen vom 16. Jahrhundert bis in die Gegenwart*. Innsbruck: Innsbruck University Press (Studien des Interdisziplinären Frankreich-Schwerpunkts der Universität Innsbruck 5).

Müller, B. (1975) Das Französische der Gegenwart. Varietäten, Strukturen, Tendenzen. Heidelberg: Carl Winter Universitätsverlag.

Pécheur, J. (2001) Nouveaux espaces pour le français. In: Ehlich, K./Osnner, J./Stammerjohann, H. (eds.) *Hochsprachen in Europa. Entstehung, Geltung, Zukunft.* Freiburg im Breisgau: Fillibach Verlag, pp. 47-72.

Richelet, C. (1680) Dictionnaire françois, contenant les Mots et les Choses, plusieurs nouvelles Remarques sur la Langue françoise. Genf: Chez Jean Herman Widerhold. Zugriff unter: http://gallica.bnf.fr/ark:/12148/bpt6k509323/f6.image [09/04/2014].

Roelcke, T. ([2]2005) *Fachsprachen. 2., durchgesehene Auflage*. Berlin: Erich Schmidt Verlag (Grundlagen der Germanistik).

Vaugelas, C. (1880) Remarques sur la Langue françoise, par Vaugelas. Nouvelle Édition, par A. Chassang. Tome Premier. Paris: Léopold Cerf.

EU-Terminologie in den einsprachigen Wörterbüchern des Deutschen

Diana Stantcheva
American University in Bulgaria
dstantcheva@aubg.bg

Abstract

Die heutige Europäische Union blickt nun schon auf eine über 50 jährige Geschichte zurück. In dieser Zeit ist die europäische Staatengemeinschaft von ursprünglich 6 auf derzeit 28 Länder gewachsen. Deutschland gehörte zu den Gründungsmitgliedern der Europäischen Gemeinschaft für Kohle und Stahl, der Europäischen Wirtschaftsgemeinschaft und der Europäischen Atomgemeinschaft, der Organisationen, die später in die Europäische Union zusammenflossen. Österreich trat 1995 der Europäischen Union bei. Diese Fakten zum Anlass nehmend, untersucht der vorliegende Beitrag die lexikographische Kodifizierung EU-bezogener Terminologie in ausgewählten allgemeinen einsprachigen Wörterbüchern des Deutschen von 1961 bis 2009.

Keywords: terminology; terminography; lexicography; specialised languages; monolingual German dictionaries; Terminologie der Europäischen Union, Deutsch-als-Fremdsprache-Unterricht; Europa-Studien; einsprachige Wörterbücher

1 Einleitung

Mit dem Begriff *Terminus* und synonym dazu *Fachwort* bzw. *Fachwendung* wird im vorliegenden Beitrag zusammen mit Wendt (1997: 63) „[...] eine Funktion beschrieben, die jede Einheit einer natürlichen und künstlichen Sprache (Wort, Wortverbindung, Abkürzung, Symbol) unter bestimmten Bedingungen übernehmen kann". Im Falle der EU-Terminologie wären diese Bedingungen das Auftreten einer sprachlichen Einheit in der fachlichen Kommunikation über das vereinte Europa und die Europäische Union mit ihren Institutionen und Organen oder die Zugehörigkeit dieser Einheit zum fachlichen Gegenstand *Amtssprache und Arbeitssprache der Europäischen Union*. Ausgehend von dieser Definition verstehe ich im Folgenden unter EU-Terminologie nicht nur genuine EU-Termini der „supranationalen Varietät ‚EU-Deutsch'"[1], wie *Euro*, *Europäische Union*, *EU*, *EU-Richtlinie*, *Eurobarometer*, *Europäisches Parlament*, *Europäische Kommission*, *Europäische Zentralbank*, *Europäische Integration*, sondern auch Fachwörter und -wendungen aus den Sachgebieten Wirtschaft, Handel, Politik, Recht, wie z.B.

1 Zum Begriff „supranationale Varietät ‚EU-Deutsch'" vgl. Markhardt (2004).

Supranationalität, *Sicherheitspolitik*, *Binnenmarkt*, *Mehrheitsentscheidung*, *Minderheitenschutz*, *Verordnung*, die in Bezug auf das Zusammenwachsen Europas und/oder die Europäische Union verwendet werden. Die Basis der vorliegenden empirischen Analyse der lexikographischen Kodifizierung und Darstellung von EU-bezogenen Termini bildet die Wortstrecke *eu – ev* in einsprachigen Wörterbüchern des Deutschen. Es handelt sich dabei um eine verhältnismäßig kurze Wortstrecke mit einer großen Konzentration an EU-relevanten Wörterbucheinträgen. Die ausgewählten Wörterbücher decken den Zeitraum von 1961 bis 2009 ab und lassen damit Veränderungen in der Terminologie und deren Darstellung im Wörterverzeichnis festhalten.

Der Untersuchung liegen folgende allgemeine einsprachige Wörterbücher des Deutschen zugrunde:

Wörterbuch der deutschen Gegenwartssprache (1961-1977) [WDG]

Ullstein Lexikon der deutschen Sprache (1969) [Ullstein-LdS]

Der Sprachbrockhaus (1972) [SBH_8]

Duden. Das große Wörterbuch der deutschen Sprache (1976-1981) [Duden-GWB_1]

Brockhaus-Wahrig. Deutsches Wörterbuch (1980-1984) [BW]

Duden. Deutsches Universalwörterbuch (1983) [Duden-UW_1]

Knaurs Großes Wörterbuch der deutschen Sprache (1985) [Knaurs-GW]

Gerhard Wahrig. Deutsches Wörterbuch (1986) [Wahrig-DW_4]

Duden. Das große Wörterbuch der deutschen Sprache (1993-1995) [Duden-GWB_2]

Duden. Das große Wörterbuch der deutschen Sprache (1999) [Duden-GWB_3]

Gerhard Wahrig. Deutsches Wörterbuch (2000) [Wahrig-DW_7]

Duden. Deutsches Universalwörterbuch (2003) [Duden-UW_5]

Duden. Deutsches Universalwörterbuch (2007) [Duden-UW_6]

Gerhard Wahrig. Deutsches Wörterbuch (2008) [Wahrig-DW_8]

Darüber hinaus werden auch drei Lernerwörterbücher, ein Rechtschreibwörterbuch und ein historisches Wörterbuch des Deutschen herangezogen, um ein umfassenderes Bild von der Behandlung der EU-Terminologie in den Wörterbüchern des Deutschen in der untersuchten Zeitspanne zu bekommen:

Langenscheidt GWDaF (1998) [$LGWDaF_1$]

Grimm. Deutsches Wörterbuch (Neubearbeitung, Band 8, 1999) [Grimm-DW]

Pons GWDaF (2004) [PGWDaF]

Duden. Die deutsche Rechtschreibung (2006) [Duden-DR_{24}]

Langenscheidt GWDaF (2008) [$LGWDaF_2$]

Gerhard Wahrig. Großwörterbuch DaF (2008) [Wahrig-$GDaF_1$]

Duden. Die deutsche Rechtschreibung (2009) [Duden-DR_{25}]

Die vorliegende Untersuchung ist im Zusammenhang mit der studienbegleitenden fachsprachlichen Deutschausbildung im Rahmen des Bachelorstudienganges Europa-Studien[2] an der American University in Bulgaria (= AUBG) entstanden und ist Teil eines größeren Projekts, das zum Ziel hat, das Curriculum für diesen Unterrichtstyp zu optimieren, Übungsmaterialien mithilfe von Korpustexten geschriebener Sprache zu erstellen sowie ein Minimum an deutschen EU-Termini für die Zwecke der studienbegleitenden Fremdsprachenausbildung zu ermitteln.

2 Lexikographische Kodifizierung EU-bezogener Terminologie in den untersuchten Wörterbüchern

Die EU-Terminologie zeichnet sich durch Breite und Dynamik aus. Das macht sich auch in den durchgesehenen Nachschlagewerken der letzten zwei Jahrzehnte bemerkbar. Jedes neue Wörterbuch und jede weitere Wörterbuchauflage verzeichnen neue Termini. Die Anzahl der aufgenommenen Lemmata der Wortstrecke *eu – ev* ist in den herangezogenen Nachschlagewerken kontinuierlich gewachsen: Von fünf Begriffen (*Europa*, *Europäer*, *europäisch*, *europäisieren*, *Europäisierung*) zum damaligen Zeitpunkt noch ohne gemeinschaftseuropäischen Bezug im mehrbändigen WDG 1961-1977 (2. Band, 1964-1967) bis hin zu 47 Termini mit gemeinschaftseuropäischem Bezug im einbändigen DUDEN-DR$_{25}$ 2009. Auffällig viel sind dabei Neulexeme mit den Komponenten *EU-*, *Europa-* und *Euro-* als Erstglied, die vor allem politische und wirtschaftliche Neuentwicklungen innerhalb der Europäischen Union bezeichnen, wie z.B. *EU-Bürger*, *EU-Erweiterung*, *EU-Staat*, *Eurocent*, *Eurozone* sowie terminologische Wortverbindungen mit dem Attribut *europäisch*, wie z.B. *Europäischer Gerichtshof*, *Europäische Union*, *Europäische Währungseinheit*, *Europäisches Parlament*, *Europäische Zentralbank*, *Europäische Kommission*. Der Entwicklung der EU-Terminologie tragen auch die von manchen Lernerwörterbüchern angebotenen kurzen enzyklopädischen Zusatzinformationen in Form eines Informationskastens außerhalb des Wörterbuchartikels Rechnung, wie z.B. zu den Begriffen *Euro* und *Europäische Union* im PGWDaF 2004 und zu *EU* und *Euro* im WAHRIG-GDaF$_1$ 2008.

Das Veralten von EU-Termini lässt sich anhand der durchgesehenen Wörterbücher ebenfalls beobachten, auch wenn die Wörterbücher dies nicht immer explizit deutlich machen: Bestimmte EU-Termini, die mittlerweile zu Historismen geworden sind, weil das Denotat außer Gebrauch kam, sind in manchen der Wörterbücher noch vorhanden, allerdings nicht immer entsprechend markiert, wie z.B.: *Europäische Gemeinschaft für Kohle und Stahl* im WAHRIG-DW$_8$ 2008 (unter *europäisch*). Dieser Begriff ist nicht als Historismus markiert und nur durch die Zeitangaben in der semantischen Paraphrase

2 Den Bachelorstudiengang Europa-Studien gibt es an der AUBG seit 1999 als Nebenfach (*European Studies Minor*) und seit 2001 als Hauptfach (*European Studies Major*). Im Rahmen des Hauptfaches ist seit 2004 eine obligatorische studienbegleitende Fremdsprachenausbildung in einer der Sprachen Deutsch, Französisch oder Spanisch vorgesehen. Näheres zum Studiengang Europa-Studien an der AUBG siehe unter http://www.aubg.bg/template5.aspx?page=4419&menu=001001002003 und http://www.aubg.bg/template5.aspx?page=4428&menu=001001003003.

wird signalisiert, dass die Europäische Gemeinschaft für Kohle und Stahl nicht mehr existiert: „1952 gegründete u. 2002 wieder aufgelöste Gemeinschaft europäischer Länder mit dem Ziel eines gemeinsamen Marktes für Kohle u. Eisen, Montanunion".

Insgesamt unklar bleiben die Kriterien für die Aufnahme der EU-Termini in die Wörterverzeichnisse, denn alle Wörterbücher weisen Lemmalücken auf. So verzeichnen z.B. WAHRIG-DW$_7$ 2000, WAHRIG-DW$_8$ 2008 und Grimm-DW 1999 den Begriff *Euro* als Lemma nicht. *Eurogeld* ist gebucht nur im DUDEN-DR$_{24}$ 2006 und DUDEN-DR$_{25}$ 2009, *Euromünze* nur im DUDEN-DR$_{24}$ 2006, DUDEN-DR$_{25}$ 2009 und WAHRIG-DW$_8$ 2008, *Eurowährung* nur im WAHRIG-DW$_8$ 2008.

Eine Recherche in den Textkorpora des Deutschen[3] bringt darüber hinaus eine Reihe von Termini zu Tage, die allesamt als Lemma nicht in den untersuchten Wörterbüchern zu finden sind: *Europarecht*, *europarechtlich*, *Europahymne*, *Europabegeisterung*, *Europamüdigkeit*, *Europapolitiker*, *europapolitisch*, *Europa-Studien*, *europakritisch*, *Europawahlkampf*, *Eurokratie*, *EU-Ebene*, *EU-Haushalt* usw., um hier nur einige zu nennen.

3 Lexikographische Defizite bei der Darstellung von EU-Termini in den untersuchten Wörterbüchern

In diesem Abschnitt soll auf die Darstellung der EU-Terminologie in den Wörterverzeichnissen der untersuchten Wörterbücher eingegangen werden. Exemplarisch soll dabei die lexikographische Darstellung folgender EU-Begriffe näher betrachtet werden: *Europa*, *Europäer*, *Europäerin*, *Euro*, *Euroskeptiker*, *Europäische Zentralbank*, *EU*, *Eurokrat*, *EU-Erweiterung* und *Euroland*.

Bei *Europa*, *Europäer*, *Europäerin* und *Euro* handelt es sich einerseits um Termini, die eine Bedeutungserweiterung im Rahmen der EU-Terminologie erfahren haben, wie *Europa*, *Europäer*, *Europäerin*, und andererseits um ein Neulexem *Euro*, dessen Denotat genuin mit der Europäischen Union verbunden ist. Im Folgenden soll anhand der lexikographischen Darstellung dieser Begriffe deutlich gemacht werden, dass die durchgesehenen Nachschlagewerke, von wenigen Ausnahmen abgesehen, den Bedeutungswandel der Begriffe *Europa*, *Europäer*, *Europäerin* nicht reflektieren und den terminologischen Neologismus *Euro* uneinheitlich beschreiben.

In sechs der durchgesehenen Wörterbücher wird *Europa* gemäß der lexikographischen Tradition, Eigennamen nicht zu verzeichnen, als Lemma nicht aufgenommen: So im BW 1980-1984, Knaurs-GW 1985, WAHRIG-DW$_4$ 1986, WAHRIG-DW$_7$ 2000, WAHRIG-DW$_8$ 2008 und Grimm-DW 1999. In acht Wörterbüchern wird *Europa* als monosemes Lexem dargestellt: WDG 1961-1977 (2. Band, 1964-1967), Ullstein-LdS 1969, LGWDaF$_1$ 1998, LGWDaF$_2$ 2008 verzeichnen nur die Bedeutung ‚Kontinent'. Der Sprachbrockhaus 1972, DUDEN-DR$_{24}$ 2006 und DUDEN-DR$_{25}$ 2009 listen lediglich die Bedeutung ‚griechische

3 Z.B. im ZEIT-Korpus (http://www.dwds.de), im Korpus der Berliner Zeitung (http://www.dwds.de) oder in den Korpora geschriebener Sprache des Instituts für Deutsche Sprache in Mannheim (http://www.ids-mannheim.de/cosmas2/).

weibliche Sagengestalt' auf und PGWDaF 2004 nur die Bedeutung ‚Staatenkomplex'. DUDEN-GWB$_1$ 1976-1981 ist das erste der durchgesehenen Wörterbücher, das *Europa* als polysemes Wort darstellt und die Unterbedeutung ‚Staatenkomplex' auflistet:

> Staatenkomplex, der durch einen Zusammenschluß der europäischen Staaten entstehen soll: Konrad Adenauer will ein E[uropa], das sich in eine größere Atlantische Gemeinschaft einfügt (Dönhoff, Ära 128); sich für E[uropa] (einen Zusammenschluß der europäischen Staaten) begeistern, einsetzen.

Diese Zeilen sind in die neueren DUDEN-Wörterbücher (DUDEN-GWB$_2$ 1993-1995, DUDEN-GWB$_3$ 1999, DUDEN-UW$_5$ 2003 und DUDEN-UW$_6$ 2007) mit Ausnahme der Tilgung des Dönhoffs Zitats und des Verbs „begeistern" unverändert übernommen, was insbesondere im Hinblick auf die Formulierung, dass der Staatenkomplex auch nach einer über 50 jährigen Geschichte noch „entstehen soll", problematisch ist. Im DUDEN-GWB$_2$ 1993-1995, DUDEN-GWB$_3$ 1999, DUDEN-UW$_1$ 1983, DUDEN-UW$_5$ 2003 und DUDEN-UW$_6$ 2007 sind auch je zwei Lemmata *Europa* verzeichnet: Das erste ist polysem und hat die Unterbedeutungen ‚Kontinent' und ‚Staatenkomplex', das zweite monosem mit der Bedeutung ‚weibliche Gestalt der griechischen Mythologie'.

LGWDaF$_1$ 1998, LGWDaF$_2$ 2008 und GRIMM-DW 1999 listen eine Reihe von Zusammensetzungen mit *Europa* als Erst- oder Zweitglied auf, allerdings ohne eine Bedeutungserläuterung oder Illustrationsbeispiele. Dabei wird auch kein Unterschied zwischen Europa als ‚Kontinent' und Europa als ‚Staatengemeinschaft' gemacht, sodass beispielsweise *Europa-Parlamentarier*, *Europareise*, *Europa-Währung* im GRIMM-DW 1999 und *Nordeuropa*, *Westeuropa*, *Europapolitik* im LGWDaF$_1$ 1998 und LGWDaF$_2$ 2008 nebeneinander und ohne Bedeutungsdifferenzierung stehen.

Keines der untersuchten Wörterbücher führt phraseologische Wortverbindungen mit dem Grundelement *Europa* als ‚Staatengemeinschaft' auf, wie z.B. *(die) Festung Europa*, *das offizielle Europa*, *das gemeinsame Haus Europa*, *Europa à la carte*, die in Textkorpora des Deutschen mehrfach belegt sind.

Noch weniger reflektieren die durchgesehenen Nachschlagewerke den Bedeutungswandel bei den Termini *Europäer* und *Europäerin*. Der Begriff *Europäer* ist im GRIMM-DW 1999 als Lemma nicht gebucht. WDG 1961-1977 (2. Band, 1964-1967), DUDEN-DR$_{24}$ 2006 und DUDEN-DR$_{25}$ 2009 verzeichnen *Europäer* ohne eine Bedeutungserläuterung. Als monosemes Wort mit der Bedeutung 'Bewohner des europäischen Kontinents' findet man *Europäer* im Sprachbrockhaus 1972, Ullstein-LdS 1969, WAHRIG-DW$_4$ 1986, WAHRIG-DW$_7$ 2000 und WAHRIG-DW$_8$ 2008. PGWDaF 2004 hat den Begriff ebenfalls nur mit einer Bedeutung aufgenommen, nämlich mit „jmd., der Bürger eines europäischen Staates ist". Der erste Eintrag von *Europäer* als polysemes Wort in den durchgesehenen Wörterbüchern ist erneut im DUDEN-GWB$_1$ 1976-1981 zu finden, mit den Unterbedeutungen „Vertreter des abendländischen Kulturkreises" und „Politiker, der für einen Zusammenschluß der europäischen Staaten eintritt". Seitdem ist *Europäer* als polysemes Wort mit variierenden Bedeutungserläuterungen verzeichnet, wie:

1. Einwohner Europas
2. Anhänger einer Politik, die den Interessen der Europäischen Gemeinschaft Vorrang einräumt (BW 1980-1984)

1. Bewohner Europas
2. jmd., der für den Zusammenschluß der Staaten Europas ist (KNAURS-GW 1985)

1. j-d, der in Europa geboren ist u. zu e-r europäischen Nation gehört
2. verwendet für j-n, dessen Denken u. Handeln die Einheit Europas zum Ziel hat (LGWDaF1 1998, LGWDaF2 2008)

DUDEN-UW$_1$ 1983 ist das erste der untersuchten Wörterbücher, das *Europäer* auch als 'Einwohner des europäischen Staatenkomplexes' definiert, auch wenn diese Definition durch einen Verweis auf das Lemma *Europa* zustande kommt:

1. Ew. zu [1]Europa [d.i. 'Einwohner des europäischen Kontinents' und 'Einwohner des Staatenkomplexes' – meine Anmerkung D. S.]
2. Politiker, der für einen Zusammenschluß/ss der Staaten Europas eintritt.

DUDEN-GWB$_2$ 1993-1995, DUDEN-GWB$_3$ 1999, DUDEN-UW$_5$ 2003 und DUDEN-UW$_6$ 2007 haben diese Artikelstruktur und diese Bedeutungserläuterung ohne Veränderungen übernommen. Abgesehen von den fünf DUDEN-Wörterbüchern erwähnen die anderen durchgesehenen Nachschlagewerke die Unterbedeutung 'Einwohner des Staatenkomplexes in Europa' bzw. 'EU-Bürger' zu *Europäer* nicht. In Textkorpora des Deutschen ist diese Verwendung von *Europäer* dagegen etwa seit den 70er Jahren des 20. Jahrhunderts belegt.

Die Movierung *Europäerin* ist in den durchgesehenen Wörterbüchern zum ersten Mal im DUDEN-GWB$_2$ 1993-1995 als Lemma mit der Bedeutungserläuterung „w[eibliche] Form zu ↑ Europäer" gebucht (so auch im DUDEN-GWB$_3$ 1999, DUDEN-UW$_5$ 2003, DUDEN-UW$_6$ 2007). GRIMM-DW 1999, WAHRIG-DW$_7$ 2000, LGWDaF$_1$ 1998 verzeichnen *Europäerin* nicht, LGWDaF$_2$ 2008, DUDEN-DR$_{24}$ 2006 und DUDEN-DR$_{25}$ 2009 haben *Europäerin* ohne eine semantische Paraphrase aufgenommen. PGWDaF 2004 listet *Europäer* und *Europäerin* nebeneinander mit dem bestimmten Artikel, der Genitivendung und der Pluralform nur des männlichen Substantivs auf[4] und definiert beide Begriffe zusammen als „jmd., der Bürger eines europäischen Staates ist". Im WAHRIG-DW$_8$ 2008 und im WAHRIG-GDaF$_1$ 2008 ist die Definition „weibl[icher] Europäer" zu finden. Mit anderen Worten: Die Unterbedeutung 'EU-Bürgerin' ist ebenfalls nur in den DUDEN-Wörterbüchern (DUDEN-GWB$_2$ 1993-1995, DUDEN-GWB$_3$ 1999, DUDEN-UW$_5$ 2003, DUDEN-UW$_6$ 2007) vorhanden, auch wenn sie sich auch hier durch eine Zirkeldefinition erschließen lässt: *Europäerin* → *Europäer* → *Europa*.

Die Währungsbezeichnung *Euro* ist in den untersuchten Wörterbüchern erst seit 1998 gebucht, allerdings nicht in allen, denn GRIMM-DW 1999, WAHRIG-DW$_7$ 2000 und WAHRIG-DW$_8$ 2008 verzeichnen den *Euro* als Lemma nicht. Die Wörterbücher, die den Begriff aufgenommen haben, geben voneinander abweichende, unpräzise und zuweilen unkorrekte Definitionen an. Das LGWDaF$_1$ 1998 erläutert

4 Diese Darstellungsform ist im gesamten Wörterverzeichnis bei der Bildung einer weiblichen Personenbezeichnung mit einem Suffix von einer männlichen Form zu finden, z.B. auch „Europagegner, Europagegnerin der <-s, ->". So eine Darstellung mag platzökonomisch sein, ist aber höchst problematisch besonders in Anbetracht der Tatsache, dass es sich bei PGWDaF 2004 um ein Lernerwörterbuch handelt.

z.B. *Euro* als „Bezeichnung für die gemeinsame Währung in den Staaten der Europäischen Union (ab 1999)". Diese Definition ist nicht korrekt, denn nicht alle EU-Länder haben den Euro als Währung. Im LGWDaF$_2$ 2008 ist diese Bedeutungserläuterung korrigiert in: „Bezeichnung für die gemeinsame Währung in den Staaten der Europäischen Währungsunion". DUDEN-UW$_5$ 2003, DUDEN-UW$_6$ 2007, DUDEN-GWB$_3$ 1999 beschreiben den Euro als „Währungseinheit der Europäischen Währungsunion", DUDEN-DR$_{24}$ 2006 und DUDEN-DR$_{25}$ 2009 als „europ[äische] Währungseinheit", und PGWDaF 2004 als „die europäische Währungseinheit".

Wenn man die Wörterbucheinträge zum Lemma *Euro* vergleicht, fällt nicht nur die unterschiedliche Bedeutungserläuterung auf, sondern auch die variable Darstellung der Formen des Plurals und des Genitivs in den einzelnen Wörterbüchern:[5]

Eu|ro, der; -[s], -s <aber: 10 Euro> (DUDEN-UW$_5$ 2003, DUDEN-UW$_6$ 2007, DUDEN-GWB$_3$ 1999)

Eu|ro, der; -[s], -s ... 30 Euro (DUDEN-DR$_{24}$ 2006 und DUDEN-DR$_{25}$ 2009)

Eu·ro *der*; -(*s*), -(*s*); (LGWDaF$_1$ 1998)

Eu·ro *der*; -(*s*), -(*s*); ... *Dieses Buch kostet 10 Euro* (LGWDaF$_2$ 2008)

Eu·ro der <-s, -s> (PGWDaF 2004)

Keines der untersuchten Wörterbücher verzeichnet auch phraseologische Wortverbindungen zum Lemma *Euro*, wie z.B. *der schnelle Euro*, *jeden Euro zweimal umdrehen*, die in Textkorpora des Deutschen mehrfach belegt sind.

Der Begriff *Euroskeptiker* wird im DUDEN-UW$_6$ 2007, DUDEN-GWB$_2$ 1993-1995 und im DUDEN-GWB$_3$ 1999 als „Politiker, der dem Europagedanken, der Politik der Europäischen Gemeinschaften skeptisch gegenübersteht" umschrieben. Warum ein Euroskeptiker nur ein Politiker sein soll, leuchtet dabei nicht ein. Nicht ganz unproblematisch in dieser Definition ist auch die Fachwendung „Europäische Gemeinschaften", denn die drei Gemeinschaften (EGKS, EURATOM und EG) sind seit 1993 in der Europäischen Union zusammengeführt. In den untersuchten Wörterbüchern herrscht darüber hinaus Unsicherheit im Hinblick auf die Markierung des Begriffs: *Euroskeptiker* wird z.B. im PGWDaF 2004 mit dem Marker „POL" (für „Politik"), im DUDEN-GWB$_2$ 1993-1995, DUDEN-GWB$_3$ 1999 und DUDEN-UW$_6$ 2007 mit „Politik Jargon" versehen. Im DUDEN-DR$_{25}$ 2009 ist der Begriff dagegen nicht markiert und dementsprechend als neutral anzusehen.

Die Paraphrase des Terminus *Europäische Zentralbank* (unter *europäisch*) ist im WAHRIG-DW$_7$ 2000 gänzlich unkorrekt: „1950 von den Mitgliedstaaten der OEEC gegründete Union zur multilateralen Verrechnung von Zahlungen, die 1958 durch das europäische Währungsabkommen abgelöst wurde". Diese Definition wurde höchstwahrscheinlich vom BW 1980-1984 übernommen, allerdings nicht richtig eingesetzt, denn die gleiche Definition steht im BW unter der terminologischen Wortverbindung *Europäische Zahlungsunion*.

5 Die fehlende Eindeutigkeit in der Darstellung der flektierten Formen dieses Neulemmas hat HERBERG (2001) bereits in sechs anderen Wörterbüchern des Deutschen festgestellt, die im Zeitraum 1996-1999 erschienen sind.

Eine sachliche Ungenauigkeit enthält auch der folgende Text im Infokasten zu *EU* im WAHRIG-GDaF$_1$ 2008: „Mit der Abkürzung *EU* wird der große europäische Staatenbund, die *Europäische Union*, bezeichnet. Hervorgegangen ist diese aus der im Jahr 1952 gegründeten Europäischen Wirtschaftsgemeinschaft (EWG)". Die Europäische Wirtschaftsgemeinschaft wurde 1957 und nicht 1952 gegründet und die EU ist, wie unter *Euroskeptiker* bereits erwähnt, nicht einzig und allein aus der Europäischen Wirtschaftsgemeinschaft hervorgegangen.

Problematisch ist auch die Definition des Lemmas *EU* im LGWDaF$_1$ 1998 und LGWDaF$_2$ 2008: „e-e Union von europäischen Staaten, die in allen politischen Bereichen eng zusammenarbeiten u. e-e politische Einheit Europas wollen". Diese Definition lässt Bereiche wie Wirtschaft und Kultur außer Acht und reduziert die Rolle der EU auf den politischen Bereich.

Das Lemma *Eurokrat* im DUDEN-GWB$_2$ 1993-1995, DUDEN-GWB$_3$ 1999 und im DUDEN-UW$_1$ 1983, DUDEN-UW$_5$ 2003 und DUDEN-UW$_6$ 2007 wird erläutert als „Politiker, der den Interessen der Europäischen Gemeinschaft (besonders gegenüber den USA) Vorrang einräumt". DUDEN-GWB$_2$ 1993-1995 und DUDEN-GWB$_3$ 1999 führen zusätzlich zu dieser irreführenden Definition auch das folgende Illustrationsbeispiel an, das in keinem Zusammenhang zur angegebenen Bedeutungserläuterung steht: „Schon jetzt scheint klar, dass die Brüsseler [Eurokrat]en mit ihrem Haushaltsgeld ... nicht auskommen (Spiegel 41, 1980, 22)". Dieses Illustrationsbeispiel würde eher zur Bedeutung ‚EU-Beamter/EU-Angestellter' oder gar ‚EU-Bürokrat' passen, welche auch Textkorpora des Deutschen belegen. Man vergleiche die nachfolgenden Textausschnitte:

> Plötzlich scheint ein Tabu zu brechen: [...] Ausgerechnet in Belgien, wo die imposanten Gebäude der EU-Institutionen in den Himmel ragen, wo Eurokraten still und geduldig ihre bürokratischen Fesseln über den Erdteil legen und Politiker aus ganz Europa den Abgesang aufs Nationale anstimmen. (Die ZEIT 47/2007)

Immer wieder wundern sich Politiker und Beamte in Brüssel, warum sie als Eurokraten verschrien sind. (Die ZEIT 47/2007, Wirtschaft)

Unter *EU-Erweiterung* findet man im PGWDaF 2004 die Definition: „Ausdehnung der EU-Mitgliedschaft auf weitere Länder (im Osten)". Der Zusatz „im Osten" mag den vermehrten Gebrauch des Begriffs in den Medien im Zusammenhang mit der EU-Osterweiterung in den letzten Jahren reflektieren, ignoriert aber die Tatsache, dass man auch über eine EU-Erweiterung nach Norden oder nach Süden sprechen kann (1995 traten z.B. Finnland, Österreich und Schweden der EU bei, und diese drei Länder liegen keineswegs im Osten).

Euroland wird in den untersuchten Wörterbüchern unterschiedlich behandelt: Die drei Lernerwörterbücher und WAHRIG-DW$_8$ 2008 stellen den Terminus als monosemes Lexem dar, wobei jedes Wörterbuch eine andere Bedeutungserläuterung angibt: Laut LGWDaF$_2$ 2008 versteht man unter *Euroland* „ein Land mit dem Euro als Währung" (keine Information über eine mögliche Pluralform), laut PGWDaF 2004 ist damit das „Gesamtgebiet der EU-Länder" (ohne Plural) gemeint und laut WAHRIG-GDaF$_1$ 2008 und WAHRIG-DW$_8$ 2008 ist *Euroland* die „Gesamtheit der europäischen Ländern [SIC!], in

denen die europäische Währungseinheit gilt“ (ohne Plural). DUDEN-DR$_{24}$ 2006, DUDEN-DR$_{25}$ 2009, DUDEN-GWB$_{3}$ 1999, DUDEN-UW$_{5}$ 2003 und DUDEN-UW$_{6}$ 2007 verzeichnen *Euroland* mit zwei Unterbedeutungen: „an der Europäischen Währungsunion teilnehmende Staatengruppe“ (ohne Plural) und „Staat, der an der Europäischen Währungsunion teilnimmt“. Bei dieser zweiten Unterbedeutung ist im DUDEN-UW$_{5}$ 2003 und DUDEN-UW$_{6}$ 2007 keine Pluralform angegeben. DUDEN-DR$_{24}$ 2006, DUDEN-DR$_{25}$ 2009 und DUDEN-GWB$_{3}$ 1999 listen dagegen Verwendungsbeispiele mit der Pluralform „Euroländer“ bei dieser Unterbedeutung auf. Diese Uneinheitlichkeit in der Beschreibung des Terminus führt zur Verunsicherung des Benutzers. Abhilfe können in solchen Situationen die Textkorpora des Deutschen schaffen.[6]

4 Fazit

Die vorstehende kurze Analyse der Darstellung von EU-Termini der Buchstabenstrecke *eu – ev* in ausgewählten allgemeinen einsprachigen Wörterbüchern des Deutschen brachte Lemmalücken, Bedeutungslücken, fehlende feste Wortverbindungen, sachliche Ungenauigkeiten in den semantischen Paraphrasen und uneinheitliche Angaben zu Tage. Dabei kann man keinen Unterschied zwischen den älteren und den neueren der durchgesehenen Wörterbücher ausmachen. Man kann sich nur vorstellen, was eine Durchsicht der gesamten Wörterverzeichnisse in Bezug auf EU-relevante Informationen für Funde bringen würde. Für die Zwecke der studienbegleitenden Fremdsprachenausbildung lässt sich in Sachen EU-Terminologie keines der untersuchten Wörterbücher des Deutschen empfehlen. Man sollte vielmehr verschiedene allgemeinsprachliche Wörterbücher gleichzeitig konsultieren und vor allem Korpusbelege sowie Sachlexika der deutschen Bundeszentrale für politische Bildung und/oder der Europäischen Union zum Vergleich heranziehen. Im Hinblick auf die Tatsachen, dass Deutschland zu den Gründerländern der heutigen EU gehörte, dass Österreich seit 1995 Mitglied der EU ist und dass das Deutsche offiziell als Arbeitssprache der EU gehandelt wird, auch wenn mit einem gewissen Abstand hinter dem Englischen und Französischen, erstaunt der mangelhafte Niederschlag, den die EU-Terminologie in den Wörterbüchern des Deutschen findet. In diesem Zusammenhang würde sich auch eine vergleichende Untersuchung der Darstellung der EU-Terminologie in einsprachigen Wörterbüchern des Englischen und Französischen lohnen.

5 Literatur

Brockhaus-Wahrig (1980-1984). Deutsches Wörterbuch in sechs Bänden. Hrsg. von Gerhard Wahrig, Hildegard Krämer, Harald Zimmermann. Wiesbaden/Stuttgart: F. A. Brockhaus/Deutsche Verlags-Anstalt.

6 Textkorpora des Deutschen belegen beide Bedeutungen von *Euroland* sowie eine weitere Schreibweise *Euro-Land*, die in den Wörterbüchern nicht lemmatisiert ist.

Der Duden: in 12 Bänden (242006). Die deutsche Rechtschreibung. 24., völlig neu bearbeitete und erweiterte Auflage, hg. von der Dudenredaktion. Mannheim u. a.: Dudenverlag. (= Der Duden; Bd. 1).

Der Duden: in 12 Bänden (252009). Die deutsche Rechtschreibung. 25. Auflage, hg. von der Dudenredaktion. Mannheim u. a.: Dudenverlag. (= Der Duden; Bd. 1).

Duden-GWB (11976-1981). Duden Das große Wörterbuch der deutschen Sprache. In sechs Bänden. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim u. a.: Bibliographisches Institut.

Duden-GWB (21993-1995). Duden Das große Wörterbuch der deutschen Sprache. In acht Bänden. 2., völlig neu bearb. und stark erw. Aufl. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim u. a.: Dudenverlag.

Duden-GWB (31999). Duden Das große Wörterbuch der deutschen Sprache. In zehn Bänden. 3., völlig neu bearb. und erw. Aufl. Hrsg. vom Wissenschaftlichen Rat der Dudenredaktion. Mannheim u. a.: Dudenverlag.

Duden-UW (11983). Duden Deutsches Universalwörterbuch. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim u. a.: Bibliographisches Institut.

Duden-UW (52003). Duden Deutsches Universalwörterbuch. 5., neu bearb. und erw. Aufl. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion. Mannheim u. a.: Dudenverlag.

Duden-UW (62007). Duden Deutsches Universalwörterbuch. 6. überarbeitete und erw. Aufl. Hrsg. von der Dudenredaktion. Mannheim u. a.: Dudenverlag.

Grimm, Jacob/ Grimm, Wilhelm (1999). Deutsches Wörterbuch. Neubearbeitung. 8. Band, emporerheben-exzitieren. Stuttgart: S. Hirzel Verlag.

Herberg, Dieter (2001). Euro. Zur Karriere eines europäischen Neologismus in deutschen Pressetexten (1995–1999). http://www1.ids-mannheim.de/fileadmin/lexik/LexikalischeInnovationen/pdf/euroanalysis.pdf [10/04/2014].

Knaurs-GW (1985). Knaurs Großes Wörterbuch der deutschen Sprache. Der große Störig. Erarb. v. U. Hermann unter Mitarbeit v. H. Leisering und H. Hellerer. München: Droemer Knaur.

LGWDaF (11998). Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Hrsg. von Dieter Götz, Günther Haensch, Hans Wellmann. Neubearbeitung. In der neuen deutschen Rechtschreibung. Berlin u. a.: Langenscheidt.

LGWDaF (22008). Langenscheidt Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Hg. von D. Götz/G. Haensch/H. Wellmann. Berlin u. a. Langenscheidt.

Markhardt, Heidemarie (2004). Das österreichische Deutsch im Rahmen der Europäischen Union. In: Lebende Sprachen 1, S. 15-22.

Pons GWDaF (2004). Großwörterbuch Deutsch als Fremdsprache. Barcelona u. a. Klett Sprachen.

SBH (1972). Der Sprachbrockhaus. Deutsches Bildwörterbuch. 8., völlig neubearb. u. erw. Aufl. Wiesbaden: Brockhaus.

Ullstein-LdS (1969). Ullstein Lexikon der deutschen Sprache. Wörterbuch für Rechtschreibung, Silbentrennung, Aussprache, Bedeutungen, Synonyme, Phraseologie, Etymologie. Hrsg. u. bearb. von R. Köster unter Mitarbeit v. H. Hahmann, H. Hartmann u. F. Mehling. Frankfurt am Main/Berlin: Ullstein.

Wahrig-DW (41986). Gerhard Wahrig. Deutsches Wörterbuch. Hrsg. in Zusammenarbeit mit zahlreichen Wissenschaftlern und anderen Fachleuten. Völlig überarbeitete Neuausgabe. München: Mosaik.

Wahrig-DW (72000). Gerhard Wahrig. Deutsches Wörterbuch. Mit einem „Lexikon der deutschen Sprachlehre". Neu hrsg. von Renate Wahrig-Burfeind. Gütersloh/München: Bertelsmann.

Wahrig-DW (82008). Gerhard Wahrig. Deutsches Wörterbuch. 8. , grundlegend bearb. und erweiterte Auflage, Nachdruck. Rechtschreibung 2006. Hrsg. von Renate Wahrig-Burfeind. Gütersloh: Bertelsmann.

Wahrig-GDaF (12008). Wahrig-Burfeind, Renate: Wahrig Großwörterbuch Deutsch als Fremdsprache. 1. Aufl. Gütersloh: Wissen Media Verlag.

WDG (1961-1977). Wörterbuch der deutschen Gegenwartssprache. Hrsg. von Ruth Klappenbach und Wolfgang Steinitz. 6 Bände. Berlin: Akademie-Verlag.

Wendt, Susanne (1997). Terminus – Thesaurus – Text: Theorie und Praxis von Fachbegriffsystemen und ihrer Repräsentation in Fachtexten. Tübingen: Narr (= Forum für Fachsprachen-Forschung; Bd. 37).

Text Boxes as Lexicographic Device in LSP Dictionaries

Elsabé Taljard*, Danie J. Prinsloo*, Rufus H. Gouws**
*University of Pretoria, **University of Stellenbosch
elsabe.taljard@up.ac.za, danie.prinsloo@up.ac.za, rhg@sun.ac.za

Abstract

It is the duty of the lexicographer to maximally utilise all lexicographic devices at his/her disposal in the compilation of a dictionary. In modern-day dictionaries lexicographers often rely on lexicographic text boxes as a means to present lexicographic data, especially where stronger emphasis or additional data regarding a specific entry is needed. It could be argued that the use of text boxes as a lexicographic device is underutilised in general dictionaries and even more so in LSP dictionaries. They should however not be used in a haphazard way but rather according to a well-devised system. Looking at existing dictionaries one often realises that a random use of text boxes has a detrimental influence on successful data retrieval. Text boxes should be used for data that need to be presented in a position of salience. The aim of this paper is first to give an overview of the current status of the use of text boxes in LSP dictionaries; secondly, to offer some suggestions as to the effective use of text boxes in these dictionaries by drawing on an analysis of the current use of text boxes in LGP dictionaries.
Keywords: Text boxes; LSP dictionaries; LGP dictionaries

1 Introduction

As pointed out by Gouws and Prinsloo (2010) the function of text boxes as lexicographic device is to place more than the default focus on a specific data item. They are typically used for data that need to be presented in a position of salience. One of the challenges for lexicographers is to make an informed decision as to what kind of data should be presented in textboxes in order to maximize data transfer and where to include the text box in order to ensure optimal access and a successful consultation by the user. Such a decision needs to be informed by a clear understanding of the user and his / her particular lexicographic needs and reference skills.

As pointed out by Tarp (2012), modern society is characterised by an ever-growing need for lexicographical tools that provide quick and easy access to data which have been prepared and selected in such a way as to enable the prospective user to retrieve the punctual information required to satisfy his / her lexicographic needs in a specific user situation. This need has resulted in an ever-increasing number of specialized dictionaries being published; however, many of these are of dubious quality, due to a lack of relevant lexicographic data. This paper consequently investigates the role that the in-

novative use of text boxes can potentially play in producing better quality specialized dictionaries. The aim of this paper is first to give an overview of the current status of the use of text boxes in dictionaries dealing with languages for special purposes, i.e. LSP dictionaries; secondly, to offer some suggestions as to the effective use of text boxes in these dictionaries by drawing on an analysis of the current use of text boxes in dictionaries dealing with language for general purposes, i.e. LGP dictionaries. For the purpose of this paper, the focus of the second point will be on LSP dictionaries for South African school learners. Within the current South African educational context, this will imply a perspective that includes the South African Bantu languages, since by far the majority of school learners and therefore of prospective dictionary users have a Bantu language as home language. The paper thus has both a contemplative and a transformative approach with regard to the use of text boxes.

2 The Nature of Text Boxes

Text boxes are typically included within the article structure of a dictionary, albeit that text boxes could also be included as article-external entries and immediate constituents of an article stretch. As article-internal entries they can prevail in a variety of article positions, cf. Wiegand and Gouws (2011). The article structure typically consists of the microstructural items and various indicators to identify article slots and item types. Text boxes are not part of these default constituents of dictionary articles. Like items they can be functionally positionally segmented but as another type of entry, mostly an inserted inner text. In this regard they usually are immediate constituents of the article but they could also be immediate constituents of a specific part of the article, e.g. the comment on semantics. They can be positioned where the lexicographer deems it necessary and can have any part of the article in their scope. They don't need to be addressed at an item in the article but can function as non-addressed entries. The nature of inserted inner texts gives lexicographers the freedom to utilize text boxes in different ways to contain whatever text the lexicographer regards as of enough importance to be allocated to a text box.

3 Text Boxes in LSP Dictionaries

It would seem that, generally speaking, text boxes are extremely rarely used in LSP dictionaries. Perusal of a random sample of five monolingual standard English LSP paper dictionaries revealed that none of the selected dictionaries utilizes this device. These dictionaries include the *Oxford Dictionary of Science*, *Oxford Dictionary of Chemistry*, *Oxford Dictionary of Economics*, *Concise Oxford Dictionary of Music* and *Stedman's Medical Dictionary*. One can only speculate as to what the possible reasons for this may be. It could be that the kind of information typically provided in text boxes in LGP dictionaries, is not deemed to be important by LSP lexicographers. In fact, provision of this kind of data in LSP dictionari-

es, even as items in the normal microstructure, is very rare, since these dictionaries do not normally intend to inform about issues such as grammar, pronunciation, spelling, etc. Another possible reason for the sparse provision of this kind of data could be that the dictionary has been designed by experts in the subject field in question, possibly with less than adequate lexicographic background, especially regarding article structures. In the third instance, the rudimentary treatment of comment on form could be a space-saving consideration, especially in the case of paper dictionaries. One should also consider the possibility that many changes introduced in LGP dictionaries might not have been reflected in LSP dictionaries because theoretical lexicographers have paid too little attention to LSP lexicography. LSP lexicographers might have used existing LSP dictionaries as their only model. A feature of the current development in theoretical lexicography is a stronger theoretical focus on LSP lexicography, cf. publications like Fuertes-Olivera (2010), Jesenšek (2013) and Tarp and Fuertes-Olivera (2014).This should result in positive changes in the planning and compilation of LSP dictionaries.
An informed decision as to whether text boxes could be a useful lexicographic device in LSP dictionaries can only be made if the skills, needs and knowledge of the target user and the function and genuine purpose of the dictionary are taken into consideration. When deciding on the possible use of text boxes in LSP dictionaries one should negotiate the subtypological diversity based on target users in the field of LSP lexicography. Three distinct user groups can be distinguished, i.e. experts, semi-experts and laypeople. Text boxes could be extremely helpful for users introduced into a new subject field who rely on their LSP dictionary as an important source of guidance. Consequently, guidance by means of text boxes is particularly important for users of school dictionaries. This information would also be decisive when deciding on the nature of the content of text boxes. We return to this issue below.

4 Typical Content of Text Boxes in LGP Dictionaries

As pointed out by Gouws and Prinsloo (2010), care should be taken as to what kind of data should be presented in a text box, and what data should form part of the normal microstructure of an article, since overuse of this device can detract from its usefulness. An analysis of a number of LGP dictionaries reveals that the typical data types found in text boxes include (but are not restricted to) contrasting related words, application range, pronunciation, register, grammar, spelling, collocations, metaphor, syntactic restrictions and contrasting British English and American English.
Obviously, some of the types of data provided in text boxes in LGP dictionaries are indeed irrelevant for LSP dictionaries: it is for example highly unlikely that data on metaphorical use of lemmata in an LSP dictionary would be needed, as would data on offensive use be. However, other data types can indeed be relevant for LSP dictionaries as well, and the possibility that additional data types might be necessary specifically for LSP dictionaries in specific usage situations by specific users needs to be investigated. Since the focus of our investigation is on LSP dictionaries for South African school learn-

ers, our assumption is that an analysis of the use of text boxes in LGP school dictionaries could provide useful information with regard to the potential use of this device in LSP dictionaries.

5 Use of Text Boxes in South African LGP School Dictionaries

It seems that the utilization of text boxes is gaining in popularity in the compilation of South African school dictionaries. In the most recent editions of three flagship school dictionaries, the *HAT Afrikaanse Skoolwoordeboek*, the *OXFORD Bilingual School Dictionary (English-Northern Sotho)*, and the *Oxford Bilingual School Dictionary (Afrikaans – English)* text boxes are used to provide additional data to the user. Compare the following examples from the *Oxford Bilingual School Dictionary (Afrikaans – English)* and *HAT Afrikaanse Skoolwoordeboek* in figures 1 and 2 respectively:

Figure 1: *Lekker* in *Oxford Bilingual School Dictionary (Afrikaans – English*.

Figure 2: Laf in HAT Afrikaanse Skoolwoordeboek.

Of the three dictionaries mentioned above, the *HAT Afrikaanse Skoolwoordeboek* is the only one in which the text boxes are explicitly labelled in order to provide the user with some guidance as to the nature of the data provided in the text box. Typical labels are *Tesourus* 'Thesaurus', *Gebruik* 'Use', *Spelling* 'Spelling', *Skryfwyse* 'Conjunctive and/or disjunctive writing' and *Uitdrukking* '(Fixed) expression'. This

could be regarded as labels but also as topics or titles for the text boxes. Such an approach is supported by the role of text boxes as inserted inner texts, occupying an article slot and not only a microstructural slot.

An analysis of the contents of text boxes found in the randomly selected alphabetical stretch 'L' in the *HAT Afrikaanse Skoolwoordeboek* is provided in table 1.

Related words and their meaning (Thesaurus boxes)	7
Spelling	4
Usage	2
Expression	1
Approximate no. of lemma signs in alphabetical stretch 'L'	321
Average of 1 text box per 23 lemma signs	

Table 1: Text Boxes for the Alphabetical Stretch 'L'.

Compare the results of a similar analysis for the *OXFORD Bilingual School Dictionary (English-Northern Sotho)* in table 2:

Northern Sotho - English		**English – Northern Sotho**	
Translation / non-translation of function words	2	Register	1
Composition of multiword lemmas	1	Pronunciation	4
Morphologically shortened forms	12	Contrast related words	2
Range of application	1	Correct usage	1
Discourse pragmatic information	5	Cross reference to another note	1
Offensive use	1		
Tense	1		
Additional information on part of speech	2		
Syntactic information	1		
Approximate no. of lemma signs in alphabetical stretch 'L'	416	Approximate no. of lemma signs in alphabetical stretch 'L'	178
Average of 1 text box per 16 lemma signs		Average of 1 text box per 19 lemma signs	

Table 2: Text Boxes for the Alphabetical Stretch 'L' in the *Oxford Bilingual School Dictionary*.

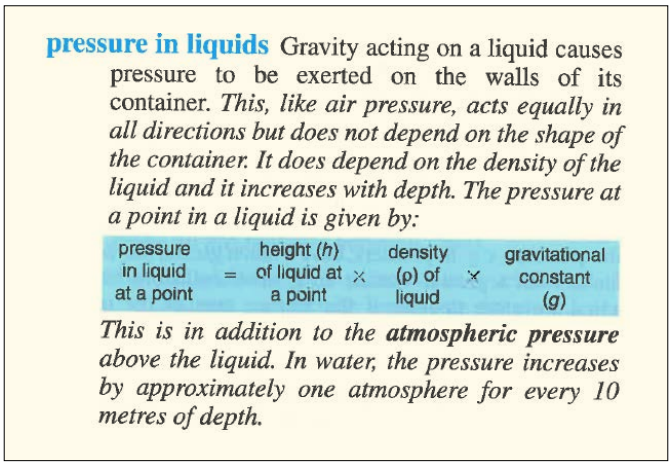
Table 3 presents the results for *The Oxford Bilingual School Dictionary (Afrikaans – English):*

Afrikaans - English		**English – Afrikaans**	
Attributive vs. predicate use of adjectives	2	Register	1
Conjunctive vs. disjunctive writing	1	Pronunciation	6
Singular vs. plural use	2	Contrast related words	1
		Cross reference to another note	1
		Correct usage	1
		Singular vs. plural use	4
Approximate no. of lemma signs in alphabetical stretch 'L'	171	Approximate no. of lemma signs in alphabetical stretch 'L'	176
Average of 1 text box per 34 lemma signs		Average of 1 text box per 12 lemma signs	

Table 3: The Oxford Bilingual School Dictionary (Afrikaans – English).

6 Text Boxes in South African LSP School Dictionaries: Some Suggestions

A logical point of departure for the LSP lexicographer would be to ascertain which of the data types provided in text boxes in LGP dictionaries could also be utilized in LSP dictionaries. Perusal of existing LSP dictionaries reveals that none of the currently available ones makes use of text boxes. A lexicographic device which seems popular is the use of shaded blocks, especially in math and science dictionaries. Formulae, equations and chemical reactions mostly make up the contents of these blocks. The data provided in these blocks are however not additional to the comments on form and semantics, but form an essential and inherent part of it. The function of these shaded blocks is mainly to act as typographical markers of a specific data type. Compare the following example from the *Oxford Physical Sciences Dictionary Grades 10-12* in figure 3:


pressure in liquids Gravity acting on a liquid causes pressure to be exerted on the walls of its container. *This, like air pressure, acts equally in all directions but does not depend on the shape of the container. It does depend on the density of the liquid and it increases with depth. The pressure at a point in a liquid is given by:*

pressure in liquid at a point = height (*h*) of liquid at a point × density (ρ) of liquid × gravitational constant (*g*)

This is in addition to the ***atmospheric pressure*** *above the liquid. In water, the pressure increases by approximately one atmosphere for every 10 metres of depth.*

Figure 3: Pressure in Liquids in Oxford Physical Sciences Dictionary Grades 10-12.

These shaded blocks contain microstructural data and are not inserted inner texts. Taking the function of text boxes as explained by Gouws and Prinsloo (2010) into consideration, there seems to be no reason why they cannot be utilized as a lexicographic device to assist with enhanced data retrieval, within an extended obligatory microstructure, from an LSP dictionary. Further motivation for the use of text boxes is the fact that they are increasingly utilized in LGP school dictionaries. It can therefore be assumed that the target user is already familiar with text boxes and knows what kind of data he / she can expect to find in them.

When deciding on the data types to be provided in text boxes for LSP dictionaries for South African school learners, care should be taken not to provide data that should form part of the default obligatory microstructure in text boxes. The fact that comments on form are generally inadequate in these dictionaries does not automatically qualify them as good candidates for presentation in a text box. Useful data on, for example, (regular) singular / plural forms, morphological derivations and pronunciation would indeed add value to LSP dictionaries as good lexicographic products, but these data categories should rather form part of the obligatory microstructure.

One data type that appears in text boxes in all three dictionaries mentioned above is information on related words, which is specifically aimed at distinguishing between words which are semantically and conceptually related. LSP dictionaries essentially deal with terminology in that the lemma signs are in actual fact terms. Apart from their function as subject specific linguistic labels for concepts in LSP, terms are often also used by laypersons as words in LGP, in which case the conceptual distinction between related concepts is often fuzzy. Words referring to related but distinct concepts are often used interchangeably, precisely because they are used by lay persons. Examples include 'temperature' and 'heat', 'weight' and 'mass', and 'power' and 'force'. It is likely that learners' first acquaintance with these concepts would have been in the non-technical sense. Using a text box in an LSP dictionary to explicate the distinction between conceptually related terms would add to the cognitive unravelling of subject specific concepts. A typical example of a text box in an LSP school dictionary would highlight the difference between the related concepts 'weight' as 'force experienced by an object due to gravity' and 'mass' as 'the amount of matter in an object'.

South African English is strongly based on British English. However, an increasing use of American English, also in subject field literature, is noticeable in South African English. American and British English do not always use the same terms. Lexicographers will do well to make provision for this situation by discussing the British and American use where applicable in the text boxes. Besides British and American English it is sometimes also necessary to refer to International English because these terms may differ from those found in both British and American English.

LSP lexicographers furthermore need to distinguish between culture-dependent and culture-independent terms. In South African dictionaries terms, e.g. in the field of medicine where terms from traditional healers still prevail, need additional treatment which falls outside the default microstructural slots. Text boxes could assist the lexicographer in this regard to ensure an optimal retrieval of information by the user.

An issue which is particularly relevant to the South African situation is standardization of terminology, especially with regard to the nine official South African Bantu languages. Official structures which are responsible for development, standardization and validation of terminology are largely non-functional, particularly those dealing with the Bantu languages. Consequently, one finds an almost unchecked proliferation of terminology, resulting in multiple terms for a single concept. Compare for example table 4:

	Northern Sotho equivalents	**Zulu equivalents**
	lerathana	inhlayiya
particle	karolonyana	intwanyana
	seripana	iphathikili
	sekgawana	

Table 4: Northern Sotho and Zulu Equivalents for *Particle*.

Apart from standard treatment of the lemma 'particle' in the dictionary, which may include listing and labelling of variant equivalent forms, data in the text box should provide the user with adequate guidance with regard to the use of standardized versus non-standardized forms and also with regard to the status of a specific term as having been standardized or not.

7 Conclusion

The use of text boxes as a lexicographic device in LSP dictionaries is under-utilized, even more so than in LGP dictionaries. Text boxes can potentially add the same value to LSP dictionaries as to LGP dictionaries, although the nature and content of these boxes will differ. The function of text boxes, i.e. to place more than the default focus on a specific data item, should be the main consideration when deciding on data inclusion in or exclusion from text boxes.

8 References

8.1 Dictionaries

HAT. Afrikaanse Skoolwoordeboek. Cape Town: Pearson Education. 2009.
Oxford Dictionary of Chemistry. Oxford: Oxford University Press. 2008
Oxford Dictionary of Economics. Oxford: Oxford University Press. 2009.
Oxford Dictionary of Music. Oxford: Oxford University Press. 2007.
Oxford Dictionary of Science. Oxford: Oxford University Press. 2010.
Oxford School Dictionary Afrikaans-Engels / English-Afrikaans. Cape Town: Oxford University Press. 2007.

Oxford School Dictionary Northern Sotho – English / English – Northern Sotho. Cape Town: Oxford University Press. 2007.
Oxford Physical Sciences Dictionary Grades 10-12. Cape Town: Oxford University Press. 2013.
Stedman's Medical Dictionary. Philadelphia, Pa.: Wolters Kluwer Health/Lippincott Williams & Wilkins. 2012.

8.2 Other References

Fuertes-Oliviera, P. (ed.) 2010. *Specialised Dictionaries for Learners*. Berlin: De Gruyter.
Gouws, R H and Prinsloo, D J. 2010. *Thinking out of the box – perspectives on the use of text boxes*. Euralex Conference proceedings, 14th Euralex Conference, Leeuwarden, 6 -10 July, 2010.
Jesenšek, V. (ed.) 2013. *Specialised Lexicography*. Berlin: De Gruyter
Tarp, S. 2012. Specialised lexicography: 20 years in slow motion. *Iberica* 24, 117-128.
Tarp, S. & Fuertes-Olivera, P. 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin: De Gruyter.
Wiegand, H.E. and Gouws, R.H. 2011. Theoriebedingte Wörterbuchformprobleme und wörterbuchformbedingte Benutzerprobleme I: Ein Beitrag zur Wörterbuchkritik und zur Erweiterung der Theorie der Wörterbuchform. *Lexikos* 21: 232-297

Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources

Izabella Thomas[1], Blandine Plaisantin Alecu[2], Bérenger Germain[3], Marie-Laure Betbeder[4]
[1]Centre L.Tesnière, Université de Franche-Comté
[2]Prolipsia, France,
[3]Share and Move Solutions, France
[4]Institut Femto-ST, Université de Franche-Comté
izabella.thomas@univ-fcomte.fr, blandine.alecu@prolipsia.com,
berenger.germain@shareandmove.fr, marie-laure.betbeder@univ-fcomte.fr

Résumé

Dans cet article nous décrivons l'architecture générale d'une plateforme web paramétrable, modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non-terminologiques : la Station Sensunique. Conçue dans l'objectif initial de faciliter et d'accélérer le processus de constitution du lexique d'une Langue Contrôlée (LC), son champ d'application peut être élargi à l'acquisition de tout type de ressources termino-ontologiques. La Station s'articule autour de deux points de vue du processus d'acquisition de ressources : (1) chronologique (centré processus) : import des textes d'entrées, analyse automatique, analyse manuelle approfondie, validation, et enfin export; (2) ergonomique (centré utilisateur-analyste) : mise en adéquation de l'analyse selon le corpus et l'application visée, visualisation et gestion des unités lexicales candidates (ULC), recherches complexes en corpus ou dans la liste d'ULC, modification ou enrichissement des descriptions ou relations des ULC, validation progressive, demande de validation par l'expert-métier, etc. La Station dispose de deux interfaces utilisateurs faciles à manipuler ; son utilisation se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées.

Mots-clés : lexique; langue contrôlée; ressource terminologique; extraction des termes; acquisition des termes; plateforme terminologique

1 Introduction

La Station Sensunique est une plateforme web paramétrable, modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non-terminologiques créée à l'Université de Franche-Comté dans le cadre du projet ANR Sensunique[1]. Conçue dans l'objectif initial d'assister le

1 Projet ANR -EMMA-2010-039 (2010-12), http://tesniere.univ-fcomte.fr/sensunique.html [08/04/2014].

processus de constitution du lexique d'une Langue Contrôlée (LC), telle que définie par (Renahy et al. 2011 ; Renahy et al. 2009 ; Vuitton et al. 2009), son objectif premier est de diminuer le temps (et donc le coût) nécessaire à la conception d'une LC. Automatiser ce processus implique deux types de contraintes : (1) liées au travail terminologique et (2) liées à la conception d'une LC (Renahy et al. 2009), à savoir recenser l'ensemble du lexique d'une LC (qu'il soit terminologique ou non), gérer les constructions particulières, notamment les structures lexicales, et respecter les principes communs à toute LC (non-ambigüité et non-redondance). Ceci présuppose la gestion de relations entre les unités lexicales (UL), qu'elles soient lexico-sémantiques (synonymie, antonymie), morphologique (flexionnelle, dérivationnelle, de variation morphosyntaxique faible, etc.) ou syntaxico-lexicales (grâce à la recherche de collocations par patterns prédictifs) (Plaisantin Alecu et al. 2012).

L'implémentation logicielle de ce processus, i.e. la Station Sensunique, automatise l'extraction d'UL candidates (ULC) à partir de corpus. Elle est configurable en finesse pour répondre aux multiples contextes d'utilisation possibles des ressources à construire, en termes de : domaines, types de textes, publics cibles des textes rédigés en une LC, ressources terminologiques préexistantes, ou plus généralement ressources linguistiques existantes et accessibles (notamment avec le courant fortement émergent des linked open linguistics data[2] (Chiarcos et al. 2012)). Elle offre aussi les fonctionnalités adéquates aux étapes suivantes du processus, à savoir, premières sélection et validation des UL par un analyste, seconde validation par l'expert métier et export de la ressource finale exploitable. Les interfaces utilisateurs (*interface de gestion* et *interface de travail*), très ergonomiques, ne nécessitent aucun savoir-faire technique et sont faciles à prendre en main et à explorer. L'application exploitant les lexiques d'une LC à concevoir (besoin initial de la Station) est un logiciel d'aide à la rédaction de textes techniques en LC sur mesure. La Station a été évaluée et validée dans ce cadre précis, sur l'intérêt du procédé de multi-extraction, implémenté dans la Station, pour le recensement du lexique d'une LC (Plaisantin Alecu et al. 2012).

Le processus métier d'acquisition du lexique d'une LC est très proche de l'acquisition de ressources termino-ontologiques (RTO) tel que décrit par Bourigault (2003). L'acquisition de dictionnaires, glossaires, lexiques, thesaurus[3] à partir de corpus doit répondre à une double contrainte de pertinence, vis-à-vis du corpus et de l'application visée, e.g. aide à la traduction, extraction d'information, indexation, etc. (Bourigault 2003).

La Station répondant à ces contraintes, son champ d'application initial (lexique d'une LC) peut être élargi à l'ensemble des RTO. Ses fondements méthodologiques et son architecture logicielle donne à la Station le potentiel d'un outil générique pouvant produire des ressources variées tout en étant fonction de l'application visée. Dans ce sens, elle suit le principe d'adéquation de Slodzian (2003) : *Qu'il s'agisse d'indexation, de mémoires de traduction bi- ou multilingues, d'aide à la rédaction de docu-*

2 Qu'il faudra toutefois télécharger puis convertir au format des ressources intégrables compatibles à la Station.

3 Pour la liste complète des RTO, voir Bourigault (2003).

ments experts, les outils proposés doivent présenter un degré d'adéquation suffisant avec le problème que l'utilisateur cherche à résoudre.

Dans la suite de cet article, nous allons tout d'abord situer notre travail (&1) dans le contexte des LC (&1.1) pour ensuite définir les besoins qui ont guidé la conception de la Station Sensunique (&1.2). Nous présenterons ensuite l'architecture générale de la Station Sensunique (&2), en mettant l'accent sur ses aspects modulaire et paramétrable. Nous décrirons les multiples possibilités de paramétrage de l'analyse automatique, en fonction du corpus d'entrée et de l'application visée. Puis nous présenterons le module de gestion qui, à partir de la liste des ULC issues du module d'analyse automatique, offre un ensemble de manipulations visant à faciliter les processus de sélection et de validation manuelle de ces ULC. Nous finirons par la présentation du module d'export qui permet également un paramétrage fin des ressources à constituer en fonction de l'application visée. Dans la conclusion, nous présenterons les possibles évolutions de la Station Sensunique.

Nous nommons « analyste » tout linguiste, terminologue, ingénieur des connaissances ou autre utilisateur de la Station et « ressources » tout type de RTO et lexique d'une LC.

1.1 Contexte

Les recherches concernant les Langues Contrôlées, sous leur multiples dénominations (langues simplifiées, langues construites, etc.) ne sont pas nouvelles, même si largement sur l'anglais : Kuhn (2013) recense plus de 100 LC conçues dans cette langue. Les travaux portant sur le français restent extrêmement rares : on peut citer l'initiative du COSLA (Comité de Simplification de la Langue Administrative ou le Français Rationalisé du GIFAS (Groupement d'Industries Françaises Aéronautiques et spatiales mais qui avait pour bases les règles du Simplified English de l'AECMA (Association Européenne des Constructeurs de Matériel aérospatial) (GIFAS, 1990).

Pour situer notre approche sur le panorama des travaux entrepris sur les LC, nous reprendrons une récente enquête sur l'ensemble des LC anglaises, dans laquelle Kuhn (2013 : 3) propose la définition suivante : *A controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics while preserving most of its natural properties.*

Nous sommes en accord avec cette définition, dans le sens où elle met l'accent sur les caractéristiques essentielles des LC telles que nous les concevons : une LC est toujours construite à partir d'une langue naturelle et en conserve les propriétés. Si l'on adopte la classification des LC proposée par Kuhn (2013), les LC que nous concevons sont de type CTWDAI, puisque : elles sont conçues dans l'objectif d'améliorer la compréhensibilité (*Comprehensibility)* ; elles augmentent la traductibilité (*Translability*) ; elles sont destinées à être écrites (*Written*) ; elles sont spécifiques à un domaine (*Domain-dependent*) ; elles sont initiées par une recherche académique (*Academic*) ; elles sont aussi industrielles (*Industrial*) dans la mesure où l'applicabilité des LC en industrie est un critère prépondérant des travaux présentés dans cet article.

Plus qu'une LC, nous cherchons à mettre en place un cadre méthodologique de conception assistée de LC *sur mesure* . Nous appelons LC sur mesure une LC reposant sur les besoins précis d'une structure particulière ayant pour objectifs l'amélioration de la qualité de son système documentaire et l'amélioration de la fiabilité de ses textes afin de diminuer les risques liés à leur mauvaise interprétation/application. Une telle LC est donc circonscrite à un domaine et à un environnement de rédaction précis, c'est-à-dire une activité précise, un public défini et un type de textes particulier (Renahy et al. 2009). Elle repose sur une analyse de corpus délimité, lequel doit recenser l'ensemble des textes en vigueur pour l'activité et le public concernés[4] (Plaisantin Alecu et al. 2012). Enfin, la LC conçue doit permettre aux personnes en charge de la rédaction technique au sein d'une structure de rédiger des documents en conformité avec les principes de cette LC.

Le cadre méthodologique d'établissement d'une LC sur mesure doit intégrer une collaboration étroite entre les linguistes et les experts du domaine et de l'activité concernés (experts métier). Il doit, de plus, prendre en considération le coût et temps de conception. Ce temps a été estimé par Jeff Allen (2005) à 5 à 10 ans dans un contexte industriel. Jusqu'à aujourd'hui, ce coût de conception était un frein à l'exploitation de LC par des structures autres que les grandes industries[5]. L'accessibilité des LC à des structures plus modestes par la diminution de leur coût de conception a orienté nos travaux sur le cadre méthodologique de conception des LC.

Les travaux présentés ici concernent le premier verrou que nous avons souhaité lever, à savoir le recensement du lexique d'une LC sur mesure. La spécificité du recensement de ce lexique par rapport à la conception de ressources terminologiques est qu'il se doit d'être exhaustif : toutes les unités lexicales nécessaires lors de l'écriture effective de documents, qu'elles soient ou non terminologiques, doivent être encodées (pour être utilisables) dans le dictionnaire de la LC. Cependant, ce critère d'exhaustivité lexicale ne doit pas permettre n'importe quel emploi des unités recensées par le futur rédacteur : l'emploi de certaines unités lexicales doit être contraint. Ceci implique de distinguer au moins deux types de dictionnaires pour chaque LC sur mesure : un dictionnaire des unités lexicales d'une LC et un dictionnaire des structures lexicales, deux notions que nous précisons par la suite.

Le développement de la Station Sensunique s'inscrit dans cet objectif précis : accélérer l'acquisition du lexique pour la construction du Lexique d'une Langue Contrôlée (LLC).

1.2 Lexique d'une Langue Contrôlée: recensement des besoins

La notion du lexique d'une LC telle que nous la considérons mérite quelques précisions dans la mesure où elle ne correspond pas à la définition du 'vocabulaire contrôlé' ('lexique contrôlé', 'termes normalisés') communément employée dans la communauté scientifique :

4 Pour donner un exemple de corpus délimité, le système documentaire du laboratoire d'immunologie avec lequel nous avons travaillé comporte environ 40 modes opératoires.

5 Une liste récente des entreprises ayant investi dans la création d'une LC est donnée par Uwe Muegge et disponible ici : http://www.tekom.de/upload/2750/tcw02_2009.pdf [03/04/2014].

Vocabulaire contrôlé (ou lexique contrôlé ou liste de termes normalisés) : Un vocabulaire contrôlé est un ensemble de termes reconnus, fixés, inaltérables, normalisés et validés par un groupe (une communauté de pratiques) utilisés pour indexer ou analyser le contenu et pour rechercher de l'information dans un domaine d'information défini (...). [6]

La première différence entre un LLC et un vocabulaire contrôlé vient de son objectif. Comme le montre la définition précédente, un vocabulaire contrôlé est dans la majorité des cas défini pour l'indexation de documents dans le but d'en faciliter la recherche. Par exemple, le MeSH considéré comme vocabulaire contrôlé (Névéol, 2004) sert à l'indexation de ressources de santé.

La deuxième différence vient de leur périmètre respectif. L'ensemble des unités composant un vocabulaire contrôlé renvoie uniquement aux termes spécifiques d'un domaine. Le LLC, quant à lui, doit permettre la rédaction d'un texte technique dans sa globalité tout en respectant l'ensemble des contraintes d'une LC. Il devra donc recenser bien plus que les termes afin de pouvoir rédiger un texte en entier. En ce sens, (Møller et al. 2006) parle de « mots » (référant alors à des unités monolexémiques comme multilexémiques) afin de ne pas confondre les unités d'un LLC avec des unités terminologiques. Nous choisissons, quant à nous, de considérer comme *unités lexicales*[7] (UL) toutes les unités d'un LLC.

Pour recenser l'ensemble du vocabulaire contenu dans une collection de textes techniques, plusieurs types de vocabulaires sont nécessaires, comme le souligne également Camlong (1996). Ensemble, ils constituent un continuum allant du vocabulaire terminologique du domaine jusqu'au vocabulaire général. En effet, pour écrire un protocole dans le domaine d'immunobiologie[8], par exemple, plusieurs types de vocabulaire sont nécessaires :

- les termes du domaine (simples et complexes) : nominaux (*anticorps monoclonaux, réactif de lyse, tampon de fixation*), verbaux (*numéroter (les cellules), centrifuger (la suspension cellulaire)*) et adjectivaux (*aneuploïde, mononucléé*) ;
- les termes d'un autre domaine (*fenêtres informatiques, répartitions gaussiennes*) ;
- les unités du lexique général: soit entrant dans la composition des termes (*anticorps de souris*) ; soit 'autonomes' (*échantillons, divers, en particulier, étude*, etc.) ; soit potentiellement ambigües, puisque possédant un sens spécifique dans le domaine traité (*solution, population* dans le domaine de l'immunobiologie, par exemple).

Une LC exige le respect, entre autres, des principes de non-ambiguïté et de non-redondance : une unité lexicale ne peut avoir qu'une seule définition ; et une définition ne peut correspondre qu'à une seule unité lexicale dans un domaine choisi. Pour être en conformité avec ces exigences, il s'avère

6 http://appui.upmf-grenoble.fr/wiki/index.php/Vocabulaire_contrôlé [03/04/2014].
7 Nous reprenons ici la notion d'unité lexicale telle que définie par L'Homme (2005).
8 Tous nos exemples se basent sur le corpus-test établi dans le domaine d'immunobiologie, composé de 14 protocoles, pour un total de 10 064 mots. Ce corpus a été soumis à des extracteurs de termes, ce qui a permis de produire une liste de 2945 unités lexicales candidates (ULC), parmi lesquelles 1512 unités lexicales ont été finalement validées par un analyste.

nécessaire de contrôler l'ensemble du lexique utilisé pour la rédaction de la documentation dans un domaine. Pour exemple, il est indispensable d'éviter d'employer le mot *solution* au sens général (*Ensemble des opérations mentales, intellectuelles susceptibles de fournir une réponse théorique ou pratiques visant à la résolution, l'analyse, la compréhension d'un problème (...), TLFi*[9]) dans les protocoles d'immunobiologie, dans lesquels *solution* prend un sens très spécifique (*Liquide formé par la dissolution d'une substance solide (p. ex. médicament) dans un solvant*, GDT[10]). Il est également nécessaire d'identifier de multiples relations entre les unités lexicales ou leurs formes telles que :

- relations morphologiques (relation flexionnelle, relation dérivationnelle, relation de variation morphosyntaxique faible, etc.) ;
- relations lexico-syntaxiques (grâce à la recherche de collocations par patterns prédictifs) ;
- relations lexico-sémantiques (par exemple, relations de synonymie, homonymie).

1.2.1 Structures lexicales d'une Langue Contrôlée

Nous introduisons la notion de structure lexicale pour répondre au critère de non-ambigüité tout en conservant le caractère exhaustif du lexique et la nécessité de restriction d'emploi selon le contexte. La notion de structure lexicale dépasse la définition d'unité lexicale à strictement parler puisqu'elle s'appuie sur la combinatoire lexico-syntaxique entre plusieurs unités lexicales, se situant ainsi à la frontière du lexique et de la syntaxe. Cette notion est à rapprocher de celles de classes de sélection distributionnelles, classes d'objets, fonctions lexicales, cadres prédicatifs, pour ne citer que quelques unes des dénominations décrivant ces types de construction dans différentes théories linguistiques. On définit une Structure Lexicale (SL) comme un patron morphosyntaxique imposé et contrôlé par un lexème, souvent prédicatif, composée d'une partie figée (lexicalisée, variable uniquement en flexion) et d'une partie variable (mais contrainte par des traits morphosyntaxiques et sémantiques). Par exemple, *marquage* est le lexème prédicatif dans *marquage des cellules, marquage des cellules leucocytaires, marquage des cellules endothéliales vasculaires animales, marquage des cellules en suspension*. Le besoin de définir des structures lexicales vient, d'une part, de l'impossibilité d'encoder ces constructions dans un dictionnaire de termes (puisque ce ne sont pas des UL) et, d'autre part, de la nécessité de contrôler leur distribution et leur variabilité dans un environnement de rédaction d'une LC. C'est pour ces raisons que nous proposons de les recenser dans un dictionnaire spécifique, sous un format décrivant leurs principales caractéristiques :

Exemple

marquage de < *NOM : CELLULE*>

La partie variable, introduite par les chevrons (<>), est généralement définie par sa catégorie fonctionnelle (ici : NOM), qui peut être en plus caractérisée par son appartenance à une classe sémantique (ici : CELLULE).

9 Trésor de la Langue Française Informatisé, atilf.atilf.fr [03/04/2014].
10 Grand Dictionnaire Terminologique, http://gdt.oqlf.gouv.qc.ca/Resultat.aspx [03/04/2014].

La notion de structure lexicale est primordiale lorsque, nous éloignant de la théorie terminologique classique, nous considérons comme termes des syntagmes autres que les syntagmes nominaux. En effet, des verbes ou des adjectifs peuvent renvoyer à des concepts bien spécifiques dans des domaines précis. Certains dictionnaires terminologiques recensent d'ores et déjà des termes de nature verbale. Par exemple, on trouve aussi bien le nom 'centrifugation' que le verbe 'centrifuger' dans Le GDT. Simplement, la description de ce verbe, en s'arrêtant à l'identification de sa catégorie verbale, ne nous renseigne ni sur la présence ni sur la nature de ses compléments : pourtant, on *centrifuge* toujours *quelque chose, du sang total, du plasma sanguin* etc. Nous proposons donc de recenser ce verbe dans un dictionnaire de structures, en indiquant clairement qu'il doit être accompagné de compléments d'une certaine classe fonctionnelle et sémantique : centrifuger <NOM : SANG> .
Un autre avantage concernant l'identification des structures lexicales est l'établissement des relations entre des UL dérivées et la vérification de la cohérence du recensement du vocabulaire. En théorie, les UL prédicatives en relation de dérivation ne peuvent introduire dans leurs structures que des compléments appartenant à des classes sémantiques identiques :

Exemple
numéroter < NOM : CELLULE> ; <NOM : CELLULE> numéroté(es) ; numération de <NOM : CELLULE>
Pour rédiger *: numération des populations leucocytaires, numéroter les lymphocytes T, B et NK*

L'avantage du recensement de ces structures est double : d'une part, cela permet de contrôler que *populations leucocytaires* et *lymphocytes T, B et NK* portent bien la contrainte sémantique *CELLULE* et que *numéroter, numération* (voire le participe passé adjectival *numéroté*) renvoient toujours à la même classe sémantique.
En résumé, le recensement du LLC implique la création de dictionnaires pour quatre types de données : les unités lexicales terminologiques, les unités lexicales non-terminologiques, les structures lexicales terminologiques et les structures lexicales non-terminologiques.

2 Architecture générale de la Station Sensunique

A notre connaissance, il n'existe pas d'outil spécifique dédié à l'aide au recensement du LLC. Par contre, il existe de nombreux outils d'extraction de termes, tâche à laquelle s'apparente l'établissement du LLC. Par conséquent, après avoir testé l'hypothèse que les extracteurs de termes peuvent aider au recensement d'un LLC (Plaisantin Alecu et al. 2012), nous avons décidé de les intégrer à la Station Sensunique.

Comme toute plateforme terminologique (par exemple : HyperTerm[11], Terminae[12], Terminus[13]), la Station intègre la mise en séquencement de plusieurs outils TAL (étiquetage, lemmatisation et extraction de termes). Sa spécificité repose sur ses autres fondements méthodologiques. Le premier est la multi-extraction ou coopération de plusieurs extracteurs. Ce procédé donne des résultats significativement meilleurs que l'utilisation d'un seul extracteur et il permet de réduire le silence et filtrer automatiquement le bruit. Plus précisément, cumuler les résultats de 3 extracteurs de termes permet de couvrir 79 % des termes (par opposition à 58% de rappel pour le meilleur extracteur), et le meilleur moyen d'aider à déterminer le statut terminologique d'une ULC est de se baser sur les résultats communs aux 2 extracteurs (Yatea et Termostat dans l'étude) avec une précision de 37 % par opposition à 28% d'un seul extracteur (Plaisantin Alecu et al. 2012). Ce procédé reprend celui des systèmes à base de vote (Fiscus 1997 ; Brunet-Manquat 2004 ; Matusov 2007 ; Serp et al. 2008), mais n'a jamais été employé avant nos travaux pour l'acquisition de ressources.

La seconde spécificité de la Station est le recoupement des résultats d'extraction avec des ressources lexicales et terminologiques existantes interrogées automatiquement. Ceci permet, d'une part, d'augmenter le potentiel terminologique d'une ULC déjà recensée comme terme dans une ressource externe, et d'autre part d'attribuer un statut non-terminologique à des ULC présentes dans les ressources lexicales intégrées à la Station.

Le dernier fondement méthodologique est le calcul de trois pondérations, en fonction de diverses informations recueillies automatiquement par la Station : (1) le Poids Terminologique (PT) ou potentiel d'une ULC à être un terme ; (2) le Poids de Structure Lexicale (PSL) ou potentiel d'une ULC à être transformée en une structure lexicale ; et (3) le Poids d'Unité Lexicale (PUL) ou potentiel d'une ULC à être une unité lexicale bien formée. Le calcul de ces pondérations organisent le travail de validation et facilitent la prise de décision et l'établissement de consensus entre plusieurs analystes ou entre l'analyste et l'expert métier.

Bien que chacun de ces procédés (multi-extraction, interrogation des ressources existantes, pondération) ne soient pas nouveaux, ils n'ont jamais été combinés, à notre connaissance, pour cumuler leurs bénéfices au sein d'une seule et même plateforme de recensement de ressources terminologiques ou non terminologiques.

La Station s'articule sur deux points de vue du processus d'acquisition de ressources : (1) chronologique (centré processus) : import des textes d'entrées, analyse automatique[14], validation, et enfin export ; (2) ergonomique (centré analyste) : mise en adéquation de l'analyse selon le corpus et l'application visée par la ressource, visualisation des ULC (fiche lexicale et contextes d'occurrence), analyse d'un groupe d'ULC (pour l'organisation du travail ou pour des actions en masse pertinentes), re-

11 http://www.tedopres.com/hyperterm-terminology-management [03/04/2014].
12 http://lipn.univ-paris13.fr/terminae/index.php/Main_Page [03/04/2014].
13 http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl [03/04/2014].
14 L'analyse automatique comprend : étiquetage, lemmatisation, racinisation, extraction des ULC, interrogation des ressources externes et internes, calcul des pondérations.

cherches complexes en corpus ou dans la liste d'ULC, modification ou enrichissement des descriptions ou relations des ULC, validation progressive, demande de validation par l'expert-métier, etc.
De cette double articulation résultent 4 modules distincts : (1) Module de configuration de l'analyse automatique, (2) Module d'analyse automatique, (3) Module de gestion des ULC et (4) Module d'export, ainsi que 3 étapes successives d'établissement du lexique (Figure 1) :

- Etape 1 : Analyse automatique, qui extrait, à partir d'un corpus textuel, une liste composée d'unités terminologiques et non-terminologiques classées en fonction de leur statut et de leur potentiel terminologique ;
- Etape 2 : Analyse manuelle approfondie, qui consiste en un premier filtrage de la liste opéré par l'analyste pour ne retenir que les unités potentiellement valables et un second filtrage réalisé avec l'aide de l'expert métier aboutissant à des ressources validées ;
- Etape 3 : Etablissement et export paramétré des ressources établies.

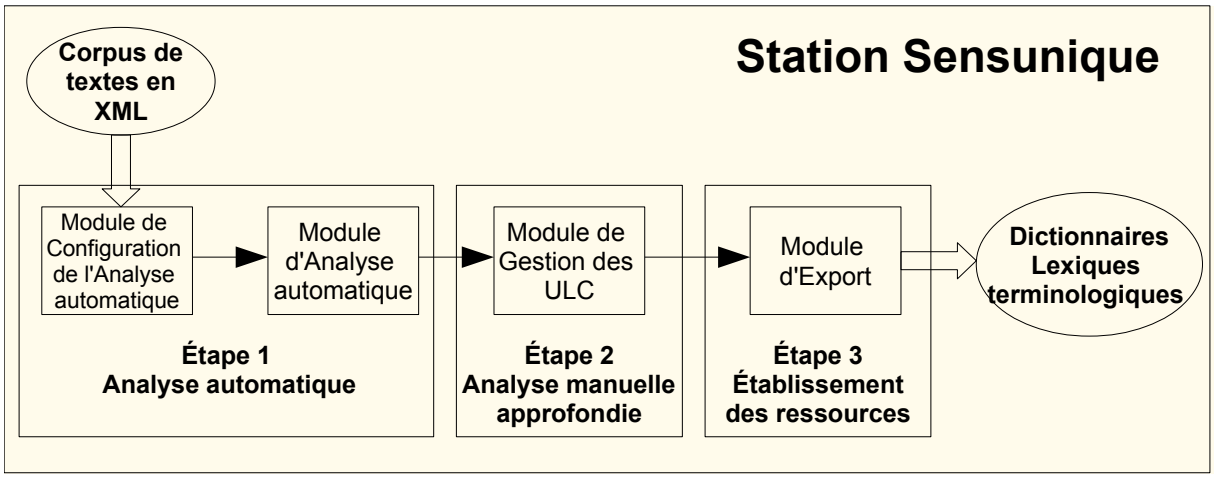


Figure 1: Schéma de la Station Sensunique.

2.1 Module Configuration de l'analyse automatique: Paramétrer l'analyse en fonction de la ressource visée

L'analyse automatique doit être configurée en fonction de l'application visée. L'analyste peut choisir ce qu'il souhaite exploiter comme types de corpus, outils, ressources et valeurs initiales de l'algorithme de pondération, selon leur adéquation au corpus et à la ressource visée. La qualité des résultats de l'analyse-extraction dépend de ces paramètres.

2.1.1 Sélection de corpus

Pour le même projet, trois types de corpus textuels[15] peuvent être simultanément analysés par la Station :

(1) le Corpus d'Analyse (CA) : c'est un corpus obligatoire duquel sont extraites les ULC à analyser ;

(2) le Corpus Support (CS) : c'est un corpus facultatif, du même domaine que le CA. En recoupant les ULC retrouvées dans les deux corpus (CA et CS), l'algorithme de pondération renforce leur potentiel terminologique. Ce procédé est inspiré de l'hypothèse de Drouin (2003) prouvant qu'une UL extraite de deux corpus différents du même domaine a plus de probabilité d'être un terme du domaine ;

(3) le Corpus Contrastif (CC) : c'est un corpus facultatif, contenant des textes généralistes, non relatifs au domaine analysé. L'exploitation d'un CC permet à l'algorithme de pondération d'augmenter la qualité des résultats en diminuant le potentiel terminologique des ULC issues du CA et du CC à la fois. De nouveau, ce procédé est inspiré de Drouin (2003) qui prouve qu'une UL extraite d'un corpus de domaine et d'un corpus généraliste a plus de probabilité d'être une unité du lexique général qu'un terme du domaine.

Les corpus sont (ré)utilisables dans plusieurs projets. En outre, un corpus n'est pas intrinsèquement lié à un statut particulier (CA, CS ou CC) : ce statut lui est attribué en fonction du projet, par un analyste. Par conséquent, le même corpus peut être utilisé comme un CA dans un projet particulier et comme un CC dans un autre projet. Ceci permet une meilleure exploitation de différents corpus constitués dans un groupe de travail ayant des projets différents.

2.1.2 Sélection des outils

Pour effectuer une analyse automatique, la Station intègre un certain nombre d'outils, à savoir :

- les étiqueteurs morphosyntaxiques : statistique Treetagger (Schmid, 1994) et à base de règles Brill[16] (Brill 1992) ; l'annotation de chaque forme fléchie du corpus par sa catégorie morphosyntaxique et ses traits morphosyntaxiques est utile non seulement à l'analyse flexionnelle et à l'extraction de termes mais également aux diverses recherches en corpus que l'analyste peut effectuer via le concordancier intégré à la Station Sensunique ;
- l'analyseur flexionnel du français Flemm v2 et v3 (Namer, 2000) : l'annotation de chaque forme fléchie du corpus par sa formé lemmatisée est utile non seulement aux extracteurs mais également aux diverses recherches en corpus que l'analyste peut effectuer via le concordancier intégré à la Station Sensunique ;
- les extracteurs de termes Acabit (Daille, 1994), TermoStat (Drouin, 2003) et YaTeA (Aubin et al., 2006) : les extracteurs de termes fournissent chacun des propositions de termes assortis d'une ma-

15 Mis au préalable au format XML TEI P5, http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_fr.html [04/04/2014].

16 Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.

trice morphosyntaxique ; de plus, Acabit regroupe des variantes du même terme ; Acabit et YaTeA découpent les termes composés en tête et expansion ; enfin, d'autres informations fournies par les extracteurs permettent de calculer certaines types de collocations (ULC incluses, composées et associées, cf. & 2.2) ;

- le racinisateur Lingua::Stem[17]: les racines ajoutées grâce à cet outil permettent d'identifier les relations dérivationnelles entre les ULC et sont également exploitées pour une recherche en corpus via le concordancier.

Les outils sont reliés en chaînes de travail indépendantes et parallèles. L'analyste peut sélectionner de 1 à 3 chaînes d'outils parmi : (1) TreeTagger - Termostat ; (2) Brill - Flemm v2 - Acabit ; (3) TreeTagger - Flemm v3 - YaTeA. Bien que la sélection d'une seule chaîne suffise pour lancer une analyse automatique, la Station est optimisée lors de l'emploi des 3 chaînes grâce au procédé de multi-extraction. Les résultats d'analyse de toutes les chaînes sélectionnées sont cumulés et recoupés et les informations obtenues affichées dans la liste des ULC résultant de l'analyse.

2.1.3 Sélection de ressources terminologiques externes (prédéfinies)

Deux ressources externes sont actuellement prédéfinies dans la Station :

- TermSciences[18], portail terminologique multidisciplinaire développé par CNRS-INIST ;
- IATE[19], base de données terminologique de l'Union Européenne.

L'interrogation automatique par web service de ces deux ressources externes permet de vérifier si une ULC proposée par les extracteurs est déjà recensée en tant que terme. Pour IATE, l'interrogation peut être restreinte à un domaine ou un sous-domaine précis (selon le référencement en domaines et sous-domaines EuroVoc[20]). Seuls les termes qui atteignent une certaine fiabilité (selon le paramètre «reliability» défini par IATE) sont retenus. Pour TermSciences, l'interrogation permet de vérifier si les constituants d'une ULC composée (sa tête ou son expansion) sont recensés indépendamment comme terme.

L'interrogation des ressources externes influe sur les pondérations, en renforçant le potentiel terminologique d'une ULC attestée dans une (ou plusieurs) ressource(s), renforcement plus ou moins fort selon si l'ULC est attestée dans sa globalité, ou si sa tête et /ou son expansion sont attestés. Elle permet ainsi de structurer le processus de validation des ULC. De plus, elle participe à l'enrichissement des informations rattachées à chaque ULC, puisque sont importées dans la Station des informations supplémentaires telles que définitions, synonymes et classes sémantiques/conceptuelles auxquelles appartient le terme attesté.

L'analyste peut choisir d'intégrer ou non l'interrogation automatique des ressources à l'analyse.

17 http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm [04/12/2011].
18 http://www.termsciences.fr/ [03/04/2014].
19 http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load [03/04/2011].
20 eurovoc.europa.eu [03/04/2011].

2.1.4 Intégration de nouvelles ressources (dites internes)

En plus de ressources externes prédéfinies, la Station permet d'intégrer à chaque nouveau projet d'autres ressources spécifiques, moyennant leur mise au format prédéfini dans la Station. Il peut s'agir aussi bien de ressources terminologiques (e.g. des dictionnaires spécialisés) qui augmentent le potentiel terminologique des ULC, que des ressources non-terminologiques (e.g. Morphalou 2.0[21]) qui augmentent le poids d'unité lexicale d'une ULC tout en diminuant son potentiel terminologique. Par ailleurs, des ressources constituées au préalable dans la Station, résultant d'autres projets, peuvent aussi être intégrées en tant que ressources internes.

Du fait de l'intégration dynamique des ressources, la Station peut être considérée comme évolutive, puisque chaque analyse peut être enrichie grâce à un ensemble de ressources spécifiques et appropriées.

2.1.5 Paramétrage des pondérations

Trois pondérations servent à faire ressortir la fiabilité des ULC et à les classer en vue d'organiser le travail de filtrage et de validation :

(1) Poids Terminologique (PT) : potentiel terminologique d'une ULC calculé selon 7 critères :

- le nombre des extracteurs ayant proposé l'ULC ;
- le seuil du statut terminologique, c'est-à-dire la valeur à partir de laquelle les ULC sont considérées comme termes ;
- présence dans le CS ou le CC (cf &2.1.1) ;
- le nombre des ressources choisies ayant attesté l'ULC ;
- le type d'attestation dans une ressource (l'attestation d'ULC globale ayant plus de poids que l'attestation de la tête et/ou l'expansion seulement) ;
- la fiabilité de la ressource externe (TermSciences ou IATE)[22] dans le domaine analysé ;
- et la présence d'une ULC dans une ressource terminologique interne (cf. & 2.1.4).

(2) Poids de Structure Lexicale (PSL) : potentiel d'une ULC à être transformée en une structure lexicale, calculé selon 8 critères dont :

- l'attestation d'une ULC globale dans une ressource terminologique (qui influe négativement sur sa possibilité d'être une structure lexicale) ;
- la matrice morphosyntaxique d'une ULC (les verbes et les participes ayant plus de probabilité de constituer les structures lexicales) ;
- le nombre de dérivées et/ou de collocations construites autour d'une ULC.

21 Lexique de formes fléchies du français développé par ATILF, http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php [03/04/2014].

22 Estimée par l'analyste.

(3) Poids d'Unité Lexicale (PUL) : potentiel d'une ULC à être une unité lexicale bien formée, calculé selon 2 critères :

- le nombre d'extracteurs l'ayant proposé ;
- la présence d'une ULC dans une ressource interne non-terminologique (cf. & 2.1.4).

A chacun de ces critères correspond une valeur jouant dans le calcul global de chacune des 3 pondérations. Des valeurs préexistent par défaut, mais sont ajustables par l'utilisateur.

2.2 Module d'Analyse automatique

Ce module a deux fonctions. La première fonction est d'annoter linguistiquement le corpus d'analyse par incorporation des résultats des étiqueteurs, lemmatiseurs et racinisateur intégrés. Sa deuxième fonction est d'extraire de ce corpus des ULC (par multi-extraction), de les décrire (résultat des extracteurs et de l'interrogation des ressources définies) et de les pondérer (résultat de l'algorithme de pondération de la Station).

Les informations issues de l'analyse automatique sont, pour chaque ULC :

- Forme canonique : correspond, la plupart de temps, à la suite de lemmes de chaque élément d'une ULC, ex. *membrane cellulaire* ;
- Statut lexical : terminologique ou non, selon le seuil du PT paramétré par l'analyste ;
  > Domaine(s) (uniquement si le statut est terminologique ; correspond dans ce cas au domaine renseigné par l'analyste dans le descriptif du projet ; ex. *immunobiologie*) ;
- Usage : "préconisé" ou "interdit", selon les spécifications d'une LC ;
- Catégorie(s) sémantique(s) : proposée(s) par les ressources externes (ex. *Structures cellulaires,* d'après TermSciences) ;
- Fréquence : nombre d'occurrences des formes fléchies de l'ULC en corpus ;
- Indices de confiance :
  > Pondérations internes : PT, PSL, PUL (cf & 2.1.5) ;
  > Indices des extracteurs externes : indices de confiance fournis par les extracteurs, ex. *loglike* pour Acabit ;
- Tête : régisseur syntaxique d'une ULC, ex. *membrane* ;
- Expansion : complément/modifieur d'une Tête, ex. *cellulaire* ;
- Catégorie morphosyntaxique fonctionnelle: en général, catégorie de la Tête d'une ULC, ex. *NOM* ;
- Matrice morphosyntaxique : suite des catégories morphosyntaxiques de chaque élément de l'ULC., ex. *Adj Nom* ;
- Formes fléchies : si trouvées en corpus, assorties des traits morphosyntaxiques et fréquence ;
- Variantes : provenant soit du corpus analysé, soit des ressources externes, ex. *membrane plasmique* ;
- ULC dérivées: ULC dont un des composants appartient à la même famille dérivationnelle, ex. *membrane cellulaire, marquage de cellule* ;

- ULC homonymes: ULC homographes d'une autre catégorie morphosyntaxique que l'ULC analysée;
- Collocations (ULC liées) ;
  > ULC incluses : une ULC incluse est une ULC dont l'intégralité se retrouve dans l'ULC analysée ; par exemple, pour l'ULC *anticorps monoclonal de souris*, les ULC incluses sont : *anticorps monoclonal, anticorps* ;
  > ULC composées : une UL composée est une ULC contenant plus que l'intégralité de la ULC analysée ; par exemple pour l'ULC *anticorps monoclonal*, les ULC composées sont *anticorps monoclonal conjugué, anticorps monoclonal de souris, anticorps monoclonal HLA-B27* [23];
  > ULC associées : une ULC associée est une ULC non incluse et non composée contenant un même lemme que l'ULC analysée ; exemple : pour l'ULC *anticorps monoclonaux*, ULC associée est *solution d'anticorps* ;
- Sources :
  > Outil(s) ayant proposée une ULC (exemple : Termostat, Acabit) ;
  > Ressource(s) externe(s) l'attestant (exemple : TermSciences) ;
- Définition(s) (provenant de ressources externes).

Partant du principe que chaque proposition faite lors d'une analyse automatique peut être modifiée, tous les résultats du module d'analyse (excepté les indices de confiance calculés par les extracteurs et les sources) sont éditables dans le module de Gestion des ULC.

2.3 Module de Gestion des ULC: Faciliter le processus de sélection et de validation

Le module de gestion des ULC rassemble des fonctionnalités facilitant la seconde phase du processus d'acquisition des ressources, à savoir l'analyse manuelle approfondie. Elle consiste en un premier filtrage des ULC par un analyste et en l'établissement du consensus final avec les experts métier (Fig.1). Le parti pris fondamental de la Station est que l'analyste peut effectuer tout changement nécessaire concernant l'ensemble de résultats proposés par l'analyse automatique. Un espace dédié, appelé *interface de travail* lui sert à visualiser, à approfondir et à élargir (si besoin) les résultats afin de les valider pour construire la ressource finale.

Dans l'interface de travail, les résultats de l'analyse automatique peuvent être visualisés sous 3 modes :

- liste des ULC contenant des informations utiles pour trier et filtrer les résultats ;
- fiche lexicale de chaque ULC détaillant toutes les informations ;

23 Les UL incluses et composées fonctionnent de manière symétrique : si une ULC1 est ULC incluse d'une ULC2, alors l'ULC2 sera ULC composée de l'ULC1.

- fiches de relations, détaillant l'ensemble de relations entre l'ULC analysée et d'autres ULC (telles que variantes, collocations, homonymes, ULC appartenant à la même famille dérivationnelle).

L'analyste peut ajouter, modifier, compléter, valider ou supprimer toute ULC ou information à partir d'un mode de visualisation approprié. Chaque proposition/modification de données est toujours tracée, c'est-à-dire, assortie du nom de son auteur (qu'il soit analyste, outil ou ressource).
Ce module réunit également des fonctionnalités d'exploration (des ULC et de leurs informations descriptives) et d'aide à la décision (aux rejet, modification, enrichissement, validation):

- **tri et filtre** sur la liste des ULC selon 21 paramètres différents, dont fréquence, PT, PUL, extracteur(s) d'origine, ressources attestant l'ULC, matrice morphosyntaxique, catégorie sémantique etc. ; les filtres sont cumulatifs, c'est-à-dire qu'on peut filtrer les ULC selon plusieurs paramètres à la fois (par exemple, ULC proposées par Termostat, ayant atteint un certain seuil de PT et d'une matrice morphosyntaxique particulière) ;
- **projection** pour visualiser une ou plusieurs ULC en contexte d'origine (en corpus ou par phrases) ;
- **regroupement** d'ULC dans les fiches de relations ; certaines ULC sont regroupées automatiquement, mais l'analyste peut aussi établir de nouvelles relations ;
- **concordancier évolué** offrant différents types de recherche sur le corpus[24] : (a) simple : sur une chaîne de caractères ; (b) morphologique simple : sur un (ou une suite de) lemme(s) permettant d'identifier toutes ses formes fléchies d'une ULC ; (c) morphologique complexe : sur un (ou une suite de) radical(aux) permettant d'identifier les familles dérivationnelles ; (d) morphosyntaxique : sur une suite d'étiquettes morphosyntaxiques ; (e) recherche dite combinée permettant de coupler les types de recherches précédents. Combiner des critères appartenant à différents niveaux d'analyse linguistique permet d'imposer des contraintes plus ou moins fortes sur les motifs recherchés, et ainsi cibler ou, au contraire, élargir le champ des résultats. Par exemple, la recherche '[e]Nom [c] de [l] cellule' (exprimée sous forme d'expression régulière Sensunique) permet de cibler les groupes dont le premier élément est le Nom suivi de la préposition 'de' et d'une forme fléchie du mot 'cellule' (ex. *nombre de cellules, greffon de cellules, analyse de cellules* etc.).

L'établissement des SL se fait manuellement, à partir du regroupement de plusieurs ULC. La fonctionnalité de dégradation permet de définir une nouvelle SL (et ses différentes informations associées, telles que statut lexical, catégorie sémantique, catégorie fonctionnelle etc.) et de l'ajouter à une liste des SL. Les opérations de tri et de filtrage peuvent être effectuées sur la liste des SL comme sur la liste des ULC.
Enfin, 7 statuts de validation, correspondant à différentes étapes d'analyse ('Non validé', 'En cours d'analyse', 'A valider par les experts', 'Invalidée par les experts', 'Validé par les experts', 'Validée', 'Invalidé') permettent de suivre le processus d'établissement du lexique.

24 Sous forme d'Expressions Régulières (selon http://fr2.php.net/manual/fr/book.pcre.php) adaptées à la Station Sensunique.

2.4 Module d'Export: Paramétrer les ressources produites en fonction d'une application

Ce module permet d'exporter en dictionnaires les données recensées dans la station au format XML afin de :

- créer des ressources terminologiques diverses ;
- exploiter les données dans d'autres applications ;
- durant l'analyse, valider les données nécessitant des compétences spécifiques par des experts métiers.

En fonction de son objectif, l'utilisateur peut paramétrer les dictionnaires de sortie, en choisissant le(s) type(s) d'informations qu'il souhaite exporter. Toute la finesse de description d'une ressource produite dans la Station n'est pas forcément utile à l'application qui va exploiter cette ressource. De même, on peut n'être intéressé que par un périmètre restreint des UL recensées.
La sélection s'effectue à l'aide des filtres cumulatifs servant à restreindre le périmètre des données exportées selon deux axes :

- sélection des propriétés des ULC (parmi les 17 propriétés proposées, telles que définition, synonymes, matrice morphosyntaxique, catégorie sémantique, colocations, statut de validation, etc.) :

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Définition et Variantes
Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Matrice morphosyntaxique et Fréquence

- sélection des propriétés des ULC et des valeurs de propriétés :

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Classe Sémantique, Définition, Statut de Validation ; ET le Statut de Validation est « Validée »

Le même projet permet de créer plusieurs ressources en fonction d'une application visée. Le principe est le même pour les dictionnaires de SL.

3 Conclusion

Conçue dans l'objectif d'optimiser (en termes de qualité et de coût) l'acquisition du lexique d'une langue contrôlée, les possibilités d'exploitation de la Station Sensunique dépassent considérablement ce champ d'action. En effet, l'éventail des configurations d'analyse (choix des outils et ressources, intégration de nouvelles ressources, personnalisation des bases de calcul des pondérations, paramétrage de l'export) en fonction de nombreux contextes d'utilisation, fait d'elle non seulement un outil d'acquisition du lexique d'une langue contrôlée, mais aussi une plateforme pertinente pour tout travail de constitution de RTO à partir de corpus.

Sur le plan méthodologique, la multi-extraction permet à la Station Sensunique d'offrir à ses utilisateurs les points forts de chaque extracteur de termes. Renforcée par l'interrogation des ressources existantes et par le principe des 3 types de corpus, la Station pondère ses résultats et permet ainsi d'organiser le processus de validation des ULC. L'interrogation des ressources existantes permet d'enrichir automatiquement la description morphologique, syntaxique et sémantique des ULC. Par ailleurs, la Station est conçue pour respecter et faciliter le processus métier d'acquisition de ressources : elle prend en compte les différentes phases de ce processus et modélise l'implication de plusieurs acteurs, y compris la validation finale par un expert-métier. De plus, l'utilisation de la Station se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station. Enfin, la Station Sensunique est dotée d'une interface utilisateur facile à manier et à explorer.

En ce qui concerne les futurs développements de la Station, plusieurs directions sont envisagées. Premièrement, nous considérons l'ajout d'autres chaînes d'outils ou le développement d'outils propres pour améliorer les performances de la Station, ainsi que l'intégration d'autres ressources externes, bien que ceci pose problème concernant les licences d'utilisation des fois difficiles à obtenir : d'où l'importance d'interagir avec les courants tels que linked open data. Deuxièmement, nous souhaitons améliorer le traitement du contenu sémantique des textes, principalement la détection des relations sémantiques et conceptuelles entre les unités lexicales. Une autre direction de recherche est l'exploitation de la plateforme pour la construction d'ontologies de domaine.

4 Références bibliographiques

Allen J. (2005). How are we responding to industrial and business needs for Controlled Language and Machine Translation, *Journées Linguistiques – Langues contrôlées, traduction automatique et langues spécialisées : 5-6 May 2004 Besançon, France*, http://web.science.mq.edu.au/~rolfs/controlled-natural-languages/papers/Jeff-Allen.pdf [08/04/2014].

Aubin S. et Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing*, 5th International Conference on NLP (FinTAL'2006), Springer, 2006, p. 380-387.

Brill E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing* (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.

Camlong (1996). Méthode d'analyse lexicale textuelle et discursive, Paris, Orphrys.

Bourigault, D., & Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues* (pp. 27-50).

Brunet-Manquat F. (2004). Fusionner pour mieux analyser : Conception et évaluation de la plate-forme de combinaison. In *Actes de TALN-2004*. Fez, Maroc, 19-22 avril 2004. vol. 1/1, pp. 111-120.

Chiarcos Ch., Hellmann S. and Nordhoff S. (2012). Linking linguistic resources: Examples from the Open Linguistics Working Group, In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*, Springer, Heidelberg, p. 201-216.

Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Workshop at the 32nd Annual Meeting of the ACL (ACL'94), Las Cruces, New Mexico, USA.

Drouin P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. In *Terminology*, vol.9, n°1, John Benjamins Publishing Company: Amsterdam/Philadelphia, p. 99-115.

Fiscus J.G.(1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), Automatic Speech Recognition and Understanding, 1997. In *Proceedings IEEE Workshop*, pp.347-354.

GIFAS, Guide du rédacteur. Groupement des Industries Françaises Aéronautiques et Spatiales, Paris, France, 1990.

Kuhn T. (2013) A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language and Information*, 22(1).

Kuhn T. (2014) A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1), 2014.

L'Homme M-C. (2005). Sur la notion de terme. In Meta: journal des traducteurs / Meta: Translators' Journal, vol. 50, n° 4, p. 1112-1132. http://id.erudit.org/iderudit/012064.

Matusov E. et al. (2007). System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1222–237.

Møller M.H., Christoffersen E., Hansen M. (2006). Building a Controlled Language Lexicon for Danish. In *LSP and Professional Communication*, vol. 6, Nr. 1, p. 12-38.

Namer, F. (2000). FLEMM: un analyseur flexionnel du français à base de règles. In *Traitement Automatique des Langues*; vol. 41/2, p. 523-547.

Névéol A. (2004). Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. In *RECITAL 2004*, Fès.

Plaisantin Alecu B., *Thomas I., Renahy J.* (2012). *La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques*. In Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, ATALA/AFCP, pp.511 518, http://www.aclweb.org/anthology/F/F12/F12-2047.

Renahy J., Devitre D., Thomas I., Dziadkiewicz A. (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability. In *Proceedings of the 11th International Symposium on Social Communication*, Santiago de Cuba, Cuba, 19-23 January 2009, pp. 289-293.

Renahy J., Thomas I., Chippeaux G., Germain B., Petiaux X., Rath B., De Grivel V., Cardey S., Vuitton DA. (2011). La langue contrôlée et l'informatisation de son utilisation au service de la qualité des textes médicaux et de la sécurité dans le domaine de la santé. In P. Staccini, A. Harmel, S. Darmoni, R. Gouider, *Systèmes d'information pour l'amélioration de la qualité en santé*, Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM'2011), Tunis, 23-24 septembre 2011, Springer-Verlag.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (Vol. 12, pp. 44-49).

Thomas I., Betbeder M.L., Renahy J., Vuitton DA.(2012). *Optimisation d'un logiciel pour la rédaction de textes techniques de qualité : application-pilote au domaine de la santé*. Projet ANR -EMMA-2010-039 (2010-12), rapport final (non-publié).

Serp C., Cazal E., Laurent A., Roche M. (2008).TERVOTIQ : un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval. In *9èmes journées internationales d'analyse statistique de données textuelles (JADT'2008),* Lyon, 2008.

Slodzian, M. (2000). L'émergence d'une terminologie textuelle et le retour du sens. Le sens en terminologie, 61-85.

Vuitton DA., Aishan A., Renahy J., Jin G., Wu X., De Grivel V., Cardey S. (2009). Controlled language: a Linguistic Concept to Improve Health Care Safety in a "Globalised" World? Application to Medical Proto-

cols Written within the Hospital Accreditation/Certification Framework in France and China. In *ISMTCL Proceedings, International Review BULAG*, PUFC, ISBN 978-2-84867-261-8, pp. 260-268.

Remerciements

Nos travaux ont été financés par l'Agence Nationale de la Recherche, programme Emergence 2010. Nous remercions toute l'équipe du projet Sensunique, les auteurs des outils intégrés et les organismes gérant les ressources terminologiques prédéfinies dans la Station Sensunique.

Station Sensunique: une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques (orientée Langues Contrôlées)

Izabella Thomas[1], Blandine Plaisantin Alecu[2], Bérenger Germain[3], Marie-Laure Betbeder[4]
[1]Centre L. Tesnière, Université de Franche-Comté
[2]Prolipsia, France
[3]Share and Move Solutions, France
[4]Institut Femto-ST, Université de Franche-Comté
izabella.thomas@univ-fcomte.fr, blandine.alecu@prolipsia.com,
berenger.germain@shareandmove.fr, marie-laure.betbeder@univ-fcomte.fr

Résumé

Dans cet article, nous présentons le fonctionnement et les services proposés par la Station Sensunique, une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et du lexique d'une Langue Contrôlée. La Station prend en charge et facilite l'ensemble de ce processus, à travers une analyse automatique du corpus, puis la possibilité d'approfondir l'analyse manuellement, pour aboutir à la création des fiches terminologiques/lexicales et des ressources exportables. La Station est un outil prêt à l'emploi, ergonomique, facile à prendre en main et à exploiter à travers ses différentes interfaces. Elle allie les avantages d'une analyse automatique (rapidité, coût) avec l'exactitude et la fiabilité d'une analyse humaine.

Mots-clés: lexique; langue contrôlée; ressource terminologique; extraction des termes; acquisition des termes; plateforme terminologique

1 Description générale

La Station Sensunique est une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques. Elle a été conçue à l'Université de Franche-Comté durant le projet ANR-EMMA-2010-039 intitulé Sensunique[1] (2011-2012) dans l'objectif d'accélérer le processus d'établissement du lexique d'un domaine ou d'une Langue Contrôlée (LC). Elle prend en entrée un corpus de textes et produit en sortie des diverses ressources (dictionnaires, lexiques, glossaires) enrichies de multiples informations linguistiques. Elle s'appuie sur une analyse automatisée de corpus, dont les résultats sont la base d'une phase de validation manuelle effectuée par un analyste, puis par un expert d'un domaine. La spécificité de cette station par rapport à d'autres

1 http://tesniere.univ-fcomte.fr/sensunique.html [08/04/2014].

plateformes de travail terminologique (HyperTerm[2], Terminae[3], Terminus[4]), repose (Thomas et al. 2014) :

- d'une part, sur les choix méthodologiques sur lesquels elle est fondée : la collaboration de plusieurs outils TAL (Plaisantin Alecu et al. 2012), l'interrogation automatique des ressources terminologiques existantes, l'intégration et l'interrogation des ressources terminologiques ou lexicales propres; ceci en vue de faciliter le travail de l'analyste en lui proposant une liste d'Unités Lexicales Candidates (ULC) pondérées et enrichies de multiples informations linguistiques acquises automatiquement ;
- d'autre part, sur les objectifs spécifiques desquels elle découle, dont notamment le recensement du Lexique d'une Langue Contrôlée (LLC), définie comme une *Langue Contrôlée sur mesure* (Renahy et al. 2009, 2011).

La spécificité d'un tel lexique est qu'il se doit d'être exhaustif : toutes les unités nécessaires lors de l'écriture de documents, qu'elles soient ou non terminologiques, doivent être encodées (pour être utilisables) dans le dictionnaire d'une LC. De plus, cette contrainte d'exhaustivité du niveau lexical d'une LC implique de distinguer au moins deux types de dictionnaires : un dictionnaire du lexique d'une LC et un dictionnaire des structures lexicales (Thomas et al. 2014), chacun pouvant être soit terminologique soit général. Une autre contrainte liée à la conception de LLC provient des principes d'une LC : de non-ambigüité (à une Unité Lexicale (UL) ne correspond qu'un sens) et, inversement, de non-redondance (à un sens correspond une et une seule UL). Ceci présuppose la gestion de la synonymie, et plus généralement la gestion des relations entre plusieurs unités lexicales (telles que homonymie, dérivation, collocations etc.).

La station est orientée analyste – utilisateur. Tous les résultats (ULC ou informations associées) sont des propositions que l'utilisateur peut modifié (ajouter, modifier, compléter, valider ou invalider). Il est assisté dans ce processus par un ensemble de fonctionnalités d'exploration des résultats, à savoir : visualisation des ULC sous plusieurs modes, tris et filtres sur la liste des ULC, projection des ULC sur le corpus d'origine, regroupement de différentes ULC, recherches sur le corpus à l'aide d'un concordancier avancé, etc.

L'utilisation de la Station se fait, sans contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station. Enfin, la Station Sensunique est dotée d'une interface utilisateur facile à manier et à explorer.

2 http://www.tedopres.com/hyperterm-terminology-management [08/04/2014].
3 http://lipn.univ-paris13.fr/terminae/index.php/Main_Page [08/04/2014]
4 http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl [08/04/2014].

2 Architecture et services

La station Sensunique fonctionne de façon modulaire, chaque module proposant à l'utilisateur plusieurs services. Les modules sont organisés pour correspondre au processus d'acquisition de ressources, divisé en plusieurs étapes (représenté par la Figure 1).

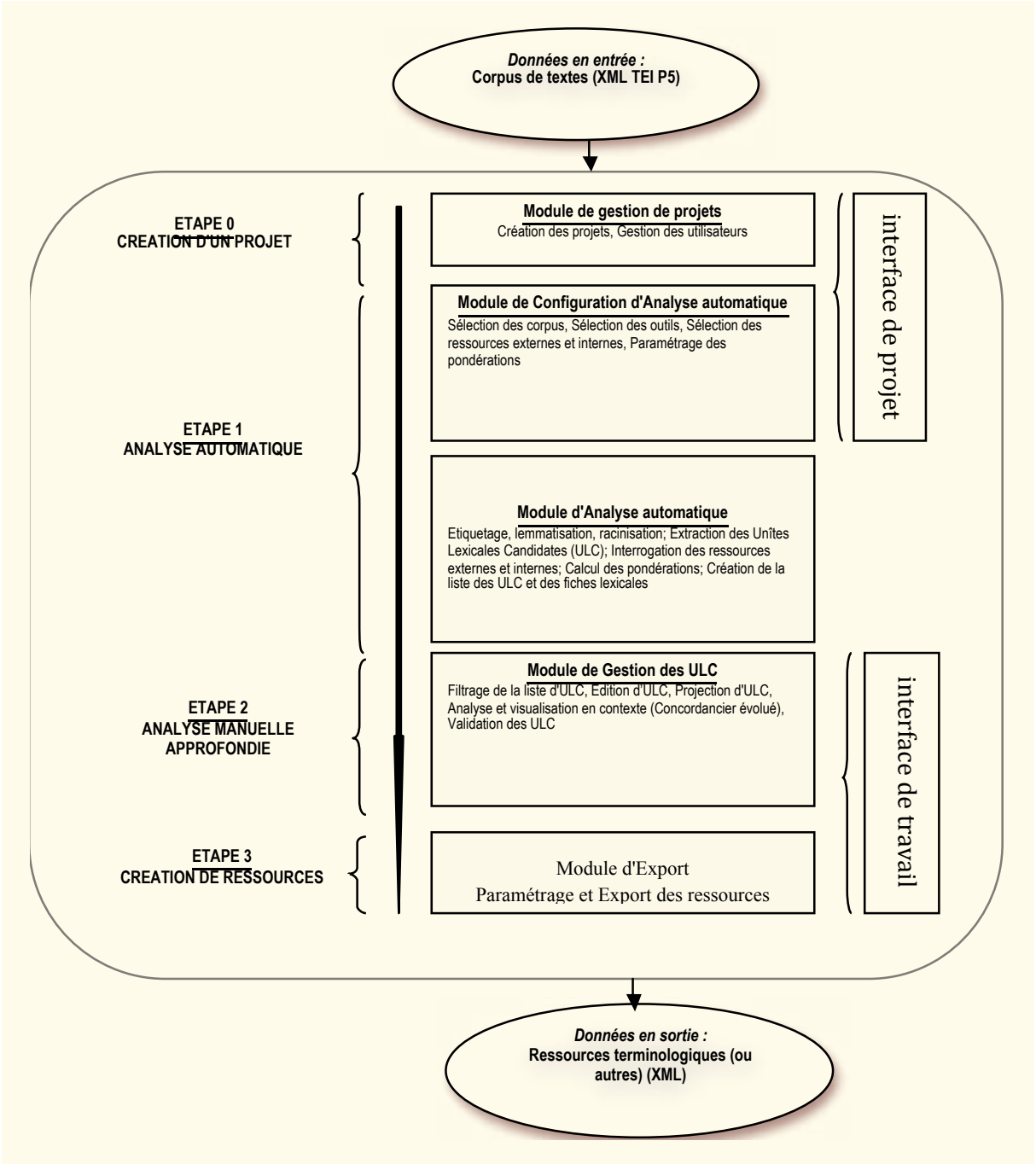


Figure 1: Schéma de la Station Sensunique.

L'utilisateur peut interagir avec la Station à partir de deux interfaces : (1) *l'interface de projet*, qui sert à définir et paramétrer un projet en termes d'utilisateurs, de corpus, d'outils et de ressources utilisés pour l'analyse automatique et (2) *l'interface de travail*, qui permet d'explorer les résultats d'analyse automatique en vue de l'établissement d'une ressource finale. Nous présentons les diverses fonctionnalités de la Station en suivant le processus chronologique d'un utilisateur souhaitant implémenter un nouveau projet.

2.1 Etape 0: Création d'un projet

La création d'un projet commence par l'établissement d'un groupe de travail, c'est-à-dire par la déclaration d'un ou plusieurs utilisateurs ayant le droit de travailler sur le projet. En effet, la Station est collaborative : elle permet à plusieurs utilisateurs d'interagir sur la même tâche. Elle assure aussi la traçabilité de toute modification (correction, modification ou complétion des données) grâce à une étiquette portant le nom de l'utilisateur concerné. Ces étiquettes de traçabilité permettent également de distinguer les données obtenues de façon automatique des données créées ou modifiées par un utilisateur.

Un groupe de travail peut créer plusieurs projets (Figure 2) ; chaque projet, en plus du nom et de sa date de création, est caractérisé par son domaine et le public auquel il est destiné. Un projet ne peut contenir qu'un corpus pour chaque type de corpus permis : Corpus d'Analyse, Corpus Support, Corpus Contrastif (Thomas et al., 2014). Les corpus doivent être chargés dans la Station au format XML TEI P5[5]; la conversion de tout document vers ce format doit être faite au préalable en utilisant, par exemple Oxgarage[6], un convertisseur automatique de format de documents en ligne.

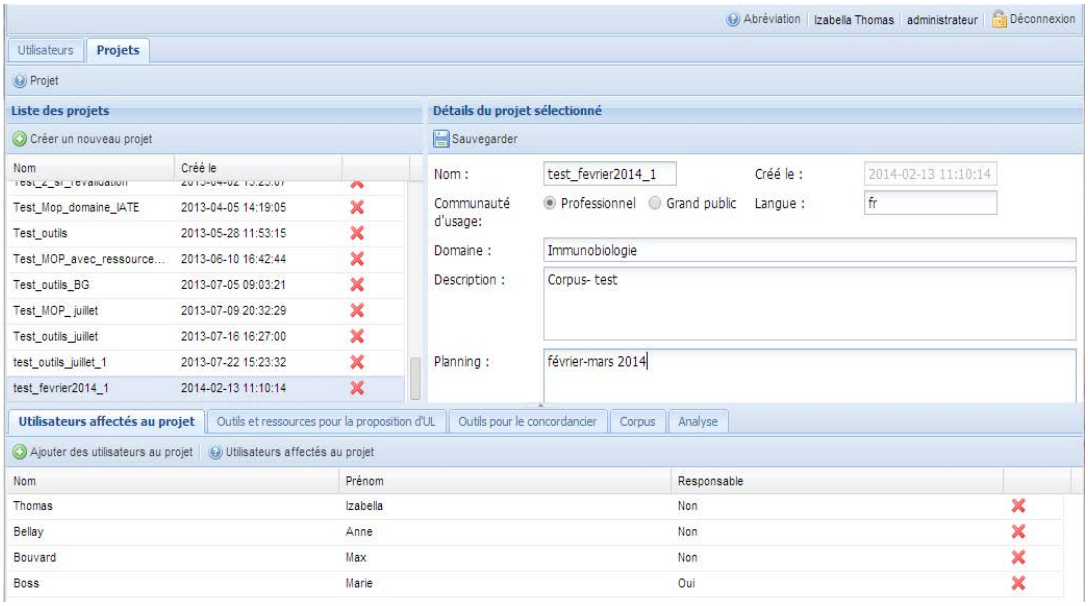


Figure 2: Déclaration d'un projet (capture d'écran).

5 http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_fr.html [04/04/2014].
6 http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient [accédé le 08/04/2014].

2.2 Etape 1: Analyse automatique

Cette étape est composée de deux sous-étapes : (1) la configuration de l'analyse automatique par l'utilisateur et (2) l'analyse automatique.

La Station Sensunique est hautement paramétrable (Thomas et al. 2014), dans l'objectif d'assurer l'adéquation de l'analyse avec la ressource à construire. Les paramétrages se font à partir de l'interface de projet (Figure 3): en fonction de son objectif, l'utilisateur peut choisir les chaînes d'outils et les ressources externes à interroger, paramétrer l'algorithme de pondération et incorporer des ressources internes (moyennant leur mise en format appropriée). Le choix de ressources internes n'est pas restreint, ce qui assure à la Station son caractère évolutif.

La qualité de l'analyse et le nombre d'informations recueillies par la Station Sensunique à partir de diverses ressources dépend du paramétrage effectué par l'utilisateur.

Actuellement, les outils et les ressources intégrés à la Station Sensunique sont les suivants :

- étiqueteurs morphosyntaxiques : TreeTagger[7] et Brill[8] (Brill 1992) ;
- analyseur flexionnel du français : Flemm v2 et v3 (Namer, 2000) ;
- extracteurs de termes : Acabit (Daille 1994), TermoStat (Drouin 2003) et YaTeA (Aubin et al. 2006) ;
- racinisateur Lingua:: Stem[9] ;
- TermSciences[10], portail terminologique multidisciplinaire développé par CNRS-INIST ;
- IATE[11], base de données terminologique de l'Union Européenne.

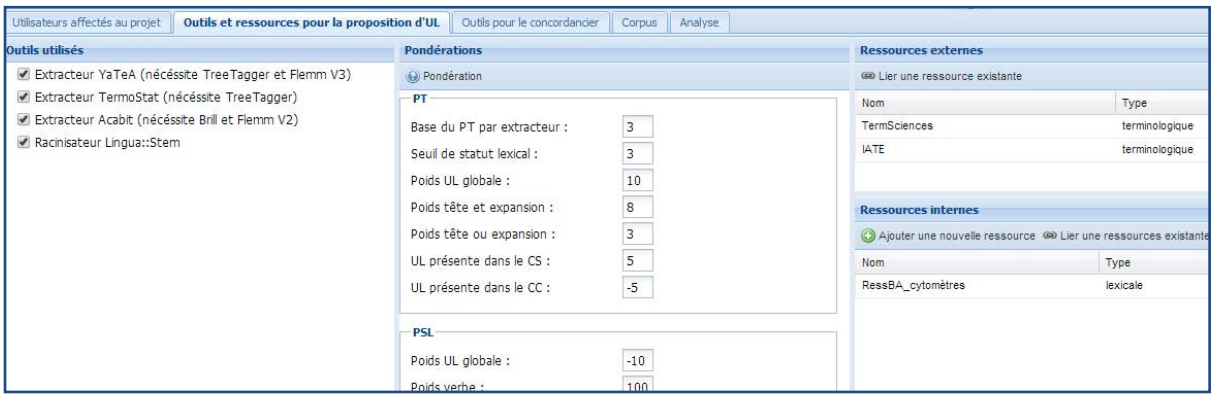


Figure 3: Sélection des outils et des ressources (capture d'écran).

En ce qui concerne la durée de l'analyse automatique, elle dépend de la taille du corpus et du paramétrage utilisateur. Pour un corpus de 50 fichiers représentant un volume de 507 Ko, soumis aux outils TreeTagger et Flemm v3, jugés représentatifs du comportement global des outils lors d'une analyse, le temps d'exécution s'élève à 7,735 s. Il n'augmente que modérément avec l'augmentation du volume

7 http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schmid/Tagger-Licence [08/04/2014].
8 Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.
9 http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm [04/12/2011].
10 http://www.termsciences.fr/ [08/04/2014].
11 http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load [08/04/2014].

d'informations à traiter (3,486 s pour un 1 corpus de 20 fichiers représentant un volume de 214 Ko). Par contre, l'interrogation des ressources externes (par web services) peut rallonger considérablement le temps d'exécution (jusqu'à plusieurs heures).

2.3 Etape 2: Analyse manuelle approfondie

Les résultats de l'analyse automatique sont affichés dans l'interface de travail. Cette interface est divisée en 4 espaces (Figure 4) que l'on peut repositionner, redimensionner, afficher ou cacher.

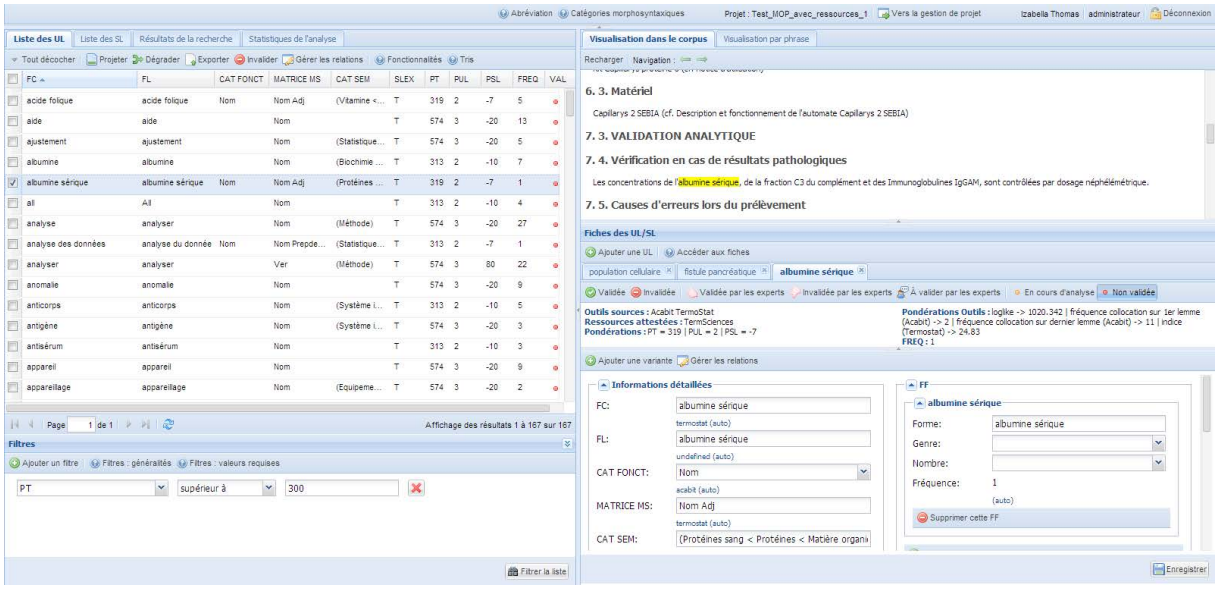


Figure 4: Interface de travail, vue globale (capture d'écran).

Dans l'espace 1, l'analyste visualise la liste des ULC assorties d'informations servant aux tris (par simple clic sur la colonne correspondante) et filtrages des résultats. L'espace 2 permet de filtrer la liste des ULC à l'aide de 21 paramètres, telles que la fréquence, l'extracteur-source ou la classe sémantique. Les filtres sont cumulatifs : il est possible de filtrer la liste selon plusieurs paramètres simultanément ; par exemple, les ULC de matrice morphosyntaxique Nom Adj, proposées par TermoStat, avec au moins 20 occurrences dans le corpus. L'espace 3 sert à visualiser les ULC dans leur contexte initial, en projetant une ou plusieurs ULC sélectionnée(s) dans la liste, en corpus ou par phrase (cf. exemple de «*albumine sérique*» sur la Figure 4). L'espace 4 sert à afficher les fiches lexicales des ULC sélectionnées dans la liste. Une fiche lexicale comporte l'ensemble d'informations concernant l'ULC, c'est-à-dire une description complète de la forme canonique d'une ULC avec des spécifications sur ses formes fléchies. L'analyste peut modifier, ajouter, valider ou enlever les informations.

Chaque ULC est aussi assortie d'une *fiche de relations* qui permet de visualiser et/ou de définir un réseau de relations qu'elle entretient avec d'autres formes recensées. Il s'agit de :

- relations morphologiques (recensement de formes fléchies (FF) d'une ULC, recensement de formes en relation de dérivation avec une (partie de) ULC (cf. UL dérivées sur la Figure 5) ;

- relations lexico-syntaxiques (cf. UL incluses, composées et associées sur la Figure 5) ;
- relations lexico- sémantiques (cf. UL homonymes, UL variantes sur la Figure 5).

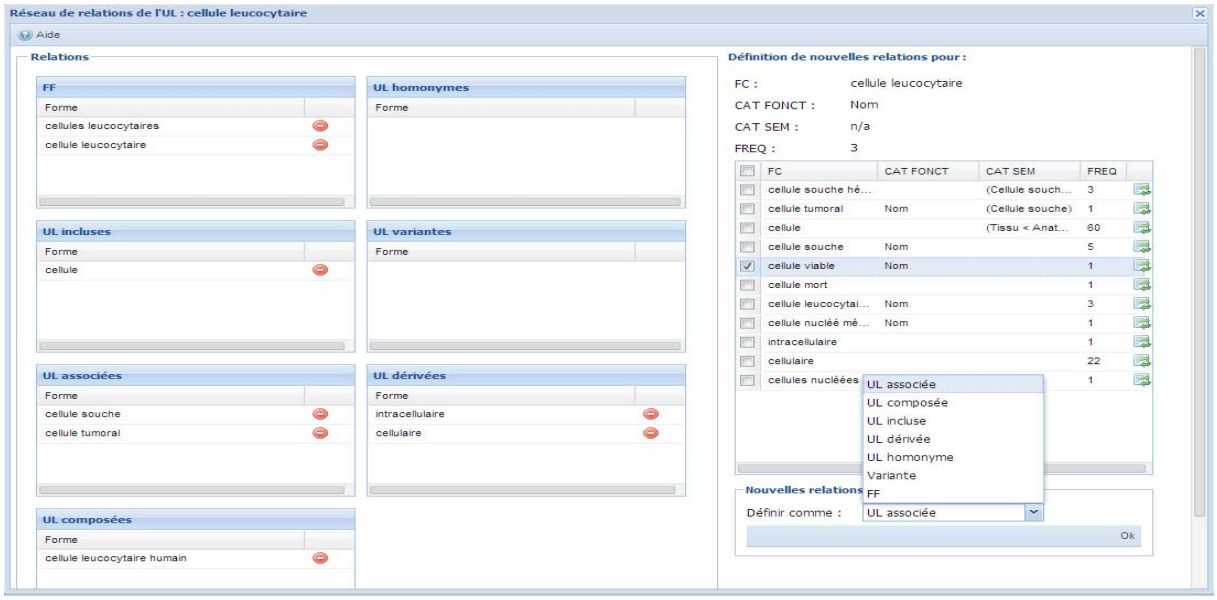


Figure 5: Exemple d'une fiche de relations pour UL *cellule leucocytaire* (capture d'écran).

Certaines relations entre les ULC sont proposées automatiquement, soit par les outils/ ressources, soit à partir de calculs effectués par la station. L'analyste peut les valider/invalider, mais aussi établir des nouvelles relations (cf. Figure 5, «*cellule viable*» sélectionnée dans la liste de gauche est définie comme une UL associée à «*cellule leucocytaire*»).

Afin d'approfondir l'analyse en explorant le corpus, la Station Sensunique propose un concordancier évolué, qui permet de visualiser les occurrences d'une forme en contexte, dans le corpus ou dans les phrases isolées (Figure 6). La recherche s'effectue soit directement à partir des informations saisies par l'utilisateur, soit en demandant à la Station de calculer les informations linguistiquement plus complexes (lemmes, racines ou catégories morphosyntaxiques) concernant les formes à rechercher. Le concordancier est dit 'évolué' au sens où il permet différents types de recherche, allant d'une simple recherche sur une chaîne de caractères jusqu'à une recherche impliquant la combinaison de différents critères linguistiques (lemmes, racines, catégories morphosyntaxiques d'une ou plusieurs unités lexicales). Combiner des critères appartenant à différents niveaux d'analyse permet d'imposer des contraintes plus ou moins fortes sur les motifs recherchés, et ainsi cibler (ou, au contraire, élargir) le champ des résultats. Par exemple, la recherche '*[l]cellule [e]Adj*' (exprimée sous forme d'expressions régulières Sensunique) permet de cibler les groupes composés d'une des formes fléchies du mot «*cellule*» suivie d'une forme à fonction adjectivale (ex. «*cellules nucléées*», «*cellules totales*», «*cellules mortes*» etc.) (cf. Figure 5).

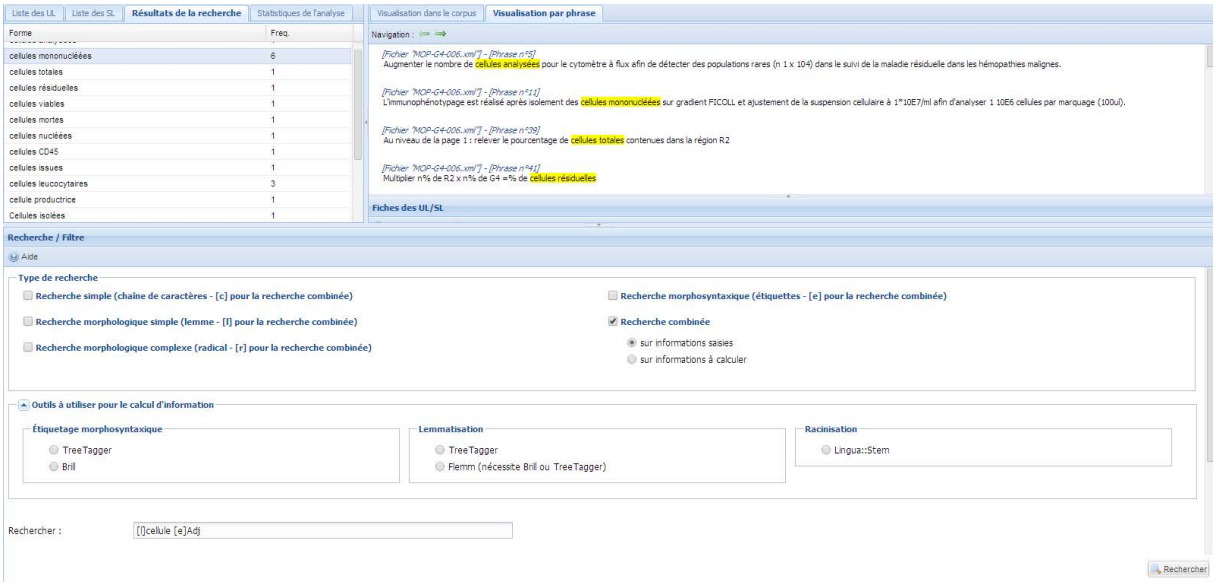


Figure 6: Concordancier évolué (capture d'écran).

2.4 Etape 3 : Création de ressources

A tout moment de son travail, l'utilisateur peut exporter les données recensées dans la station afin de les exploiter dans d'autres applications, les valider par un expert-métier ou simplement, créer une ressource terminologique finale. Les données à exporter, au format XML, peuvent être sélectionnées et restreintes à certaines valeurs grâce à un système de filtres cumulatifs (Figure 7).

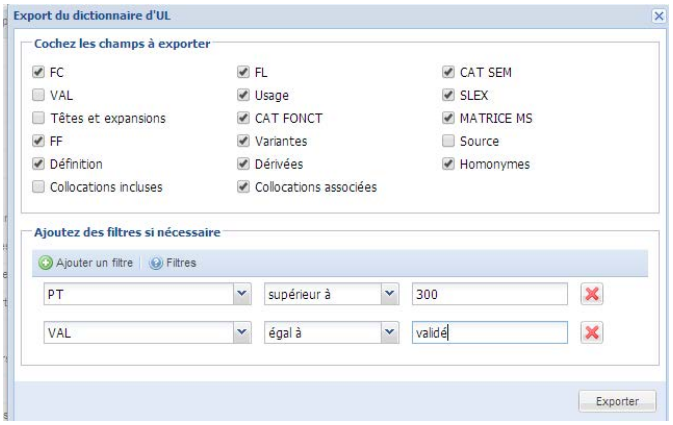


Figure 7: Paramétrage de l'export (capture d'écran).

3 Conclusion

Les fondements méthodologiques et l'architecture logicielle de la Station Sensunique lui permettent de dépasser son objectif initial (établissement du lexique d'une LC) et lui donne le potentiel d'être un

outil générique assistant l'établissement de diverses ressources terminologiques : glossaires, dictionnaires, bases de données, thesaurus, index, termino-ontologies etc., aussi bien pour une consultation directe que comme entrées pour d'autres applications en TAL (recherche et extraction d'information, systèmes d'indexation, acquisition et représentation des connaissances etc.). La facilité d'utilisation de la Station Sensunique nous semble un véritable avantage : c'est un outil prêt à l'emploi, ergonomique, facile à prendre en main et à exploiter. Elle allie les avantages d'une analyse automatique (rapidité, coût) et l'exactitude et la fiabilité d'une analyse humaine. L'utilisation de la Station se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station.

Nous projetons d'intégrer de nouvelles fonctionnalités à la Station Sensunique, concernant le traitement du contenu sémantique des textes, principalement l'amélioration de la détection des relations sémantiques et conceptuelles entre les unités lexicales. Une autre direction de recherche inclut la construction d'ontologies de domaine.

4 Références bibliographiques

Aubin S. et Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing*, 5th International Conference on NLP (FinTAL'2006), Springer, p. 380-387.

Brill E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing* (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.

Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Workshop at the 32nd Annual Meeting of the ACL (ACL'94), Las Cruces, New Mexico, USA.

Drouin P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. In *Terminology*, vol.9, n°1, John Benjamins Publishing Company: Amsterdam/Philadelphia, p. 99-115.

Namer, F. (2000). FLEMM: un analyseur flexionnel du français à base de règles. In *Traitement Automatique des Langues*; vol. 41/2, p. 523-547.

Plaisantin Alecu B., Thomas I., Renahy J. (2012). *La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques*, In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012,* volume 2 : TALN, ATALA/AFCP, pp.511 518, http://www.aclweb.org/anthology/F/F12/F12-2047.

Renahy J., Devitre D., Thomas I., Dziadkiewicz A. (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability, in *Proceedings of the 11th International Symposium on Social Communication*, Santiago de Cuba, Cuba, 19-23 January 2009, pp. 289-293.

Renahy J., Thomas I., Chippeaux G., Germain B., Petiaux X., Rath B., De Grivel V., Cardey S., Vuitton DA. (2011). La langue contrôlée et l'informatisation de son utilisation au service de la qualité des textes médicaux et de la sécurité dans le domaine de la santé, In P. Staccini, A. Harmel, S. Darmoni, R. Gouider, *Systèmes d'information pour l'amélioration de la qualité en santé*, Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM'2011), Tunis, 23-24 septembre 2011, Springer-Verlag.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of international conference on new methods in language processing (Vol. 12, pp. 44-49).

Thomas I., Plaisantin Alecu B., Germain B., Betbeder M.L. (2014). Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources. In *Preceedings of Euralex 2014*, EURAC, Institute for Specialised Communication and Multilingualism.

Lexicography of Lesser Used Languages

Towards an Integrated E-Dictionary Application – The Case of an English to Zulu Dictionary of Possessives

Sonja Bosch*, Gertrud Faaß[1]
*University of South Africa, University of Hildesheim
E-mail: boschse@unisa.ac.za, faassg@uni-hildesheim.de

Abstract

This paper describes a first version of an integrated e-dictionary translating possessive constructions from English to Zulu. Zulu possessive constructions are difficult to learn for non-mother tongue speakers. When translating from English into Zulu, a speaker needs to be acquainted with the nominal classification of nouns indicating possession and possessor. Furthermore, (s)he needs to be informed about the morpho-syntactic rules associated with certain combinations of noun classes. Lastly, knowledge of morpho-phonetic changes is also required, because these influence the orthography of the output word forms. Our approach is a novel one in that we combine e-lexicography and natural language processing by developing a (web) interface supporting learners, as well as other users of the dictionary to produce Zulu possessive constructions. The final dictionary that we intend to develop will contain several thousand nouns which users can combine as they wish. It will also translate single words and frequently used multiword expressions, and allow users to test their own translations. On request, information about the morpho-syntactic and morpho-phonetic rules applied by the system are displayed together with the translation. Our approach follows the function theory: the dictionary supports users in text production, at the same time fulfilling a cognitive function.
Keywords: Zulu; less-resourced languages; learner's dictionary, extralexicographic features; integrated e-dictionary; possessives

1 Introduction

This paper describes an integrated approach to e-dictionaries as learning tools informed by the Function Theory of Lexicography. Issues raised by information science as well as learners' levels of knowledge in text production (cf. Bothma and Tarp 2012) are taken into consideration. The case of possessive constructions in Zulu will be used as example. We will explore ways of using an e-dictionary integrated in a learning environment to translate possessives from English to Zulu. We aim to enhance the language learning experience of users in a didactically valuable manner.

1 This project is part of the activities of Gertrud Faaß as an affiliated research fellow of the University of South Africa.

Our users in focus are not only language learners, but also others who do not strictly fall into one of the categories defined by the function theory, e.g. advanced speakers of the language possibly requiring knowledge about correct grammatical constructions, in this case possessives. We combine lexicographic and morpho-syntactic information and take particular extra-lexicographical situations and user needs into consideration.

To our knowledge this research on an integrated e-dictionary for Zulu is novel work. While e.g. Radtke and Heid (2012:798) propose an approach to address the needs of text production by making use of corpus data, and Prinsloo et al. (2012:292) offer graphical decision trees as a solution for explaining and translating Northern Sotho kinship terminology and copulative constellations to users, we go a step further by integrating morpho-syntactic and lexicographical information as well. We address text production by showing users how to form possessives and by offering information about how these are formed. In addition, we address the cognitive function by assisting users to understand the morphology of the other language (Gouws 2007:85). Though the e-dictionary described here in the long term may be developed into an intelligent computer assisted language learning (ICALL) system, this e-dictionary is not as yet, comparable to what is already developed for other Bantu languages, as e.g. described by Katushemererwe and Nerbonne (2013) for Runyakitara or Hurskainen (2009) for Swahili.

In the next section, a brief background on the morphological structure of Zulu will be given, followed by an exposition of the complexities encountered with the translation of possessives from English into Zulu. Thereafter the implementation of the integrated e-dictionary application will be discussed with detailed exemplification, followed by a conclusion and notes on future work.

2 Background of the morphological structure of Zulu

Zulu [zul] belongs to the Bantu languages which have a rich agglutinating morphological structure, based on two principles, namely the nominal classification system, and the concordial agreement system. According to the nominal classification system, nouns are categorized by prefixal morphemes, which for analysis purposes have been assigned class numbers. These noun class prefixes use concordial agreement to link the noun to other words in the sentence such as verbs, adjectives, pronouns, possessives etc. (cf. Poulos and Msimang 1998).

Some degree of morpho-phonological complexity occurs which can mainly be ascribed to the phonological phenomena at morpheme boundaries resulting from the conjunctive orthography of the Zulu language. These phenomena are mostly predictable and rule-based. Because of its complexity, the construction of the possessive constitutes a particular challenge for language learners. First, information on the noun class of the possession and the possessor is required in order to determine the connecting element: it can either be a “regular” class specific possessive concord containing the possessive marker *a,* or if the possessor is a noun belonging to class 1a, then a special possessive marker

ka is used (cf. Poulos and Msimang 1998:146). After selecting the appropriate connecting element, the speaker has to know the morpho-phonetic rules for cases where the connecting element and the possession are to be fused. See some examples in the following table:

English possessive construction	Zulu possession (class)	connecting element	Zulu possessor (class)	Translation	Rule
Sipho's cats lit. 'cat's of Sipho'	amakati (6)	ka-	uSipho (1a)	amakati kaSipho	fuse connecting element with possessor by deleting initial vowel *u-*
old women's box lit. 'box of old women'	ibhokisi (5)	la-	izalukazi (8)	ibhokisi lezalukazi	fuse connecting element with possessor by means of vowel coalition *a+i > e*
mother's bread lit. 'bread of mother'	isinkwa (7)	sika-	*umama* (1a)	isinkwa sikamama	fuse connecting element with possessor by deleting initial vowel *u-*

Table 1: Examples of Zulu possessive constructions.

When a possessor is to be replaced by a pronoun, its noun class again plays an important role. For instance when translating the phrase 'their box' (lit. 'box of them') referring to 'old women' as possessor, the connecting element or possessive concord of the possession ('box') class 5 *la-* is prefixed to the (abbreviated) pronoun of class 8 *zo*, resulting in *ibhokisi lazo*.
Examples of even higher complexities are highlighted in the implementation section below.

3 Challenges for learners

The microstructure of bilingual dictionaries usually contains source language lemmas and inter alia their respective translation equivalent in the target language. The user needs in terms of the function theory could be on the level of text reception[2], text production or cognitive needs in order to understand the construction. Occasionally a language learner might want to query short phrases, such as possessive structures e.g. "Sipho's food", or "(the) doctors' medicine" without having to look up the possessor, the possession as well as the grammatical rules on how to formulate the construction.
Although Computer Aided Language Learning (CALL-) applications offer such possibilities, these applications usually only have restricted dictionaries that contain solely what is described in the learning material accompanying them. This is usually sufficient, as the learner can learn about the rules

2 The work described here, however, does not cater for receptive user needs as this issue is foreseen to be taken into account at a later stage.

for producing the constellation in the target language and only has to use a bilingual dictionary to look up translations for the words (s)he needs to know.

But what if even simple constellations show a huge variety in forms? In languages like German, English or French, there are only three parameters to take into account when formulating a possessive: person, number, and gender, the patterns are not too difficult to learn and there are few exceptions. For Zulu, a fourth parameter replaces the 3 choices of gender with up to 16 choices of noun class, some of which are illustrated in Table 1. Language learners hence struggle for quite a while to learn full paradigms for each of the day-to-day constellations. In addition, the above mentioned morpho-phonetic phenomena need to be memorized and kept track of as well. One of the constellations that is very challenging to absorb is the possessive construction, as knowledge of the following facts is required: (i) class of the possessor; (ii) class of the possession; (iii) the concord that is to be used; (iv) generic morpho-phonological rules.

To give an idea of the complexity of possessive constructions, in some Zulu learner guides, the description of the possessive construction is spread over an average of four pages which "can be very user-unfriendly or time consuming for him/her to find the right information by having to read entire sections of grammatical descriptions." (Prinsloo and Bosch 2012:300).

4 Current State of Implementation

Our system architecture comprises three parts: (1) an sql-database containing the dictionary (lexicographic content part); (2) morpho-phonological rules (linguistic part) and, lastly, (3) programming scripts that build the Graphical User Interface (GUI) and build the web page, i.e. the interface between database and user (the presentation part), cf. figure 1.

We opted for using the XAMPP[3]-environment for developing the sql-database, because it assists when designing and filling databases with data and also simplifies the creation of websites. *php* is a dynamic scripting language specifically designed to develop websites and is part of the XAMPP environment. It offers phpMyAdmin, an sql-interface, i.e. sql-commands can be executed on a respective database with *php* and allows the results to be written into variables that can then be displayed on a website. After the development has been finalized, the php scripts and the database can easily be ported onto an apache webserver. Our test interface can be found on www.uni-hildesheim.de/iwist-cl/projects/eZulu/login.php, access is currently restricted to known users[4].

3 http://www.apachefriends.org/de/xampp.html

4 Please contact G. Faaß if you should like to try out the test interface

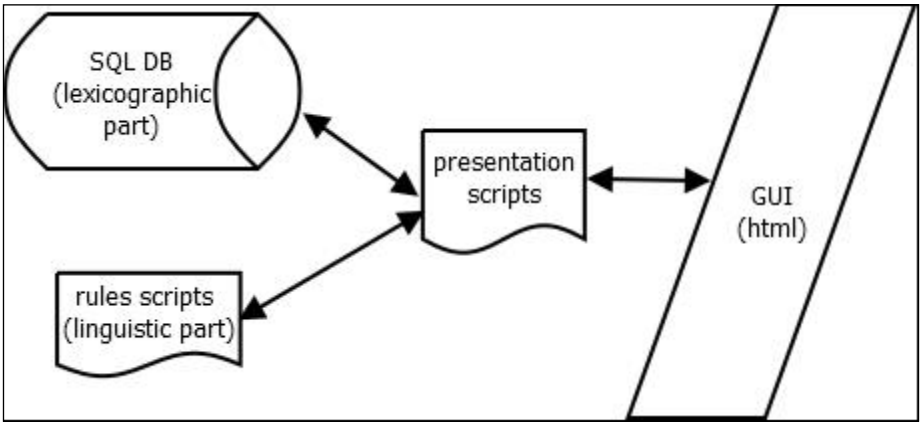


Figure 1: System Architecture.

For a start, our database is kept simple and currently contains representative data of all noun classes necessary to perform the given task. We are still working on the design of the website, as it should fulfil the users' needs in terms of the relevant issues of the function theory described above. Currently, we focus on an accurate implementation of the rules and a user-friendly way to explain their application to those users who request such an explanation by ticking the checkbox.

Though the website design is preliminary, we show screenshots in Figures 2, 3 and 4 in order to demonstrate our implementation principle: keep it simple and intuitive. In the header, the user is asked what kind of information (s)he requires over and above the translation which is done in all cases. Checkbox 1 on the left asks whether the user would like to have a pronominal version of the translation (i.e. the possessor noun being replaced by the pronoun belonging to the same noun class) – in addition to the full translation as Zulu uses different pronouns dependent on the noun class they belong to, hence the pronouns "it" or "they" has several translation equivalents, dependent of the noun class of the noun that it stands for. If we ask the user what possessor the pronoun stands for, it will be possible to identify the noun class of this noun so that the tool can choose the right pronoun.

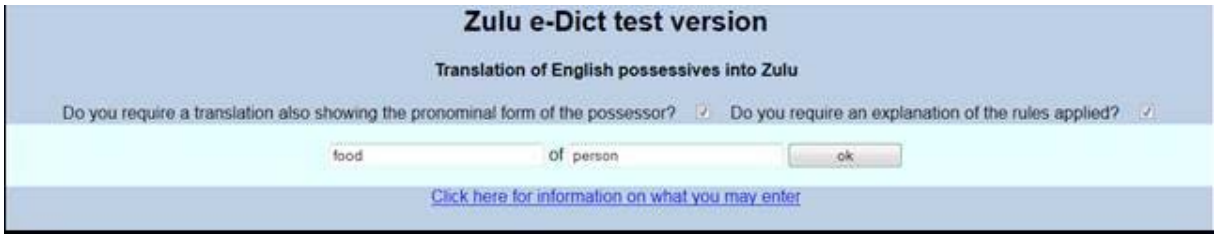


Figure 2: Start page: user requires translation of "food of person" plus rules plus pronominal translation of the possessor.

Ticking checkbox 2 on the right hand side will lead to an additional explanation how the resulting Zulu translation was formed concerning the elements and the rules applied. These additional explanations are currently still kept brief, but we plan to link these to other pages where the applied rules will be described in more detail and with additional examples.

Figure 3 demonstrates the result of the query shown in Figure 2: A dynamic webpage which is divided into 4 blocks is visibly distinguishable: The 1st block, i.e. the header consists of static information on the page as such ("Zulu e-Dict test version"), while the 2nd block – which appears in all cases of use – repeats the English expression entered ("food of person"[5]). It then shows the word-by-word translation of possession and possessor (*ukudla* and *umuntu*), and also the possessive concord that is to be used (*kwa*). Since the checkbox requiring the pronominal form was ticked, it appears as well (*ukudla kwakhe*).

Ticking the checkbox "Do you require an explanation of the rules applied" results in the popping up of a 3rd block appears (identified by a darker blue background). Here, the English expression is repeated, however now, possession and possessor are clearly categorized as such. This 3rd block then also informs the user that the elements of the English construction "food of person" are *ukudla kwa,* and *umuntu* in Zulu. In addition to the translation of each of the elements, the rules forming the correct translation equivalent, *ukudla komuntu*, are briefly explained. A second paragraph describing the rules of forming the pronominal form is added, because the respective checkbox was ticked.

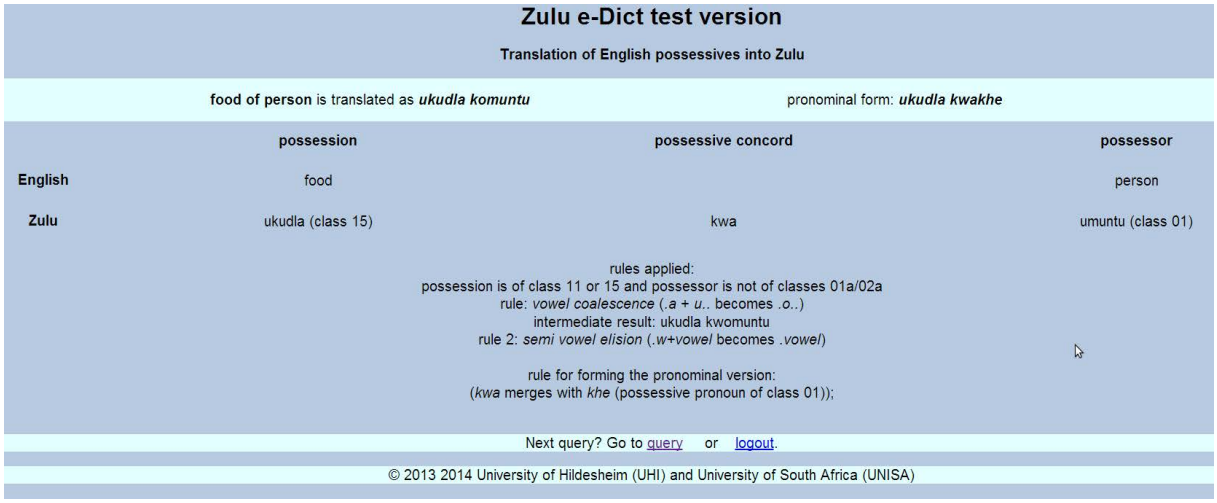


Figure 3: Webpage resulting from query shown in Figure 2.

We make use of the checkboxes mainly to show only the relevant data on the resulting page, i.e. the required units of information as they were defined by the user (information on demand, cf. Bothma and Tarp 2012:90). This is in line with Tarp's (2009:26) claim that "The study of access routes, i.e. the rapid and easy access to the relevant needs-adapted data, is of utmost importance to lexicography." Figure 4 demonstrates what happens if both checkboxes remained unticked.

5 We also consider offering the shorter form, i.e. "person's food" with the final design of the website.

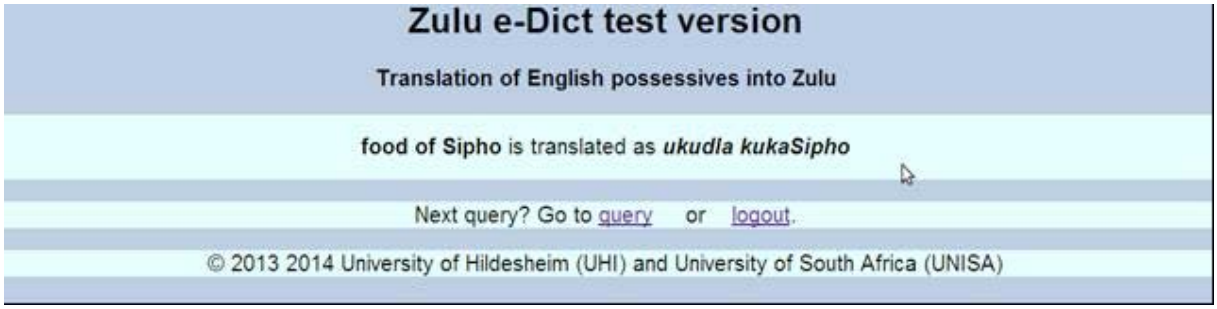


Figure 4: Webpage resulting from the query: "food of Sipho", no checkboxes ticked.

In this version of the website, no linking information to study guide pages/chapters is given, however as it is planned to become part of a free online course (cf. University of South Africa, 2010), this will be added then. As mentioned earlier on, the lexicon presently only contains a limited number of nouns (approximately 500 entries) that may be entered. However, we are currently working towards a second version of the database that will contain the vocabulary described in the study material used in the online course. The third version is then planned to become part of an e-dictionary of the African languages currently developed in the framework of the SeLA[6] project which will translate single words and other morpho-syntactic constructions in both directions

5 Challenges and Plans for further Implementation Steps

What we still consider a problem is the case where a possessive concord ends in *–a* while the possessor begins with *–o*. In this case – restricted to class 2a noun possessors – there are two ways of forming the Zulu possessive, as demonstrated by Figure 5: The first rule, when applied, merges the concord with the possessor and thereby deletes the vowel ("vowel elision") leading to *yodokotela* while the second, alternative rule adds *–w–* ("semi-vowel insertion") resulting in the constellation *yawodokotela*. As both constellations occur frequently in a Zulu corpus[7], we cannot decide prescriptively which of the constellations should be used. This is a problem in the light of the productive function that this e-dictionary should fulfil: the user might be confused not knowing which of the two constellations is "better". Currently, we need to offer both alternatives, without distinctive features, as more corpus based research will be necessary to determine whether there are contextual factors that give preference to one of the constellations above the other.

An open challenge is the didactic element that is foreseen: we would like give users the opportunity to test their knowledge of Zulu, i.e. we will allow them to enter a translation equivalent that they

6 Scientific e-Lexicography for Africa, cf. Heid (2012) and http://www.uni-hildesheim.de/iwist-cl/projects/sela/

7 We are very thankful to Prof. DJ Prinsloo for giving us access to the Zulu corpus at the University of Pretoria, cf. Prinsloo and de Schryver (2005).

deem correct in one of the next versions of the dictionary. The system is then to compare the user's input with the correct output and, in a first stage, it will inform the user whether his/her input was correct or not. However, we will ask each user whether (s)he will give his/her consent to store the given data. If this consent is given, the input plus the attempted and the correct output will both be stored in a log file. We hope that these data will help to investigate typical learners' mistakes and incorrect generalizations. Depending on the output, we hope that in future the system can be enabled to dynamically offer appropriate feedback. These feedback messages, like the log file, are not planned to be personalized, but will only depend on the given inputs.

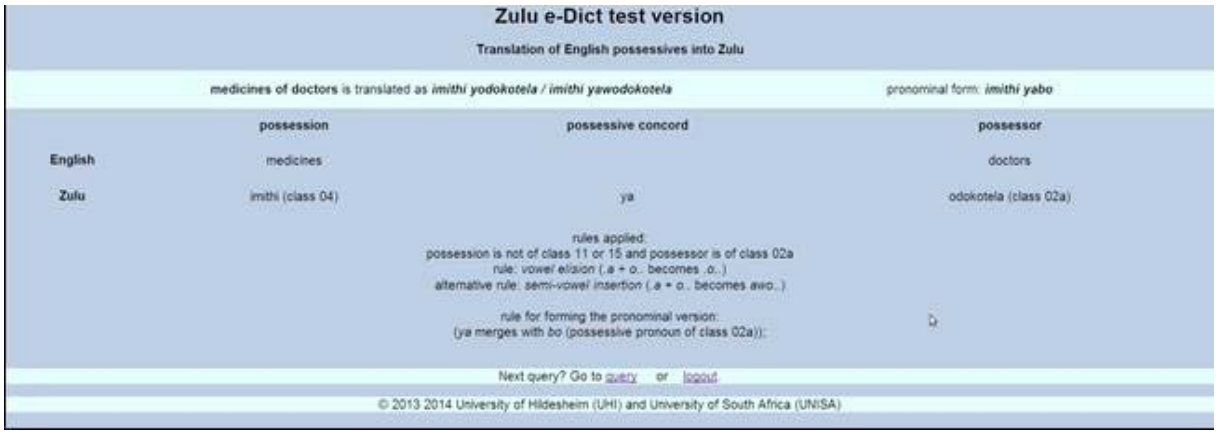


Figure 5: Webpage resulting from a query with a noun of class 2a as possessor (usually beginning with *o–*) and a noun of a class where the possessive concord ends in *–a* as possession.

6 Conclusion and Future Work

We believe that this prototype e-dictionary with integrated morpho-syntactic information will go a long way to address user needs in specific user situations, in particular with respect to complicated issues such as the translation of possessives from English to Zulu. A next stage of development will allow for a bi-directional translation, where Zulu possessives given as input will be translated into English. This method will serve as a starting point for adding further constructions, while the vocabulary contained will be extended continuously. A parallel activity will be a thorough evaluation of the system; so far, the performed random sample tests have not discovered any errors.

At a later point, we also foresee an expansion of the prototype to closely related Bantu languages such as Xhosa, Swati, Southern Ndebele and Zimbabwe Ndebele. The prototype will also be testing for usability and user preferences, and therefore a test access has been made available for interested users.

7 References

Bothma, T. and Tarp, S. (2012). Lexicography and the Relevance Criterion. *Lexikos* 22, pp. 86-108.

Gouws, R. (2007). A Transtextual Approach to Lexicographic Functions. *Lexikos* 17, pp. 77-87.

Heid, U. (2012). SeLA – a new project on electronic lexicography. *Lexicographica* 28(1), pp. 437-440.

Hurskainen, A. (2009). Intelligent computer-assisted language learning: Implementation to Swahili. Technical Reports in Language Technology Report No 3, University of Helsinki, Finland. http://www.njas.helsinki.fi/salama. [09/04/2014]

Katushemererwe, F. and Nerbonne, J. (2013). Computer-Assisted Language Learning (CALL) in support of (re-) learning native languages: The case of Runyakitara. *Computer-Assisted Language Learning*, 1-18 DOI: 10.1080/09588221.2013.792842 [09/04/2014]

Poulos, G. and Msimang, C.T. (1998). *A linguistic analysis of Zulu*. Pretoria: Via Afrika.

Prinsloo, D.J. and Bosch, S. (2012). Kinship terminology in English–Zulu/Northern Sotho dictionaries - a challenge for the Bantu lexicographer. In: *Proceedings of the 15th EURALEX International Congress*. 7-11 August (2012), Oslo, pp. 296-303. Oslo: Reprosentralen, UiO. ISBN 978-82-303-2095-2

Prinsloo, D.J. and de Schryver, G-M. (2005). Managing eleven parallel corpora and the extraction of data in all official South African languages. In W. Daelemans, T. du Plessis, C. Snyman & L. Teck (eds.) *Multilingualism and Electronic Language Management*. Proceedings of the 4th International MIDP Colloquium, 22-23 September 2003, Bloemfontein, South Africa (Studies in Language Policy in South Africa 4), pp. 100–122. Pretoria: Van Schaik Publishers. Available: http://tshwanedje.com/publications/11ParCor.pdf [28/10/2013]

Prinsloo, D.J., Heid, U., Bothma, T., and Faaß, G. (2012). Devices for Information Presentation in Electronic Dictionaries. *Lexikos* 22 pp. 290-320

Radtke, J. and Heid, U. (2012). Word formation in electronic language resources: state of the art analysis and requirements for the future. In*: Proceedings of the 15th EURALEX International Congress*. 7-11 August 2012, Oslo, pp. 794-802. Oslo: Reprosentralen, UiO. ISBN 978-82-303-2095-2

Tarp, S. 2009. Beyond Lexicography: New Visions and Challenges in the Information Age. In Bergenholtz, H., Nielsen, S. & Tarp, S. (eds), *Lexicography at a Crossroads Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow.* Linguistic Insights 90, pp.17-32. ISBN 978-3-03911-799-4 br

University of South Africa Free Online Course. 2010. Available: http://www.unisa.ac.za/free_online_course/ [26/08/2013].

Zur (Vor-)Geschichte der saamischen Lexikografie: ein lateinisch-saamisches Wörterverzeichnis aus dem 17. Jahrhundert

Eino Koponen
Institut für die Landessprachen Finnlands
eino.koponen@kotus.fi

Zusammenfassung

Die Saamen (früher Lappen) sind ein nordisches Minoritätsvolk, das auf dem Gebiet von vier Staaten (Norwegen, Schweden, Finnland, Russland) lebt. Die Wurzeln der auf Saamisch gedruckten Literatur reichen ins 17. Jahrhundert zurück, und aus demselben Jahrhundert datiert der erste Versuch eines saamischen Lexikons. Dabei handelt es sich um ein von dem schwedischen Pastor Z. Plantinus verfasstes handschriftliches Wörterverzeichnis von ca. 850 lateinischen Wörtern mit saamischen Äquivalenten, das 1888 in Stockholm wiederentdeckt und später publiziert wurde. Das uns überlieferte Dokument scheint eine korrumpierte Abschrift des Originals zu sein, deren zahlreiche Fehler das Erkennen der saamischen (und bisweilen auch lateinischen) Wörter stellenweise (fast) unmöglich machen. Hier wird das Wörterverzeichnis einer genaueren Analyse unterzogen. Die Arbeit wurde im Rahmen eines größeren Forschungsprojekts, das eine etymologische Datenbank der saamischen Sprachen und der verwandten uralischen Sprachen darstellt, am Institut für die Landessprachen Finnlands durchgeführt. Alle saamischen Wörter des Verzeichnisses sind jetzt in die Datenbank eingespeist. Auch die auf den ersten Blick obskuren Wörter werden dadurch der künftigen Forschung zugänglicher. Bei der Präsentation werden die Resultate der Analyse anhand der Álgu-Datenbank veranschaulicht.

Stichworte: Geschichte der Lexikografie; etymologische Datenbank; Saamisch; Latein

1 Einleitung

1.1 Die saamische Literatur im 17. Jahrhundert

Die ersten auf Saamisch gedruckten Bücher, ein ABC-Buch und ein Gesangbüchlein, erschienen 1619 in Schweden. Der Verfasser war Nicolaus Andreae, Pastor der Gemeinde Piteå. Auch wenn man alle Schwierigkeiten in Acht nimmt, die der Bahnbrecher zu beseitigen hatte, zeugen diese Werke nicht über besonders gute Sprachkenntisse des Verfassers. Noch bescheidener war die Beherrschung des Saamischen bei dem Fortsetzer seiner Arbeit, dem in Umeå ansässigen Pastor Olaus Petri Niurenius. Der von ihm 1633 herausgegebene Katechismus ist, was die Sprache anbetrifft, als das erbärmlichste

je gedruckte saamische Werk beurteilt worden (Wiklund 1922: 22). Im Gegensatz zu den oben genannten, deren Sprache als entstelltes Süd- oder Umesaamisch mit finnischer Mischung charakterisiert werden kann, repräsentiert das nächste Druckwerk, das ABC-Buch eines unbekannten Autors (1638), einen archaischen lulesaamischen Dialekt (Bergsland 1982).

Der Höhepunkt der saamischen Literatur im 17. Jahrhundert, wenigstens was den Umfang betrifft, wurde 1648 erreicht, als das 950-seitige Manuale Lapponicum von Johannes Tornaeus (Pfarrer von Nieder-Torneå) herauskam. Tornaeus hatte zur Aufgabe eine Schriftsprache zu erschaffen, die allen Ansprüchen der Saamen im schwedischen Reich genügen würde. Aus dem Blickwinkel der modernen Sprachwissenschaft kann das Resultat unterschiedlich beurteilt werden (s. zuletzt Koponen 2010 mit weiterer Literatur). Auf jeden Fall befriedigte es nicht alle Zeitgenossen, sodass Olaus Stephani Graan, Pastor in Lycksele, das Manuale (allerdings in engerem Umfang) 1669 erneut in seine eigene umesaamische Muttersprache übersetzte. Da die nächsten auf Saamisch gedruckten Bücher schon dem folgenden Jahrhundert angehören, kann dieses Werk neben einem von demselben Autor 1668 herausgegebenen Katechismus als der Endpunkt der saamischen Literatur im 17. Jahrhundert betrachtet werden. (Qvigstad & Wiklund 1899: 22-23.) Sprachlich werden die Werke von Graan dem südlichen Teil der saamischen Bevölkerung (d. h. den Süd- und Umesaamen) gefallen haben, für die anderen (d. h. für die Pite-, Lule- und Tornesaamen; von den östlichen Kemi- und Inarisaamen ganz zu schweigen) waren sie sicher mehr oder weniger unverständlich.

Wenigstens was die Lautbezeichnung anbetrifft, bedeuteten die Arbeiten von Graan einen Schritt rückwärts. Obgleich die Orthografie von Tornaeus unvollständig und inkonsequent ist, sind die Wörter bei ihm meistens auch ohne Kontext (oder Übersetzung) zu erkennen. Bei Graan (und seinen Vorgängern) hingegen ist das Lesen der Wörter nur aufgrund des Schriftbilds öfters unmöglich. Besonders markant kommt das zum Vorschein bei den Affrikaten [ts] und [tš], die bei Tornaeus (meist) durch *z* bezeichnet werden, während die übrigen Druckwerke des 17. Jahrhunderts für sie (oft) *ti, ki, gi* (oder aber auch nur *t, k, g*) benutzen. Da dieselben Buchstaben (vor allem wenn kein *i* folgt aber nicht nur) für normale Klusile stehen können, ist die Lautung der Wörter nur aufgrund der Bedeutung zu erschließen. Neben den Affrikaten sei bemerkt, dass der Sibilant [š] bei Tornaeus (meist) *sch,* bei den übrigen Autoren des 17. Jhs (meist) *hi* geschrieben wird. (Qvigstad 1899: 13; Qvigstad 1947: 19; Bergsland 1982; Sköld 1986)

1.2 Die saamischen Wörterbücher des 18. Jahrhunderts

Das erste saamische Wörterbuch erschien 1738. Der Autor war der als Schullehrer (später als Pfarrer) in Lycksele ansässige Petrus Fjellström (geb. 1697 im schwedischen Lappland), der in demselben Jahr auch die erste saamische Grammatik herausgab. Das Wörterbuch umfasst 190 Seiten und ca. 7.500 schwedische Stichwörter mit saamischen Übersetzungen. Da einerseits ein schwedisches Stichwort mehrere saamische Wörter als Übersetzungen haben, andererseits aber ein und dasselbe saamische Wort als Übersetzung für mehrere schwedische Wörter auftreten kann, beläuft sich die Anzahl der

saamischen Lexeme in diesem Wörterbuch auf 6.000 bis 9.000. Mit diesen und anderen Arbeiten von demselben Autor und seinen Zeitgenossen (s. genauer Qvigstad & Wiklund 1899: 26-) wurde in der Mitte des Jahrhunderts die erste schwedischsaamische normierte Schriftsprache entwickelt, die sich im Lexicon Lapponicum (LL 1780) von Ericus Lindahl und Johannes Öhrling manifestierte. Dieses Großwerk besteht aus einem 584-seitigen saamisch-lateinisch-schwedischen Teil mit reichlichen Beispielen aus der saamischen Phraseologie und aus einem 130-seitigen schwedisch-saamischen Index, der als ein selbständiges kleineres Wörterbuch betrachtet werden kann (Larsson 1997: 107).
Während die Schriftsprache sich in Schweden auf der Basis der südlichen Dialekte entwickelte, basierte die Schriftsprache in Norwegen auf einem nördlichen (dem schwedischen Tornesaamisch nahestehenden) finnmarksaamischen Dialekt. Obgleich das Saamische auf der norwegischen Seite erst seit dem 18. Jahrhundert als Schriftsprache benutzt wurde, wurde der erste Teil eines norwegischsaamischen Lexicon Lapponicum bipartitum von Knud Leem schon 1768 herausgegeben. Dieses Werk, das mehr mit dem Wortreichtum als mit systematischer Darstellung imponiert, umfasst über 1600 Seiten und gibt die Bedeutungen der saamischen Stichwörter auf Dänisch und Lateinisch an. Der zweite Teil mit dänischen Stichwörtern und mit einem lateinischen Index erschien erst 1781, sieben Jahre nach Leems Tod, und wurde von seinem Kollegen Gerhard Sandberg herausgegeben. (Larsson 1997: 106-107; zu den alten und neueren saamischen Wörterbüchern s. auch Magga 2012.)

1.3 Die lateinisch-schwedischen gedruckten Wörterverzeichnisse des 17. Jahrhunderts

Da die Bildungssprache am Anfang der Neuzeit auch im schwedischen Reich Lateinisch war, wurden nach dem mitteleuropäischen (vor allem deutschen) Vorbild Wörterverzeichnisse als Hilfsmittel für das Studium des Lateinischen herausgegeben. Die bekanntesten sind das 4-sprachige (lateinisch-schwedisch-deutsch-finnische) Lexicon Latino-Scondicum von Ericus Schroderus (1637) und das „Variarum Rerum Vocabula Latina“, das im 16. und 17. Jahrhundert in mehreren Auflagen (zuerst Lateinisch-Schwedisch, später unter Hinzufügung des Finnischen) herausgegeben wurde. Für unser Thema ist nicht ohne Interesse, dass man ein handschriftliches lateinisch-finnisches Wörterverzeichnis aus dem Jahre 1669 kennt, das eine Abschrift des finnischen Materials aus der letzten Auflage des letztgenannten Werkes (VR 1668) darstellt. Es handelt sich um eine Arbeit, die der deutsche Arzt und Gelehrte Martinus Fogel auf die Aufforderung des Großfürsten von Toscana Cosimo III., ihm Materialien über die finnische Sprache zu verschaffen, durchführen ließ. (Hierzu genauer Stipa 1990: 78-80.)

2 Die etymologische Datenbank Álgu

Am Institut für die Landessprachen Finnlands wurde im Laufe der letzten 12 Jahre eine etymologische Datenbank der saamischen Sprachen aufgebaut, die vor allem der Erforschung des Ursprungs und der Geschichte des saamischen Wortschatzes dienen soll. Die Datenbank namens Álgu (nordsaamisch: ‚Beginn, Herkunft') hat neben einer finnischen und nordsaamischen Benutzeroberfläche auch eine auf Deutsch und Englisch. Die Datenbank steht im Internet allen Interressierten frei zur Verfügung. Alle saamischen Wörter aus dem hier zu besprechenden Wörterverzeichnis sind jetzt in Álgu eingespeist und dort (unter der Sprache Schwedischlappisch) leicht aufzufinden. Aus der Datenbank geht hervor, in welcher Relation (Äquivalenz, Ableitung usw.) jedes Wort zu dem uns aus anderen Quellen bekannten saamischen Wortmaterial steht. Auch die auf den ersten Blick obskuren Wörter des Verzeichnisses werden dadurch einleuchtend und der künftigen Forschung zugänglicher, vorausgesetzt natürlich, dass ihre Rätsel richtig gelöst sind. Bei Wörtern, die vorläufig nicht oder nur mit Vorbehalt identifiziert worden sind, erfährt der Benutzer auch diese Information. (Zum Aufbau der Datenbank s. genauer Aapala et al. 2010.)

3 Das Wörterverzeichnis von Zacharias Plantinus

3.1 Die Entdeckung, Entstehung und Herausgebung des Verzeichnisses

Das Wörterverzeichnis, das hier einer genaueren Analyse unterzogen wird, wurde 1888 im Nachlass des schwedischen Schriftstellers und Sprachforschers Georg Stiernhielm (1598–1672) in der Königlichen Bibliothek zu Stockholm wiederentdeckt. Die Entdeckung wurde von dem finnischen Finnougristen E. N. Setälä gemacht, der das Verzeichnis etwas später auch herausgab (Setälä 1890). Der Verfasser des Wörterverzeichnisses ist Zacharias Olai Plantinus, geboren in den zwanziger Jahren des 17. Jahrhunderts in Umeå als Sohn des oben erwähnten Olaus Petri Niurenius (1580–1645). Über den späteren Lebenslauf von Z. Plantinus kann erwähnt werden, dass er nach seinem Studium in den fünfziger und sechziger Jahren u. a. als Lektor der griechischen Sprache am Hernösander Gymnasium arbeitete und zweimal auch als Rektor des Gymnasiums tätig war. 1672 wurde er zum Pastor des Kirchspiels Offerdal und Probst über ganz Jämtland ernannt. Plantinus starb 1688. (Setälä 1890: 86-87) Vater und Sohn waren Pastoren in Gemeinden mit (auch) saamischer Bevölkerung und engagierten sich beide in der Erforschung der Saamen und ihrer Sprache. Wie mehrere seiner Kollegen hat Z. Plantinus Johannes Schefferus bei seinem berühmten Werk „Lapponia" (1673) Hilfe geleistet. Neben einem von Vater Niurenius verfassten und von Sohn Plantinus ergänzten handschriftlichen Lappmarkbericht hat Schefferus das Vorwort eines ungedruckt gebliebenen saamischen Lexikons von Plantinus benutzt (Löw 1956: 14, 409). Auch dieses handgeschriebene Vorwort wurde von Setälä gefunden, allerdings an einer ganz anderen Stelle, sodass es seiner Meinung nach unsicher sei, ob das Vorwort

und das Wörterverzeichnis zusammengehören. Auf jeden Fall kennt man kein anderes lexikalisches Werk von Plantinus, zu dem dieses Vorwort gehören könnte (Setälä 1890: 90).

Was den Zeitpunkt der Enstehung betrifft, stellt Setälä (1890: 86) fest, dass Stiernhielm bereits 1672 starb, sodass das Wörterverzeichnis nicht aus einer späteren Zeit stammen kann. Setälä verbindet die Entstehung des Wörterverzeichnisses einerseits mit der Ernennung von Plantinus zum Pastor in Offerdal und andererseits mit der Aufforderung des schwedischen Antiquitätsarchivs, Schefferus für sein Werk Nachrichten mitzuteilen, und hält deswegen für am wahrscheinlichsten, dass das Wörterverzeichnis auch nicht früher sondern eben um 1672 geschrieben wurde. G. J. Stipa (1990: 83, 145) hingegen verlegt den Zeitpunkt der Entstehung in das zweite Drittel des Jahrhunderts, d. h. zeitlich wenigstens sechs Jahre früher. Laut Stipa wurde das Wörterverzeichnis von Stiernhielm, Schefferus und späteren Forschern benutzt. Falls Stipa recht hat, muss das Wörterverzeichnis früher in mehreren Exemplaren existiert haben. Dies ist an sich sehr gut möglich, zumal das einzige uns bekannte Exemplar mit seinen zahlreichen Schreibfehlern eine korrumpierte Abschrift des Originals zu sein scheint. Im Gegensatz zu dem eigenhändigen Begleitschreiben ist das an Stiernhielm versandte Exemplar des Wörterverzeichnisses nicht von Plantinus selbst, sondern von unbekannter Hand „ins Reine geschrieben" worden (Setälä 1890: 86). Löw (1956: 434) zufolge hätte Setälä es „nicht ganz adäquat" herausgegeben, was bedeuten würde, dass die Schreibfehler wenigstens zum Teil erst dann entstanden wären. Setälä (1890: 90) hingegen versichert, dass er und sein Mitarbeiter, der schwedische Lappologe K. B. Wiklund, keine Mühen beim gewissenhaften Dechiffrieren der Handschrift gescheut hätten, die Handschrift allerdings oft beinahe unlesbar gewesen sei, sodass er nicht zu behaupten wage, dass alle schwierigen Stellen richtig interpretiert worden seien.

3.2 Bekannte und rätselhafte Wörter

Das Verzeichnis enthält ca. 850 alphabetisch geordnete lateinische Stichwörter mit saamischen Übersetzungen. Für ca. 50 lateinische Wörter werden zwei (oder drei) saamische Übersetzungen aufgeführt, sodass die Anzahl der saamischen Wörter ca. 900 ist. Etwa 60 saamische Wörter begegnen im Verzeichnis als Übersetzungen für zwei (oder drei) lateinische Wörter. Wenn man ein und dasselbe Wort nur einmal mitrechnet, beläuft sich die Anzahl der saamischen Lexeme im Verzeichnis auf etwa 840. [1] In Wortklassen teilen sie sich folgendermaßen auf: ca. 370 Substantive, ca. 270 Verben, ca. 140 Adjektive, ca. 40 Partikeln, ca. 10 Numeralien und ca. 10 Pronomina.

Etwa 720 der saamischen Wörter sind mit Sicherheit oder mit einigem Vorbehalt mit Lexemen in LL (und im heutigen Schwedischsaamisch) gleichzusetzen oder enthalten wenigstens einen in LL belegten Wortstamm, d. h. stehen in einem (korrelativen) Ableitungsverhältnis zu einem dort vorkommenden Lexem. Etwa 40 Wörter sind mit größerem Vorbehalt mit einem in LL belegten Wort(stamm) zu vergleichen, und etwa 140 Wörter haben keine Vergleichspunkte in LL. Einige obsolete Wörter sind

1 Laut Setälä (1890: 90) ist die Anzahl der saamischen Wörter 820; er erklärt nicht, wie er zu diesem Ergebnis gekommen ist.

mir bei Tornaus oder Graan aufgefallen.[2] In Fjellströms Wörterbuch habe ich passende Wörter aufgrund ihrer anzunehmenden schwedischen Bedeutung gesucht,[3] und in gleicher Weise wurden von mir alle Wörterbücher der heutigen saamischen Sprachen sowie der lateinische Index in Leems Wörterbuch durchgearbeitet. Die meisten und sichersten Entsprechungen für in LL unbelegte Wörter habe ich im Südsaamischen gefunden.[4] Im heutigen Lulesaamisch (sowie im Nord- und/oder Ostsaamischen, nicht aber im Südsaamischen) begegnende Entsprechungen haben *såbge* ‚radix' (*suobde* ‚Wurzel eines gefällten Baumes') und *loivijth* ‚frangere' (*lajggit* ‚losreißen, abziehen, abreißen'), und zwei Wörter wären vielleicht mit dem Kildinsaamischen zu vergleichen: *wout* ‚divitiae' (*vāptõk* ‚reich; Reichtum') und *vetzeth* ‚currere' (*vīdže ‚laufen'); das letztere könnte aber gleich gut mit LL qwotset ‚currere' identisch sein. Auch wenn die Vergleichungen mit den kildinsaamischen Wörtern zutreffend sind, beweisen sie natürlich nicht, dass Plantinus auch ostsaamische Wörter in seine Liste aufgenommen hätte, sondern, dass diese Wörter früher auch im Schwedischsaamischen (Süd- oder Lulesaamischen) existiert haben. Gleichfalls kann es sich bei einigen dunkel gebliebenen um alte Wörter handeln, die später spurlos verlorengegangen sind, bei anderen wiederum um Wörter, deren Erkennung aus verschiedenen Gründen (noch) nicht gelungen ist.*

Wie aus meiner Darstellung zu ersehen ist, ist es meistens zwecklos, Plantinus' Wörter in den alphabetisch geordneten saamischen Wörterbüchern dem Schriftbild nach aufzusuchen. Am leichtesten findet man sie aufgrund der Bedeutung. Der lateinische Index in Leem 1781 gibt z. B. für ‚aqua' finnmarksaamisch *zhjatze* an, welches im heutigen Nordsaamisch *čáhci* geschrieben wird. Dasselbe Wort begegnet in verschiedenen phonetischen und orthografischen Varianten in allen saamischen Sprachen, im heutigen Südsaamisch *tjaetsie,* Lulesaamisch *tjáhtse* usw.; LL *tjatse* ‚aqua; vatten'. Bei Plantinus steht für ‚aqua' *kiatie,* welches ohne Zweifel dasselbe Wort ist. Mehr Mühe erfordert die Identifizierung des nächsten Wortes *kiergitt* für ‚aquila'. Leem gibt für dieses Wort *guasskem* an, ein im Saamischen weit verbreitetes Wort für Adler, offensichtlich aber nicht das, was Plantinus gemeint hat. Der schwedische Index in LL hilft nicht weiter, denn dort steht sub *örn* (= ‚aquila') *arnes* und *kåskem*. In dem norwegisch-südsaamischen Wörterbuch findet man endlich (s. v. *ørn*) ein passendes Wort *gïergehtse* und erfährt dazu, dass *k(i)* hier nicht für eine Affrikate sondern für einen Klusil steht. Es stellt sich heraus, dass dieses Wort auch in LL als *kergits* ‚aquila; örn' belegt ist, obgleich im Index darauf nicht verwiesen wird.

---

2 *håividhith* ‚verberare' (Tornaeus 1649); *raik* ‚gaudium; laetitia', *kagge* ‚ira' (Graan 1668).

3 Das Resultat: *åchte* ‚genus', *laiketh* ‚deserere', *tiackohieth* ‚placare', *karrath* ‚ornare' (Fjellström 1738: *ächt* ‚slächt', *lacketet* ‚öfwergifva', *tiotzkotet* ‚stilla wreden', *garfwot* ‚bepryda', *garfwet* ‚pryda').

4 *voszes* ‚animosus' (Südsaamisch *vuasehks* ‚dreist, kühn, zudringlich'), *pullames* ‚bulla' (*bollenjes* ‚Wasserblase'), *strappo* ‚cadaver' (*straahpoe* ‚die Leiche eines vom Raubtier getöteten Rentiers'), *sveibul* ‚flamma' (*svååjpele* ‚Flamme'), *kasnåth* ‚fomes' (*gasnege* ‚Zunder'), *girbe* ‚pudenda' (*girpie* ‚Penis eines Hundes'), *vaknerteth* ‚luctari' (*viengerdidh* ‚sich mit etwas sehr anstrengen'),*miacht* ‚res' (*mij-akt* ‚irgendwas'), *kaijeh* ‚sonus' (*gaajege* ‚Widerhall'), *harithe* ‚celer' (*haerrehtje* ‚schleunig, rasch, flink'; hierzu wohl auch *varitha* ‚celeriter'), *botnadt* ‚custodire' (*båtnodh* ‚verbergen, verstecken, verwahren, aufbewahren, sparen'),*beendeth* ‚tumescere' (*bonhtegidh* ‚anschwellen'), *kockedeth* ‚allicere' (*gåvhkoehtidh* ‚locken'), *jaebeteth* ‚aperire' (*gaehpiedidh* ‚öffnen, eine Öffnung, einen Riss machen'), *veividh* ‚cogere' (*meejvedh* ‚(Rentiere) zusammentreiben'), *galdij* ‚fons' (*gaaltije* ‚Quelle'); alle bis auf das letzte mit enger südsaamischer Verbreitung.

Wörter wie *aqua* und *aquila* sind in dem Sinne unproblematisch, dass ihre Bedeutungsäquivalente im Schwedischen, Norwegischen, Finnischen und Deutschen bei Bedarf anhand der zweisprachigen lateinischen Wörterbücher leicht herauszufinden sind, was die Voraussetzung für die Benutzung der existierenden zweisprachigen saamischen Wörterbücher ist. Die meisten lateinischen Wörterbücher orientieren sich zunächst nach dem klassischen Latein, Plantinus wiederum nach dem des 17. Jahrhunderts. So ist es nicht immer eindeutig, welche Bedeutung eines polysemantischen lateinischen Wortes Plantinus gemeint hat. Das erste Wort des Verzeichnisses ist *abies,* was im Lateinischen die Benennung eines Nadelbaums ist. Laut den lateinisch-deutschen Wörterbüchern bedeute es eine Tanne oder eine Fichte, von welchen nur die letztere in Frage kommen kann, denn die Tanne wächst im schwedischen Lappland nicht. Die Fichte heißt im Saamischen aber südsaamisch *goese,* nordsaamisch *guossa* usw. Dieses Wort begegnet bei Plantinus (*håsze* geschrieben) als Äquivalent für ‚pinus', welches laut den lateinisch-deutschen Wörterbüchern ‚Kiefer' bedeuten soll. Auch in LL steht für *kuose* lateinisch ‚pinus', obgleich aus der schwedischen Bedeutung (‚gran') zu ersehen ist, dass nicht die Kiefer sondern die Fichte gemeint ist. Die Kiefer heißt südsaamisch *bietsie,* nordsaamisch *beahci* usw. In LL ist *petse* ‚abies; tall, furu' belegt, woraus eindeutig hervorgeht, dass *abies* hier für ‚Kiefer' steht. Der Baumname *lätha* bei Plantinus bezieht sich also zweifelsohne auf die Kiefer, nicht die Fichte. Eine Behauptung, woran man wenigstens auf den ersten Blick viel mehr zweifeln kann, ist, dass es sich bei *lätha* und *petse* um ein und dasselbe Wort handelt. Nicht nur der Umstand, dass für dieses Wort im Saamischen sonst keine Vergleichspunkte zu finden wären, sondern die Beobachtung, dass das Verzeichnis auch andere Wörter mit dem Kopierfehler *l* pro *p* enthält (*langsing* ‚labium', *loide* ‚adeps'; LL *panksem* id., *puoite* ‚pingvitudo, lardum'), erlauben uns hier von einer ursprünglichen Form **pätha* (?**pätia, *pätza*) auszugehen. Auf die Problematik des Vokalismus lohnt es sich hier nicht näher einzugehen (siehe Fußnote 7), es sei nur kurz darauf verwiesen, dass die Diminutivform des Wortes im heutigen Lulesaamisch *bätsasj* lautet.

Als saamisches Bedeutungsäquivalent für lat. *insidiae* ‚Hinterhalt, Versteck zum Auflauern; Falle, Fallstrick' steht bei Plantinus *biutas.* Dieses gehört ohne Bedenken zusammen mit LL *piwto(sma)* ‚tendicula, laqueus aut alius eius generis, quo captura sit', *piwtosmab takket* ‚retia ponere aliasve insidias feris struere', nordsaamisch *bivddus* ‚Fanggerät' (zu LL *piwtet* ‚capturae avium vel piscium etc. operam navare'). Ein ähnlich aussehendes Wort steht einige Zeilen weiter unten für lat. *institutum* ‚Einsetzung, Einrichtung, Sitte, Gewohnheit, Unterweisung'. Hier handelt es sich jedoch wohl um ein ganz anderes Wort, und zwar LL *piejates, piejetus* ‚mandatum; befallning, förordnande'.[5] Das Stammwort von *piejetus* ist *piejet,* das laut LL ‚ponere; lägga, sätta' aber auch ‚ordinare; befalla' (*naute le Jubmel piejam* ‚ita Deus ordinavit') und ‚instituere; instikta' (*Lådnesteje le kastateseb sisa piejam* ‚Salvator Sacramentum Baptismi instituit') bedeutet. Dieses Verb begegnet bei Plantinus als *paijeth* ‚ponere' und *paieth* ‚iubere'. Eine Ableitung von demselben Verb ist auch Plantinus *paijemat* ‚jussum'. Das Suffix *-mat* ist für das Saami-

5 Die Verwendung des Wortes wird in LL mit folgenden Beispielen beleuchtet: kalka målsotet tait piejetusit ‚mutabit instituta illa' (ein Bibelzitat Act 6:14, in der Vulgata allerdings: mutabit traditiones) und Jubmelen piejetusen melte wiesot ‚ita vivere ut jussit Deus; lefva efter Guds befallning'.

sche befremdend, sodass es nicht ausgeschlossen ist, dass es sich hierbei um eine Entstellung von *päijetom, piejetum* o. ä. handelt. Eine solche Ableitung, die von demselben Typus wie *wårnotom* ‚juratus' (worüber unten) ist, begegnet bei Graan (1668: 60): [Weralden pijemusiij pargån ja] *piejetumen piira* ‚Om [thet Werdlsliga Regementet och] Instichtelse'. LL gibt als lateinisches Bedeutungsäquivalent für *piejetus* nur *mandatum* an, aber offensichtlich wären auch *institutum* und *jussum* (sowie die Nebenform *piejetum*) hinzuzufügen. Plantinus hat auch ein drittes zu diesem Begriffskreis gehörendes Wort: *vijehtes* (*vijchtes*) ‚lex'. LL gibt für schwedisch *lag* ‚Gesetz' auch *piejetus* an. Es stellt sich die Frage, ob nicht auch *vijehtes* auf dieses zurückzuführen wäre.[6]

Wenn auch die meisten lateinischen Wörter in den mir zur Verfügung stehenden Wörterbüchern aufzufinden sind, gibt es bei Platinus einige, die ich in keiner anderen Quelle habe belegen können. Möglicherweise handelt es hierbei um falsch geschriebene (und/oder kopierte) Wörter oder Neologismen mit enger Verbreitung. Zu den erst genannten scheint *cherus* für saamisch *kierk* (vgl. LL *keres, kerok* ‚carus, dilectus; kär', auch VR und Schroderus *charus* ‚käär') zu gehören, zu den letzteren *insipidus* (auch in VR *insipidum* ‚osmakande'), worüber weiter unten. Bei einigen Fällen hat sich das lateinische Wort durch Abschreibfehler geändert: so steht bei saamisch *vösze* und *sacke* lateinisch *sanus* pro *saccus* (vgl. LL *wuoss* ‚saccus' und südsaamisch *sïehke* ‚Sack'), bei saamisch *kijrz* lateinisch *guttus* pro *guttur* (vgl. LL *karas, kirs* ‚gula, cartilaginosa pars colli'), bei saamisch *brades* lateinisch *festivus* pro *festinus* (vgl. LL *brad(es)* ‚praeruptus, acclivis, celer'),bei saamisch *kiomatz* lateinisch *aridus (?)* pro *armus* (vgl. LL *tjåmotes* ‚quadrupedum armus'), wahrscheinlich auch bei saamisch *hiäg* lateinisch *angelus* pro *angulus* (vgl. LL *tjäk* ‚angulus'; *ängel* ‚angelus'; hier muss also auch das saamische Wort beim Kopieren entstellt worden sein). Wie bei *sanus/saccus* spricht auch bei *angelus/angulus* für die Richtigkeit des letzteren Wortes der Umstand, dass es (im Gegensatz zu dem ersteren) alphabetisch an der richtigen Stelle stehen würde. Es ist erwartungsgemäß, das die lateinischen Wörter, auch wenn sie falsch gelesen und kopiert worden sind, immerhin lateinische Wörter geblieben sind, während aus den – dem Kopierer wohl unbekannten – saamischen Wörtern Kauderwelsch geworden ist. Es ist möglich, dass es unter den bisweilen dunkel gebliebenen Wörtern noch mehrere Fälle gibt, in denen vor allem das geänderte lateinische Bedeutungsäquivalent das Problem ist. Ohne Kontext und ohne (richtige) Information über die Bedeutung sind die saamischen Wörter wegen der ungenauen und launischen Schreibweise meist unerkennbar, auch wenn sie „richtig" geschrieben (und kopiert) sind.

Beim Vergleich von Plantinus' Wörterverzeichnis mit den zeitgenössischen lateinisch-schwedischen Vokabularien ist die relativ große Anzahl der Verben (ca. 190 von Plantinus' 290 lateinischen Verben begegnen weder in VR noch bei Schroderus) auffällig. Sowohl die lateinischen als auch (mit wenigen Ausnahmen) die saamischen Verben stehen in der Infinitivform, deren Endung bei den meisten saamischen Wörtern *-th* geschrieben wird (bei 8 Verben jedoch *-dh*, bei 5 *-dt*, bei 4 *-d* und zweimal *t*). Auch

6 Bei Fjellström steht für schwedisch *lag* saamisch *lag, räcktie, biäjetz* und weiter *lagenbiäjetia* ‚lagstifftare' (= Gesetzgeber) sowie *biäjet* ‚stiffta'. Es ist auch möglich (und vielleicht sogar wahrscheinlicher), dass *vijchtes* eine Entstellung von *räcktie,* LL *rektas* ‚ius, forum; rätt, domstol', *laga ja rektas* ‚lag och rätt' ist (*v* pro *r* auch in *vöxe* ‚femur', LL *ruoksje* id.).

alle oben erwähnten in Schweden gedruckten Wörterbücher (VR, Schroderus, Fjellström, LL) führen die Verben in der Infinitivform an, während sie bei Leem in der Form der 1. Person Präsens stehen. Bei 30 saamischen Verben scheint die Form bei Plantinus jedoch nicht der Infinitiv zu sein. Die Formen *matta* ‚intelligere', *kauna* ‚reperire', *sava* ‚optari [!]', *sijta* ‚velle', *ähta* ‚amare' *bijla* ‚timere', *billa* ‚metuere', *gåive* ‚haurire' sind eindeutig 3. Person Singular (die ersten zwei begegnen im Verzeichnis dazu im Infinitiv: *mattith* ‚posse', *kaunath* ‚invenire'), mit Vorbehalt auch *jore* ‚rotari' (= *joreth* ‚volvere') und *bårre* ‚comedere'.[7] Derselben Kategorie dürften weiter die Verbformen *kiotha* ‚concitare', *nålla* ‚abigere', *laggå* ‚irasci' und *naldne* ‚titubare' angehören, aber bei ihnen ist nicht klar, um welches Wort es sich überhaupt handelt.

Im Verb *tabran* ‚hærere' ist das auslautende *n* ein Ableitungssuffix. Es handelt sich um den dreisilbigen Stamm LL *tabranet* ‚adhærere'; die Form dürfte 3. Sg. sein. Ein ähnlicher Fall ist möglicherweise *hiekin* ‚fremere', wobei das ganze Wort dunkel ist. Das Verb *kanest* ‚oscitare' ist ohne Bedenken eine (3. P. Sg.) Form des Verbs LL *kawestet* id., das Verb *mijkes* ‚singultire' wohl eine des Verbs LL *niakkestet* id. Bei *kattest* ‚pendere' handelt es sich entweder um LL *katsostet* ‚suspendere' oder LL *katsahet* ‚pendere' (die Bedeutung deutet auf ersteres, das Schriftbild auf letzteres, das im Verzeichnis auch als *kittesteth* ‚suspendere' begegnet). Das Wort *sinerves* ‚ringi' gehört mit den gleichbedeutenden Verben südsaamisch *snorvet,* lulesaamisch *(s)nirvvit* (vgl. auch LL *snerdset* id.) zusammen. Als Verb würde ihm südsaamisch *snjerviestidh* am nächsten stehen, das jedoch semantisch (‚brünstig sein') weiter entfernt ist. Es wäre vielleicht nicht ausgeschlossen, dass es sich bei *sinerves* nicht um ein Verb sondern um ein Adjektiv handelte, vgl. nordsaamisch *snirvvas* ‚grinsend'. Bei *sadgus* ‚mussitare' und *sumdkus* ‚fremere' handelt es möglicherweise um ein und dasselbe Wort, das eventuell mit *sabkasath* ‚sibilare' identisch ist (vergleiche jedoch auch südsaamisch *sjodkesjidh* ‚pfeifen (Wind)'). Der Stamm von *sabkasath* ist mit dem Stamm der im LL (und anderswo) belegten Ableitung *samkelet* ‚insusurrare, in aurem dicere' identisch (*-bk-* ~ *-mk-* ist ein interdialektaler Wechsel), die ein anderes Suffix enthält. Ein suffixales *-l-* steckt auch in *kalmell* ‚frigere' (vgl. LL *kalmet* ‚gelari', *kalmes* ‚frigidus'), sowie im ersten Teil der Konnexion *gioggel pijas* ‚surge'. Hier handelt es sich also sowohl auf Lateinisch als auch auf Saamisch um eine Imperativform (*pijas* ‚auf'). Daneben enthält das Wörterverzeichnis auch die Infinitivform *gioggeleth* ‚surgere' (sowie *kioggeleth* ‚erigere' pro ‚erigi'; vgl. LL *tjuodtjelet* ‚surgere', *tjuodtjel paijas* ‚erige te', *tjuodtjaldattet* ‚erigere, exsurgere facere'). – Aus zwei Wörtern bestehende Konnexionen sind weiter

7 Der Grund für den Vorbehalt ist der Auslautvokal *e* statt des zu erwartenden *a,* vgl. modernes Südsaamisch *jårra* (3. P. Sg. aus *jårredh* ‚rollen, sich drehen'), *bårra* (3. P. Sg. aus *bårredh* ‚fressen'). Einem einzelnen Buchstaben, insbesondere einem Vokal kann jedoch keine große Bedeutung beigemessen werden. Erstens sind die Vokale im Südsaamischen paradigmatischen Wechseln unterworfen (*bïlledh – bælla – billieh; gåajvodh – gåå-jve – gåajvoeh; iehtsedh – eahtsa – iehtsieh*). Zweitens weisen sie schon innerhalb des Südsaamischen interdialektale Variation auf (z. B. *jarredh, barredh* neben *jårredh, bårredh*), und das Bild wird noch bunter, wenn man auch andere saamische Sprachvarietäten (Ume-, Pite- und Lulesaamisch) berücksichtigt. Drittens begegnen im Saamischen Vokale (vor allem Diphtonge), bei denen es gar nicht evident ist, mit welchen lateinischen (oder schwedischen) Buchstaben sie zu bezeichnen wären. Und viertens muss man immer auch damit rechnen, dass das Schriftbild beim Kopieren entstellt worden ist (so erklären sich wohl die nicht seltenen Fälle, wo anstelle eines zu erwartenden *å* ein *ä* oder *a,* anstelle eines *ä* ein *å* oder *a* und anstelle eines *a* ein *ä* oder *å* steht).

raszataieth ‚rumpere' (LL *rasta tåjet* ‚confringere' s. v. *rasta* ‚trans'), *palaidienames* (LL *ädnami palet* s. v. *begrafva*), *sorritråckelet* ‚deprecari' (vgl. LL *särrit = erit* ‚in alium locum [= weg]', Leem *erit rokkadalam* ‚deprecor') und *postäithpotheth* ‚redire' (vgl. südsaamisch *bååstede* ‚zurück').

Bei einigen Wörtern stellt sich die Frage, ob es sich bei dem lateinischen und dem saamischen Wort um ein und dieselbe Wortklasse handelt. Das Wort *svolateth* steht im Verzeichnis als Äquivalent für lat. *furax,* obgleich es sich dabei wohl nur um das Verb LL *suoladet* ‚furari' handeln kann. Das ursprüngliche lateinische Wort wird also *furari* gewesen und *furax* ein Fehler beim Abschreiben sein. Ein analoger Fehler scheint bei *holgahieth* ‚concitatus' vorzuliegen: das saamische Wort ist offenbar mit dem südsaamischen Verb *holkesidh* ‚nachfolgen' identisch. Bei genauerer Betrachtung stellt sich weiter heraus, dass das ursprüngliche lateinische Wort wohl nicht *concitare* ‚(an)treiben, jagen, aufhetzen' sondern *comitari* (od. *comitare*) ‚begleiten' gewesen ist, nicht nur weil dieses semantisch besser passt, sondern auch, weil es im Verzeichnis zwischen *comedere* und *comperdere* steht.[8] Das Verb *concitare* begegnet außerdem weiter unten (an der richtigen Stelle), wo ihm, wie oben festgestellt, ein dunkles *kiotha* als saamisches Äquivalent zugeordnet wird. Dem saamischen Wort *plaidgaseth* sollte dem Verzeichnis zufolge lateinisch *fulgus* entsprechen. Ein solches Wort ist in Wörterbüchern unbelegt; die vom Schriftbild her am nächsten gelegenen Wörter, die ich finden kann, wären *fulgor* oder *fulgur* ‚Glanz, Blitz' und *fulgere* ‚glänzen, blitzen, funkeln'. Das saamische Wort ist anscheinend ein Verb und zunächst mit südsaamisch *pleajhkasjidh* ‚glänzen, funkeln, blinken' zu vergleichen.

Während die oben aufgezählten saamischen Wörter auf *-th* trotz ihrer angeblichen nominalen Bedeutung höchst wahrscheinlich Verben sind, gibt es andere, die mehr (*tiabuth* ‚collum', LL *tjäpot* id.; *eckith* ‚vespera', LL *ekked* ‚vesper') oder weniger eindeutige Substantive sind. Zu den letzteren gehören *mäneth* ‚exitus', *kolath* ‚examen', *jelith* ‚sumptus' und *jelijth* ‚expensa', deren Stämme zwar jeweils mit einem verbalen Stamm gleichzusetzen sind (vgl. LL *mannat* ‚proficisci, iter facere', *kullet* ‚audire', *jelet* ‚vivere'), deren *-th* aber ein deverbales nominales Ableitungssuffix sein wird. Die zwei letzten sind vielleicht mit der nordsaamischen Ableitung *ealádat* ‚Nahrung' (oder *ealáhat* ‚Lebensunterhalt') identisch. Nicht ausgeschlossen ist, dass es sich in dem einen oder anderen Wort um dasselbe Suffix handelt wie bei *iamate* ‚mors' (vgl. südsaamisch *jaemede* ‚Tod') und *jelidh* ‚vita' (vgl. südsaamisch *jielede* ‚Leben').[9] Zu den Verbalsubstantiven gehört anscheinend auch *måreth* ‚cura', vgl. LL *mårraha, mårrek* id., *mårrahet* ‚curare, curam gerere', *mårretet* id. Zu dem letztgenannten Verb würde dem Schriftbild nach ideal *morreteth* passen, das im Verzeichnis jedoch nicht für ‚curare' sondern ‚conari' steht. Um einen Fehler beim Kopieren scheint es sich nicht zu handeln, denn *conari* steht hier alphabetisch an der

8 Lat. *comperdere* ist in den mir zur Verfügung stehenden Wörterbüchern unbelegt und dunkel ist auch das entsprechende saamische Wort *kieketeth*. Dem Schriftbild nach würde es ideal zu LL *tjäketet,* frequ. zu *tjäket* ‚abscondere, occultare' (lulesaam. *tjiehkat* ‚verstecken, verbergen; geheimhalten', skoltsaam. *čiekkâd* ‚verbergen; vergraben') passen. Möglicherweise ist das ursprüngliche lateinische Wort *compercere* ‚ersparen, zusammensparen' gewesen. Laut Lagercrantz (1939: 86) kann nordsaam. *čiehkat* auch in der Bedeutung ‚sparen, aufbewahren' gebraucht werden. Die semantische Zusammengehörigkeit von ‚verbergen, verstecken, vergraben' und ‚aufheben, zurücklegen, für die Zukunft aufbewahren, sparen' ist naheliegend.

9 Im Verzeichnis steht an dieser Stelle „Vita *hägke* vivere *jelidh*", wobei nicht eindeutig ist, ob *jelidh* hier ‚vita' oder ‚vivere' bedeuten soll. Auf jeden Fall steht weiter unten noch einmal „Vivere *Jälith*".

richtigen Stelle, sowie auch *curare,* dessen saamisches Äquivalent *kattith* mit LL *kattet* ‚cavere, custodire; curam gerere' identisch ist. Auf das rätselhafte *morreteth* wird am Ende des Beitrags zurückgekommen.

Einige Wörter, die im Verzeichnis die Endung *-th* haben, gehören zu den sog. Karitivadjektiven, d. h. sie enthalten ein Suffix, das dem deutschen *-los* entspricht: *kialmeth* ‚cæcus', *pelieth* ‚surdus', *maimath* ‚liber' und (trotz der angeblichen substantivischen Bedeutung wohl auch) *jwrmath* ‚imprudentia' (also „augenlos", „ohrlos", „schuldlos", „vernunftlos", vgl. LL *tjalmete/bme, -s* ‚cæcus', *peljete/bme, -s,* ‚surdus', *mainete/bme, -s, -k* ‚innocens', *jerbmete/bme, -s* ‚insipiens, imprudens' zu *tjalme* ‚oculus', *pelje* ‚auris', *maine* ‚culpa, vitium; morbus; delictum', *jerbme* ‚cerebrum; intellectus'). Möglicherweise begegnet dasselbe Suffix (allerdings anders geschrieben) auch im Adjektiv *pakott* ‚rudis'. Es könnte mit LL *lakkot(e)s* ‚dimidius', *lakkots pargo* ‚opus cujus non nisi dimida pars absoluta est' identisch sein (*p* pro *l* falsch abgeschrieben). Dieses Wort wird ein Karitivadjektiv zu LL *lakke* ‚pars dimidia' sein, d. h. „dem die Hälfte fehlt". Das Karitivsuffix hat im Saamischen (und im LL) mehrere Varianten. Die oben aufgezählten Formen auf *-th* dürften der Variante *-tek* entsprechen, das Verzeichnis enthält aber auch der Variante *-tebme* entsprechende Formen: *almatijm* ‚insipidus' (wohl: ‚geschmacklos', vgl. (neu)lat. *insipiditas* ‚Geschmacklosigkeit' zu LL *almes* ‚dulcis', südsaamisch *aelmie* ‚würziger Geschmack'), *maimetijm* ‚insons' und möglicherweise noch *iwösetim* ‚rarus' und/oder *jeszetijm* ‚tepidus', die eventuell mit LL *assetes* ‚tenuis', südsaamisch *jissehts,* inarisaamisch *asettim* id. zu vergleichen wären. Hierbei handelt es sich um ein Karitivadjektiv zu LL *asse* ‚cutis, cuticula', südsaamisch *jissie* ‚Innenseite des Felles' mit der ursprünglichen Bedeutung ‚das eine dünne Innenseite hat (vom Fell)'. Das letztgenannte Wort würde lautlich fast ideal zu dem südsaamischen Wort passen, lateinisch *tepidus* könnte ein Abschreibfehler (pro *tenuis*) sein. Bei dem erstgenannten Wort, das mit größerem Vorbehalt hierzu gehört, wäre von einem entstellten Schriftbild und einer ungenauen Wiedergabe der Bedeutung (*rarus* ‚spärlich, nicht dicht (z. B. vom Haar eines Felles)') auszugehen.[10]

10 Vgl. LL *niarbe* ‚rarus, gles; tenuis, tunn', *niarbot* ‚rarefieri, rarescere, blifva tunn och gles', *qwolga niarbo* ‚pili rarescunt'. – Neben *maimath* ‚liber' und *maimetijm* ‚insons' begegnet im Verzeichnis noch *maimetime,* dessen Bedeutung mit ‚liberare' angegeben wird. Möglicherweise hat hier ursprünglich *liberatus* gestanden. Die Bedeutung wäre hier (wie bei *maimath*) als ‚von einer Schuld freigesprochen' o. ä. zu verstehen. – Anscheinend identisch mit *jwrmath* ist *jermet* ‚stultitia' (gleichfalls mit substantivischer Bedeutung!); vgl. LL *jerbmetes-wuot* ‚dementia, imprudentia' mit einem Ableitungssuffix *-wuot,* das im Saamischen abstrakte Substantive aus Adjektiven bildet. Dieses Suffix ist häufig auch bei Plantinus: *irmes våth* ‚prudentia' (LL *jerbmeswuot* id.), *pahas vath* ‚jnimicitia' (LL *pahaswuot* ‚malitia', *paha(s)* ‚malus'), *varres våth* ‚sanitas' (LL *warreswuot* ‚valetudo, integra'; zu *varres* ‚sanus', LL *warres* id.), *sudes wåth* ‚quies' (LL *säddoswuot* id.; zu *sieddâs* ‚tutus' und ‚securus', LL *säddos* ‚tranquillus'), *båreswothe* ‚senectus' (nordsaamisch *boaris* ‚alt', *boarisvuohta* ‚Alter'), *vöridvöt* ‚sobrietas' (LL *wuorredeswuot* id.; zu *vörid* ‚sobrius', LL *wuorredes* id.), *kalgos våth* ‚negligentia' und *kolgos våth* ‚pigritia' (LL *kålkoswuot* ‚pigritia, ignavia'; zu *kalgos* ‚negligens' und *kolgos* ‚piger', LL *kålgos* ‚tardus, segnis'), *arges våth* ‚timor' (LL *argeswuot* ‚timiditas'; zu *arges* ‚timidus', LL *arge(s)* id.), *arves våtes* ‚terror' (zu *arvos* ‚pavidus'; vgl. südsaamisch *asvoeh* ‚entsetzlich, schrecklich'?), *skamoszvåth* ‚verecundus' (mit einer angeblichen adjektivischen Bedeutung ‚schamhaft, schüchtern' statt der zu erwartenden **verecundia* ‚Scham(gefühl), Schüchternheit'; falsch kopiert?; vgl. LL *skabmokes* ‚pudendus', *skabmokeswuot* ‚turpitudo'). Trotz der substantivischen Bedeutung stehen in einer adjektivischen Form anscheinend *sadnes* ‚veritas' (LL *sadneswuot* id., *sadnes* 'verus'), eventuell auch *jallo* ‚invidia' (falls es sich hierbei um dasselbe Wort handelt, wie *jallos* ‚inimicus'; zu der Wortfamilie des erstgenannten gehört jedenfalls *jallolit* ‚invidere', unklar ist hingegen, ob diese Wörter irgendwie mit LL *jall(a)* ‚stultus, stolidus', *jalla-wuot* ‚stultitia', *jallastet* ‚stultum agere' zu verbinden sind).

Während *rasatijm* ‚instructus' und *podetijm* ‚nuntius' mit keinem saamischen Wort(stamm) zu identifizieren sind, scheint *pudetijm* ‚tabes' mit folgenden Wörtern zu verbinden zu sein: LL *påd(w)o* ‚inquinamentum, sordes, omne istud, quod inquinat', *pådw(ot)et* ‚contaminare, inquinare'. Hier handelt es sich nicht um ein Karitivadjektiv, sondern um ein Verbalsubstantiv auf *-m* (zunächst des dreisilbigen Stammes *pådwotet*). Zu derselben Kategorie gehören die zwei „biblischen" Wörter *hiettetalim* ‚similitudo' (zu LL *sjättetallet* ‚applicare, congruere facere, aptare') und *hegge labma* ‚tentatio' (LL *kättjelem* ‚tentatio, periculum' zu *kättjelet* ‚periclitari, tentare'). Das letztere (fälschlich in zwei Wörtern geschriebene) Wort ist anscheinend eine Illativform und entstammt dem Vaterunser (et ne nos inducas in tentationem; im ABC-Buch 1638 *kiäggielebma,* s. genauer Bergsland 1982: 15). Das erstere begegnet (allerdings in einer kürzeren Form *hiettedem*) bei Graan 1669, nicht aber in LL, wo für ‚similitudo' *muotolwas* angegeben wird. Zwei Verben stehen in einer auf *-m* auslautenden Form, die die 1. Person Sg. sein könnte: *munnum* ‚lactere' und *vtigem* ‚minuere'.[11] Möglich wäre aber, dass es sich auch hierbei um Verbalsubstantive handelt, vgl. LL *utsanem* ‚ipse actus, dum quid minuitur, sive minus sit, imminutio' (zu *utsanet* ‚minui').[12] Neben den Verbalsubstantiven haben die (aktiven und passiven) Partizipien des Präteritums eine Endung auf *-m.* Hierher scheinen folgende Wörter zu gehören: *vaddijm* ‚donum', *valtuin* (*valtum?*) ‚conjugium', *sijlome* ‚lassus' (vgl. LL *waddes* ‚donum', *waldom* ‚nuptus', *sillom* ‚fessus' zu *waddet* ‚dare', *waldot* ‚uxorem ducere, nubere', *sillot* ‚defatigari'), sowie *pädohijm* ‚furibundus' zu *pädohieth* ‚furere' (vgl. LL *piädat* ‚insanire, stultum esse', *piädatet* ‚injicere alicui amentiam') und *wårnotom* ‚juratus' (vgl. LL *wuordnes* ‚jus jurandum', *wuordnom* id., *wuordnot* ‚jurare', *wuordnotet* ‚jus jurandum imponere').

3.3 Zur Semantik

Der Hauptteil der konkreten Substantive in Plantinus' Wörterverzeichnis kann in folgende semantische Gruppen eingeteilt werden: (i) leblose Natur (Zeit, Licht, Wetter, Gelände u. dgl.), (ii) lebendige Natur (Pflanzen, Tiere), (iii) Anatomie (Organe, Körperteile u. dgl.), (iv) Physiologie (Nahrung, Krankheiten, Sekrete), (v) Mensch (Alter, Geschlecht, Verwandschaft, Beruf u. dgl.), (vi) Materielle Kultur (Bauten, Fahrzeuge, Kleider, Instrumente u. dgl.). Bei den Substantiven mit abstrakter Bedeutung handelt es sich zum großen Teil um Ableitungen von Adjektiven und Verben. Beispiele dafür sind

11 Bei dem ersteren ist trotz des entstellten Schriftbildes (süd)saamisch *njammedh* (1. P. Sg. *njammem*) ‚saugen' zu erkennen, das letztere (oder wenigstens sein Stamm) dürfte mit LL *utsetet* ‚minuere', (süd)saamisch *uhtjiedidh* (1. P. Sg. *uhtjedem*) ‚vermindern' zu verbinden sein.

12 Mehr oder weniger sicher gehören zu den Verbalsubstantiven auf *-m* weiter folgende Wörter: *koikolim* ‚sitis' (zu *koikolith* ‚sitire'; LL *kåikeles ~ kåikelwas* ‚sitis', *kåikelet* ‚celeriter arescere', *kåikeluet* ‚sitire'), *jåkijm* ‚haustus' (LL *jukkem* ‚potatio' zu *jukket* ‚bibere'), *höfvijan* ‚laus' (zu *höffvidh* ‚laudare', südsaamisch *heevedh* ‚beloben'; LL *hewetem ~ hewetes* ‚laus', *hewetet* ‚laudare'), *sijtom* ‚voluntas' (zu *sijta* ‚velle'; LL *sit(t)em ~ situd* ‚voluntas', *sit(t)et* ‚velle'), *kailekothin* ‚jocus' (zu *kailakotith* ‚jocari'; vgl. LL *skalkestallet* ‚nequitiam perpetrare', *skalk* ‚nequam, nebulo', nordsaamisch *skálkkastit ~ skálkošit* ‚Späße machen', *skálka* ‚Spaßmacher, Schelm'), *spallim* ‚alapa' (LL *spekkestem* ‚plaga, alapa', *niärab spekket* ‚alapam impingere' s. v. *spekket* ‚vola pulsare sive percutere'; *njär* ‚genæ, mala'), *kiäres potin* ‚acceptus' (vgl. LL *påtem* ‚adventus', *keres* ‚carus, dilectus', lulesaamisch *buorisboahtem* ‚willkommen'), *maitim* ‚sapor' (zu *maitzateth* ‚gustare', LL *maistet* id.), *siöggiom* ‚reditus' (dunkel).

oben (Fußnoten 10 und 12) angeführt. Besondere Aufmerksamkeit unter den abstrakten Substantiven verdienen Wörter, die zum Begriffskreis (vii) der geistigen Kultur (Religion, Gesellschaft) gehören. (Appendix 1)

Obgleich die meisten Verben einen Menschen oder wenigstens ein Lebewesen als prototypisches Subjekt voraussetzen, kann das Subjekt bei einem Teil der Verben auch ein Ding, eine Naturerscheinung usw. sein. Zwei Gruppen zeichnen sich ab: (viii) intransitive Verben und (ix) aktitive, kausative und transitive Verben, die (auch) mit einem nicht belebten Subjekt stehen können. Ein belebtes (nicht unbedingt menschliches) Subjekt voraussetzende Verben gehören zu folgenden semantischen Gruppen: (x) Verben für physiologische Tätigkeiten, (xi) spatiale Verben (Bewegung, Dasein, Orientierung u. dgl.), (xii) faktitive, kausative und transitive Verben mit einem prototypisch belebten Subjekt. Die Verben, die einen Menschen als prototypisches Subjekt voraussetzen, bilden folgende Gruppen: (xiii) faktitive und kausative Verben (arbeiten, sich bemühen u. dgl.), (xiv) possessive und habitive Verben (besitzen, bekommen, verlieren, dienen u. dgl.), (xv) mentale und kognitive Verben (denken, fühlen, ausdrücken u. dgl.). (Appendix 2)

Die Adjektive bei Plantinus können je nach dem, ob sie sich (xvi) auf eine Eigenschaft eines Menschen (oder eines Tieres) beziehen oder (xvii) auch in Bezug auf nicht belebte Gegenstände benutzt werden können, in zwei Gruppen eingeteilt werden. (Appendix 3). Bei den Numeralien handelt es sich um die Kardinalzahlen von 1 bis 10, wobei für lateinisch *duo* saamisch *weestes* steht, welches wohl eher die Ordinalzahl ‚primus' sein wird, vgl. LL *wuostes* ‚primus', *qwekt(e)* ‚duo' (aber auch *qwektes* id.). Bei den Partikeln handelt es sich um Adverben für Raum und Zeit, grammatikalische Wörter, Präpositionen u. dgl.

4 Schlussfolgerungen

Das Wörterverzeichnis von Z. Plantinus fällt zeitlich mit dem Katechismus und dem Manuale von O. S. Graan, der Lapponia von J. Schefferus und dem Verzeichnis der finnischen Wörter von M. Fogel zusammen. Dem letztgenannten und Plantinus' Verzeichnis ist gemeinsam, dass es sich bei ihnen um eine handgeschriebene Wortliste handelt, die auf Verlangen einer bestimmten Person für wissenschaftliche Zwecke erstellt wurde. Im Gegensatz zu Fogel, der sein Material direkt aus VR entnommen hat, hat Plantinus, was das lateinische Material anbetrifft, weder dieses noch das Lexikon von E. Schroderus als Vorlage benutzt. Von den ca. 850 lateinischen Wörtern bei Plantinus sind nämlich nur ca. 360 in VR und ca. 430 bei Schroderus belegt; ca. 300 Wörter sind sowohl in VR als auch bei Schroderus unbelegt.

Eine unerklärte Frage ist, wie sich das hier analysierte Wörterverzeichnis zu der Handschrift eines Lexicon Lapponicum von Plantinus verhält, dessen Vorrede von Schefferus zitiert wird. Wie oben (Abschnitt 3.1) erwähnt, hat Setälä auch die von Schefferus benutzte Vorrede gefunden, allerdings an einer anderen Stelle, sodass er es für unsicher ansieht, ob sie zu diesem Wörterverzeichnis oder zu einer

verlorengegangenen größeren Arbeit gehört. Wie das Wörterverzeichnis ist auch die Vorrede nicht von Plantinus selbst sondern von einer fremden Hand niedergeschrieben. Obgleich Setälä das nicht ausdrücklich sagt, muss die Handschrift der Vorrede wohl eine andere als die des Verzeichnisses sein, denn sonst hätte man ja keinen Grund, an ihrer Zusammengehörigkeit zu zweifeln. Eine Möglichkeit ist, dass die erhalten gebliebenen Dokumente zwei zu verschiedenen Zeiten und für verschiedene Zwecke angefertigte partielle Kopien (die eine für Schefferus und die andere für Stiernehielm) aus der Originalhandschrift von Plantinus darstellen. Eine andere Möglichkeit ist, dass es sich bei der Vorrede, die auch Plantinus' eigenhändige Ergänzungen und Berichtigungen enthält, um den Anfang eines für die Druckerei bestimmten Dokuments handelt; ungewiss ist dabei, ob es ursprünglich länger gewesen ist und auch den Wörterbuchteil umfasst hat; falls ja, so könnte das uns überlieferte Verzeichnis daraus kopiert worden sein. Als Vorlage der Kopie hat natürlich auch irgendeine frühere oder spätere Arbeitsphase dienen können. Und wie es damit auch bestellt sein mag, ist es möglich, dass es sich bei der an Stiernhielm versandten Wortliste um einen Auszug (oder eine Auslese) aus einem größeren Wörterbuchmanuskript handelt.

Im 15. Kapitel der Lapponia (De Lingva & Sermone Lapponum) führt Schefferus (1673: 178-182) drei Wortlisten mit 52 saamischen Wörtern an. Von ihnen kommen 11 bei Plantinus nicht vor, und auch die dort begegnenden 41 Wörter sind insofern anders geschrieben, dass ihre primäre Quelle augenscheinlich nicht Plantinus ist. Zu den Informanten von Schefferus gehörten auch saamische Muttersprachler, darunter der schon erwähnte Olaus Stephani Graan (Pastor in Lycksele) und sein Namensvetter Olaus Graan (Pastor in Piteå) sowie Olaus Sirma (Student und später Komminister in Enontekis), von dem die zwei durch die Lapponia später weltberühmt gewordenen saamischen Lieder stammen. Da die von Schefferus angeführten Wörter nördliche und östliche Merkmale aufweisen (*ännä* ‚mater', *Immel* ‚Deus', *riemnes* ‚vulpes', vgl. Plantinus *ädne, Jubmel,* LL *edne, Jubmel, repe*), ist es wahrscheinlich, dass Schefferus auch die Wörter von dem in Kemi-Lappland aufgewachsenen Olaus Sirma Angaben bekommen hat. (Zur Heimat von Sirma s. auch Itkonen 1940.)

Wie gut Plantinus selbst das Saamische beherschte, wissen wir nicht. Muttersprachler war er nicht, möglicherweise hatte er aber schon von Kindheit an Kontakt mit der Sprache. Es ist auch möglich, dass das Verzeichnis in der einen oder anderen Weise eine Kollektivarbeit ist. Wie schon Qvigstad (1947: 39) bemerkt, steht die Sprache in Plantinus' Wörterverzeichnis der in den Werken von O. S. Graan nahe, was sowohl aus zeitlichen als auch geografischen Gründen natürlich ist. Dies gilt vor allem für die Schreibweise der Wörter. Über das lexikalische Verhältnis von Plantinus' und Graans Arbeiten kann nichts Absolutes gesagt werden, weil Graans Wortschatz (besonders das Manuale) noch nicht systematisch erforscht ist. Da das Hauptgewicht in Plantinus' Verzeichnis deutlich auf der profanen (nicht kirchlichen, religiösen oder biblischen) Lexik liegt, ist anzunehmen, dass die Wörter nicht aus der gedruckten Literatur exzerpiert wurden, obgleich die zeitgenössischen Druckwerke dem Zusammensteller und seinen Informanten natürlich bekannt gewesen sind. Falls Plantinus seine Arbeit an dem Wörterverzeichnis schon in Hernösand angefangen hat, könnte ein Teil des Wortschatzes eventuell von saamischen Schülern in Hernösand herstammen. Auf den heterogenen Ursprung des Mate-

rials deutet, dass ein Teil der Verben nicht im Infinitiv angeführt sind, und vielleicht auch, dass viele saamische Wörter im Verzeichnis verschieden geschrieben als Bedeutungsäquivalente für mehr als ein lateinisches Wort begegnen.

Es ist möglich, dass das Verzeichnis (wenigstens zum Teil) auf Material basiert, in dem die Metasprache ursprünglich nicht Lateinisch, sondern Schwedisch gewesen ist. Dadurch würden zwei rätselhafte Wörter einleuchtend werden. Für lateinisch *brevis* steht im Verzeichnis sowohl *aniki* (LL *åne(kes)* ‚brevis; kort') als auch *spelek.* Letzteres widersetzt sich allen Vergleichen mit Wörtern mit der Bedeutung ‚kurz', ihm kommt aber sehr nahe LL *spiäl, spjäla* ‚chartae lusoriae; kort', also ‚Spielkarte(n)'. Wie ersichtlich, werden diese zwei Wörter im Schwedischen in gleicher Weise geschrieben, sodass die Verwechslung der Bedeutung verständlich ist. Ein anderer Fall kann das oben erwähnte *morreteth* sein, das im Verzeichnis für *conari* statt eines anzunehmenden *curare* steht. Das schwedische Bedeutungsäquivalent von *mårretet* ist bei Fjellström *försörja* (= ‚(ver)sorgen'). Lateinisch *conari* (= ‚versuchen') heißt auf Schwedisch *försöka,* veraltet auch *försökja.* Diese Wörter sind zum Verwechseln ähnlich, ein falsch gedeuteter Buchstabe hätte die Bedeutung leicht ändern können.

Es sieht so aus, dass sich in das Verzeichnis in verschiedenen Arbeitsphasen Fehler eingeschlichen haben, sodass auch die weiter oben besprochenen nicht alle dem Hersteller der uns erhalten gebliebenen Kopie zuzuschreiben sind. Trotz seiner Mängel und Ungenauigkeiten zeugt das Verzeichnis von guter Kenntnis des Saamischen und Lateinischen beim Verfasser (und/oder seinen Informanten). Plantinus' Arbeit blieb seinerzeit ungedruckt, und über ihre Bedeutung für die spätere saamische Lexikografie wissen wir nichts Genaueres. Es ist möglich, dass das Manuskript noch im nächsten Jahrhundert von Fjellström benutzt wurde. Die von mir durchgeführte Analyse hat jedoch keine direkte Indizien dafür vorlegen können. Sicher ist hingegen, dass die von Schefferus angeführten saamischen Wörter nicht Plantinus entstammen. Somit ist auch wahrscheinlich, dass ihm nur die Vorrede zu Plantinus Lexicon zur Verfügung gestanden hat.

5 Literatur

Aapala, K., Koponen, E., Ruppel, K. (2010). Überblick über die Geschichte, den gegenwärtigen Stand und die Zukunftsperspektiven der etymologischen Forschung des Saamischen (Lappischen) in Finnland. In *Studia etymologica Cracoviensia,* 15, pp 7-14.

Álgu-Datenbank. Álgu-tietokanta. <kaino.kotus.fi/algu>

Bergsland, K. (1982). Den svensk-samiske ABC fra 1638 som sproghistorisk dokument. In *Språkhistoria och språkkontakt i Finland och Nord-Skandinavien. Studier tillägnade Tryggve Sköld den 2. november 1982.* Kungl. Skytteanska Samfundets handlingar 26. Umeå, pp. 11-20.

Fjellström, P. (1738). *Dictionarium Sueco-Lapponicum.* Stockholm.

Graan, O. S. (1668). Korta och Enfaldiga Spörsmåhl Öfwer Catechismum. Stockholm.

Graan, O. S. (1669). *Manuale Lapponicum.* Stockholm.

Itkonen, E. (1940). Olaus Sirman kotiseudusta ja kielestä. In *Virittäjä* 44, pp. 334-349.

JSFOu = Journal de la Societé Finno-Ougrienne. Helsinki.

Koponen, E. (2010). Johannes Tornaeuksen Manuale Lapponicumin murrepohjasta ja vaikutuksesta ruotsinlapin kirjakielen myöhempään kehitykseen. In *Sanoista kirjakieliin. Juhlakirja Kaisa Häkkiselle 17. marraskuuta 2010.* MSFOu 259, pp. 43-52.

Larsson, L.-G. (1997). Prästen och ordet. Ur den samiska lexikografins historia. In *LexicoNordica* 4, pp. 101-117.

Leem, K. (1768, 1781). *Lexicon Lapponicum bipartitum.* Nidrosiae.

LL = Lindahl, E., Öhrling, J. (1780). *Lexicon Lapponicum.* Holmiae.

Löw, B. (1956). Einleitung und Kommentare zum Text in *Johannes Schefferus, Lappland.* Översättning från latinet av Henrik Sundin. Granskad och bearbetad av John Granlund, Bengt Löw och John Bernström. Nordiska museet. Acta Lapponica 8. Uppsala, pp. 9-23, 408-408.

Magga, O.-H. (2012). Lexicography and indigenous languages. In *Euralex 2012 Proceedings,* pp. 3-18.

MSFOu = Mémoires de la Société Finno-Ougrienne. Helsinki.

Qvigstad, J. (1899). Übersicht der geschichte der lappischen sprachforschung. In *JSFOu* 16:3, pp. 11-29.

Qvigstad, J. (1947). Sproget i Graans Manuale Lapponicum. In *Studia Septentrionalia* 3. Oslo, pp. 8-39.

Qvigstad, J., Wiklund, K. B. (1899). *Bibliographie der lappischen Litteratur.* MSFOu 13.

Schefferus, J. (1673). Lapponia Id est, regionis Lapponum et gentis nova et verissima descriptio. Francofurti.

Schroderus, E. (1637). Lexicon Latino-Scondicum. Holmiae Sueonum.

Setälä, E. N. (1890). Ein lappisches wörterverzeichnis von Zacharias Plantinus. In *JSFOu* 8, pp. 85-104.

Sköld, T. (1986). Pehr Fjellström och det svensksamiska skriftspråket. In *Saga och sed* 9, pp. 15-26.

Stipa, G. J. (1990). *Finnisch-ugrische Sprachforschung von dem Renaissance bis zum Neopositivimus*. Redaktionelle Bearbeitung und Zusammenstellung der Bibliographie Klaas Ph. Ruppel. MSFOu 206.

Tornaeus, J. (1648). *Manuale Lapponicum.* Stockholm.

VR = Variarum Rerum Vocabula Latina (1579, 1668).

Wiklund, K. B. (1922). De första lapska böckerna. In *Nordisk tidskrift för bok- och biblioteksväsen* 9. Uppsala, pp. 13-28.

Appendix 1

(i) tempus, aetas, annus, ver, aestas, autumnus, hyems, dies, meridies, vespera, nox, lux, umbra, tenebrae, radius, sol, luna, stella, coelum, sonus, strepitus, ventus, tempestas, nubes, nebula, aqua, glacies, nix, grando, pluvia, gutta, ros, bulla, spuma, foetor, odor, nidor, fumus, fuligo, ignis, flamma, pruna, carbo, pulvis, lutum, coenum, arena, lapis, aes, ferrum, locus, terra, sylva, desertum, mons, clivus, vallis, fossa, littus, insula, lacus, flumen, fons;

(ii) arbor, arbuscula, abies, pinus, radix, cortex, virga, ramus, gummi, arundo, foenum, muscus, animal, alces, canis, equus, fiber, lupus, mus, mustela, taurus, sciurus, avis, aquila, corvus, piscis, perca, serpens, vermis, pediculus, musca, culex;

(iii) corpus, cadaver, latus, dorsum, gibbus, sinus, pectus, umbilicus, pudenda, venter, cor, jecur, renes, vesica, *armus, humerus, brachium, manus, pugnus, digitus, pollex, pes, femur, genu, poples, sura, cervix, collum, *guttur, caput, vultus, nasus, mala, mentum, auris, oculus, pupilla, os ‚Mund', labium, lingva, dens, os ‚Knochen', pellis, sangvis, crinis, ungvis, penna, cornu;

(iv) cibus, caro, adeps, sebum, butyrum, caseus, lac, farina, bacca, ovum, fames, sitis, somnus, angor, dolor, scabies, livor, tussis, mucus, sudor, urina;

(v) homo, vir, senex, foemina, puella, pater, mater, liberi, foetus, filius, filia, frater, soror, procus, sponsus, sponsa, conjugium, uxor, vidua, viduus, affinis, gener, socer, socerus, nuntius, dux, mercator, venator, fautor, famulus, subditus, possessor, debitor, mendicus, latro, rusticus, heros, amicus, adulter, claudus, majores, genus, populus, turba;

(vi) domus, ostium, columna, porticus, gradus, *angulus, focus, saeptum, vallum, pons, via, cymba, traha, jugum, habena, lorum, vinculum, funis, nodus, vestis, lacinia, pannus, tunica, balteus, calceus, chirotheca, anulus, *saccus, rete, insidiae, hamus, arcus, sagitta, hasta, clavis, cochlear, acus, forfex, forceps, malleus, pistillum, cuneus, ligo, ignarium, silex, fomes, cos, gluten, scipio, lectus, patina, globus, liber, icon;

(vii) Deus, oratio, spiritus, superstitio, praerogativa, spectrum, bellum, pax, rixa, auxilium, labor, merces, meritum, tributum, pars, magistratus, licentia, institutum, lex, jussum, poena, poenitentia, mos, consuetudo, gestus, opinio, mendacium, veritas, verbum, sermo, colloquium.

Appendix 2

(viii) incipere, fluere, diescere, *fulgere, splendere, sibilare, fremere, defervere(?), haerere, se incurvare, tumescere, tremere, rotari, cadere, silere, desinere;

(ix) necare, occidere, frangere, lacerere, rumpere, perfundere, siccare, incendere, concremare, urere, extinguere, inficere, minuere, lustrare, vertere, sufflare;

(x) vivere, fieri, crescere, mori, perire, respirare, comedere, mordere, mandere, rodere, bibere, lactere, sorbere, deglutire, lingere, gustare, odorari, intueri, videre, audire, spuere, ructare, vomere, cacare, tussire, sternutare, oscitare, ringi, scabere, pati, frigere, esurire, sitire, aegrotare, dormire, vigilare, expergisci, fatigere, quiscere;

(xi) ire, currere, irruere, migrare, proficisci, pergere, venire, redire, sequi, natare, volare, luctari, titubare, scandere, surgere, stare, iacēre, sedere, habitare, manere, exspectare, errare, quaerere, invenire, reperire, occurrere;

(xii) facere, parere, pascere, cibare, trahere, portare, levare, haurire, tundere, fricare, relinquere;

(xiii) laborare, conari, probare, tentare, moliri, ponere, suspendere, sepelire, occultare, iacere, iaculari, praecipitare, reicere, verberare, abigere, cogere, *comitari, concitare, congregare, deducere, pellere, venari, ludere, aedificare, statuere, tendere, erigere, tegere, suere, secare, scribere, pungere, ornare, induere, aperire, confirmare, munire, excavare, extendere, excoriare, ungere, miscere, lavare, coquere, assare, implere, consumere, uti, finire, molestare, dolore afficere, simulare;

(xiv) accipere, captivare, sumere, dare, tradere, donare, egere, possidere, tenere, retinere, *furari, decipere, mendicare, liceri, vendere, emere, mercari, solvere, liberare, servare, custodire, curare, observare, parcere, *compercere, desolare, deserere, auxiliari, parēre;

(xv) noscere, scire, recordari, intelligere, posse, consuescere, discere, opinari, cogitare, conicere, desiderare, optare, velle, poenitere, imaginari, excogitare, assentiri, approbare, despicere, detestari, suspicari, dubitare, mirari, invidere, amare, metuere, timere, irasci, dicere, narrare, depraedicare, notificare, interrogare, respondere, defendere, invocare, iubere, docere, hortari, suadere, dissuadere, negare, consultare, consulere, promittere, iurare, mentiri, fateri, orare, deprecari, gratias agere, laudare, accusare, lacessere, maledicere, illudere, ineptire, [se] iactare, suberbire, garrire, iocari, valedicere, blandiri, placare, solari, minari, terrere, ulisci, allicere, mussitare, queri, gemere, singultire, flere, deflere, clamare, vocare, rixari, furere, cachinnare, ridere, gaudere, laetari, canere, saltare, osculari, amplecti.

Appendix 3

(xvi) obligatus, reus, insons, liber, iuratus, dives, egenus, miser, tenax, *carus, honestus, gratiosus, benevolus, concors, gratus, promptus, contentus, mansuetus, docilis, prudens, diligens, castus, probus, pius, cautus, pavidus, timidus, verecundus, audax, temerarius, anomosus, furibundus, inimicus, malevolus, contumax, ferox, superbus, imprudens, fallax, negligens, petulans, impudicus, impudens, improbus, insipiens, fatuus, stultus, piger, maestus, laetus, macer, nudus, natus, robustus, fortis, imbecillus, agilis, *festinus, sanus, somnolentus, lassus, sobrius, ebrius, surdus, caecus, mutus, balbus;

(xvii) simplex, solus, vulgaris, exiguus, sinister, totus, integer, rotundus, rectus, erectus, curvus, inversus, altus, latus, longus, brevis, crassus, rarus, densus, mollis, durus, gravis, levis, vacuus, plenus, tumidus, acclivis, profundus, pallidus, luridus, obscurus, niger, albus, ruber, viridis, versicolor, novus, recens, crudus, mucidus, insipidus, amarus, acidus, dulcis, humidus, siccus, frigidus, calidus, fervidus, constans, tener, remissus, munitus, rudis, erroneus, deformis, ineptus, nequam, malus, ridiculus, extraneus, mirabilis, bonus, acceptus, pulcher, pretiosus, precipuus, sanctus, verus, selectus, utilis, notus, secretus, securus, tutus.

Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses

Gabriele Langer, Susanne König, Silke Matthes
Institute of German Sign Language and Communications of the Deaf, University of Hamburg
Gabriele.Langer@sign-lang.uni-hamburg.de,
Susanne.Koenig@sign-lang.uni-hamburg.de,
Silke.Matthes@sign-lang.uni-hamburg.de

Abstract

Nowadays lexicographic work such as lemma selection, identification of word senses and usage information are usually based on large corpora. Sign language lexicographers face the same difficulties as their colleagues of other less studied and previously not written minority languages. They cannot rely on written texts and large corpora. In addition sign language linguists have to cope with sign-language specific issues due to the visual-gestural modality, namely the iconicity of signs, a broad utilization of simultaneity of linguistic signals and the integration of lexical material of a spoken language by the way of mouthings.

Sign language specific lexicographic issues are discussed and exemplified with regard to the Basic Vocabulary of German Sign Language (DGS) that is compiled within the larger context of the DGS Corpus Project. The Basic Vocabulary is not corpus-based but based on previously published sign collections that are used as a starting point. The data is reviewed, sign senses are finer split and disambiguated and the data then undergo a validation process by the sign language community. This validation process is conducted through a feedback system especially designed for surveys involving sign language.

Keywords: lexicography of sign language; German Sign Language (DGS); word senses; feedback system; crowd sourcing; basic vocabulary; iconicity; mouthing

1 Background

1.1 Sign Language Dictionaries

Sign language lexicography is still in the process of striving for the best lexicographic methods to analyse signs, as well as solutions to document and present them in an adequate and user-friendly way. For meaning analyses and descriptions there are two aspects especially relevant in a dictionary pro-

ject: First, what kind of data is the given information based upon, and, second, how can the identified meanings of signs (senses) be made clear and best presented to the dictionary user.

In the past – due to a lack of an everyday writing system and written sign texts – sign dictionaries have hardly been in the position to base their sign selection and information on sign use on corpus analyses. Up to the present, sign dictionaries and sign collections have mainly been based on information drawn from earlier sign collections, word-to-sign elicitations, the introspection of single signers or groups of signers (and their metalinguistic discussions on signs' meanings and usage), on participant observation, or on a mixture of these methods (compare for example Johnston 1989: 8, Brien 1992: x-xi, Kristoffersen & Troelsgård 2010: 5). To our knowledge the first sign language dictionary largely and systematically based on a collection of filmed and transcribed sign language data was the ASL Dictionary published in 1965 (Stokoe, Casterline & Croneberg 1965). Considering the limited technical possibilities at the time (cf. Stokoe 1993) this work was outstanding in many regards and way ahead of its time.

Technical progress in recording equipment and transcription software now makes the collection of relatively large sign language corpora feasible. Sign corpora exist or are being collected in several countries such as Australia, Germany, Great Britain, Italy, the Netherlands, Poland, France and others. Most of these corpora are still in the stage of data collection or transcription and annotation. With the availability of larger annotated sign corpora the information in sign dictionaries – especially on the meanings and usage of signs – can reach new levels of quality.

Without a fully functioning writing system for signs (see below) some effort has to be put into the presentation of sign forms and how to provide a search by sign form in order to ensure bi-directional accessibility in sign dictionaries. However, how to best identify meaning(s) of signs and to make them accessible has not been discussed in great depth so far in sign language lexicography. Common to all general sign dictionaries that we are aware of is the bilingual solution of expressing the meanings of signs by matching translation equivalents (keywords) of the surrounding spoken language. In addition to that some dictionaries also provide meaning explanations (e.g. Johnston 1989), usage notes (e.g. Stokoe, Casterline & Croneberg 1965) or example-like contexts (e.g. Kennedy 2008, cf. esp. xvii) – all using the surrounding spoken language as meta-language – or signed examples (e.g. *Ordbog over Dansk Tegnsprog*).

In this paper we differentiate between sign dictionaries and more simple sign collections. A sign dictionary regards the signs as the linguistic units to be described with information about their form, grammar, meaning and usage. Sign language dictionaries of this kind have sign entries and – if necessary – include several senses of a given sign. Sign collections, on the other hand, are predominantly mono-directional word lists with the direction from spoken language words to signs. Usually, they contain no real sign entries and do not provide much information about the sign other than its form. For German Sign Language (DGS) there exist quite a number of different sign collections but not yet a corpus-based general dictionary.

1.2 The DGS Corpus Project

The DGS Corpus Project[1] is a long-term project of the Academy of Sciences and Humanities in Hamburg. The research is carried out at the Institute for German Sign Language and Communication of the Deaf (University of Hamburg). The project has started in 2009 and will continue until 2023. The project has two major goals: building a large general reference corpus of DGS and producing the first corpus-based general dictionary of DGS – German.

The DGS corpus is intended to be a linguistic and cultural resource for further research as well as the language community and other interested persons. It will also provide data for the first corpus-based dictionary DGS – German. For the corpus 330 signing informants from all over Germany were filmed in pairs at 12 different locations. The filmed material includes narrations of personal experiences, discussions, re-narrations, and other kinds of language use elicited in staged communicative events.[2] The raw data are comprised of 900 hours recording time with estimated 540 hours of signed activity containing estimated 3.5 million tokens of signs. A large portion of this data is being made accessible through segmentation, lemmatization, transcription, annotation and translation to become a searchable corpus. Corpus annotation work is carried out using iLex, an annotation environment that is linked to a lexical database (Hanke & Storz 2008). For lemmatization tokens are matched to types that are represented by unique glosses. The citation form of each sign type is described using the notation system HamNoSys (Hanke 2004). This is very time-consuming and cumbersome work, as none of these steps can be carried out in an automated way yet. Hence, the corpus will not have reached a relevant size for a number of years to go. This leads to the situation that at the present time DGS lexicographers have to resort to other sources and methods for dictionary compilation.

Within the DGS Corpus Project two electronic dictionaries of DGS will be produced in consecutive phases. First – while the corpus is still being built and the elicited data not yet fully accessible – a preliminary Basic Vocabulary with about 1500 entries is being produced. It is based on previously published sign collections and validated through feedback from the signing community. At a later stage the corpus data will be used as basis for the production of a general Dictionary of DGS – German with approximately 6000 sign entries to be published in 2023.

1.3 Sign Language as a Multi-channelled and Non-written Language

One challenge each sign dictionary project has to face is, that sign language is a visual-gestural language without written form. A sign stream involves signals produced by different parts of the upper body – two hands, mouth movements, facial expressions, head movements, posture, body shifts and eye-gaze. In the visual-gestural modality these articulatory features are easily produced and perceived simultaneously. Since the sign stream consists of several simultaneous but different signals it is diffi-

1 For more information on the project see: www.dgs-korpus.de.
2 For more detail see Nishio et al. (2010).

cult to devise a writing system that on the one hand can capture and represent the form and given complexity and on the other hand is simple enough to be written and read easily. In the research of spoken languages it is possible to adapt an existing writing system to a previously unwritten language – at least for a transition period – because the problem of how to write spoken languages has been solved. In sign language, however, there is no functioning writing system for any language that could simply be adapted to DGS for the purpose of dictionary making.[3] For sign language lexicography this absence of an operating writing system poses several challenges:

First, researchers cannot rely on already existing written texts to build a corpus of or to base their analyses on. Corpora have to work with data consisting of filmed signing. For corpus building such films are often recorded specifically for this purpose. Sign films recorded for other purposes are still limited to specific contexts (types of language use, genres). For most signers they are not part of everyday communication and can therefore cover only a small segment of sign language use. Second, for corpus-building and transcription purposes there is no fast and easy way of writing down signed text in a way that represents the form of the language units. More demanding and time consuming notation and transcription methods or indirect forms of annotations have to be used. This also makes it more difficult to analyse and automatically process corpus data. For instance searches on word forms have to be provided for by other means if there is no written representation of the forms available. Third, there is no direct and stable representation of form that enables the analyser to scan and review the sign and its context on a glance directly through their linguistic forms in the same way as a written text would do for example in a concordance line. Fourth, for dictionary production there is no practical way to use sign language as meta-language extensively. Film clips with signed information do not provide the same usability for scanning, browsing, searching and punctual access as does written information, since the content of a film is not permanently visible or machine-readable. Furthermore, films are very time-consuming in production and editing and therefore also for practical and financial reasons no real alternative to the use of written information.

Fifth, since there is no written form of the lemma sign, the question arises of what element to use as lemma. The lemma usually represents the form of the word, serves as an element for quick identification, sorting, searching and ordering and as the address for cross-references. All elements that could possibly be chosen (film, drawing, notation, gloss, number) do not fulfil all of these functions and have their drawbacks.[4] The best solution so far is to use a combination of one element to represent

3 There are some notation systems developed for academic purposes (e.g. HamNoSys, Stokoe-Notation). They are suited for detailed descriptions of sign forms. However, these systems do not handle facial expressions, body posture, eye gaze, mouthings or iconic locations well enough to completely rely on them in research. Also stretches of text are not easily written and read. Other systems, intended for general use such as SignWriting (Sutton et al. 2009), are not widely established and also have other drawbacks like not being machine-readable. For a further discussion of merits and shortcomings of the most common writing systems see König & Schmaling (2012).

4 For a more in-depth discussion of the advantages and disadvantages of these elements representing signs in dictionaries see Zwitserlood 2010: especially pp. 445-454 and Kristoffersen & Troelsgård, 2012a: 295-298.

the form of the lemma sign (e.g. drawing or film) and a second element for unique identification, sorting and address for cross-referencing (e.g. unique gloss or number).

2 The Basic Vocabulary of DGS

2.1 The Basic Vocabulary within the Context of the DGS Corpus Project

The Basic Vocabulary is intended to be a basic learners' vocabulary and to cover the most central signs of DGS and their core meanings. It is based on information about DGS signs contained in different sign collections and learning materials. The information is lemmatized, reviewed, edited and prepared for the signing community's feedback by which it will be validated and supplemented. The Basic Vocabulary is only a first and preliminary step in the lexicographic work within the project. It serves a number of purposes within the overall project design:

It is a means to round up and review information already published on DGS signs and put some of it to the test. This allows for a critical review and evaluation of previous lexicographic work with the chance of identifying and eliminating artefacts and to move towards a more complete and adequate description of signs and their meanings. With the Basic Vocabulary we can also develop and test lexicographic solutions to sign language specific problems that result from the visual modality as well as from the lack of a written form.

In addition, the publication of the Basic Vocabulary is a way to give something back to the signing community for their previous participation in the project, i.e. as informants in the corpus data collection, much earlier than 2023. The feedback process is also a way to involve the language community and let them further participate in this project on their language. Through the feedback they gain some influence and control over how their language is represented in the resulting dictionaries.

For the compilation of the Basic Vocabulary a set of question types to validate lexical items are developed, tested and used to collect data on signers' intuitions about signs and their meanings. Also a feedback system is specifically designed to conduct this survey via sign language. In the context of the Basic Vocabulary we can develop, test and improve both feedback tool and procedures. The gained experience and know-how can at a later stage be utilized for supplementing corpus data in the production of the Dictionary DGS – German.[5]

5 With estimated 2.5 million tokens the DGS corpus will still be relatively small in comparison to corpora of well-studied written languages. Complementary methods to gain additional insights on signs and their meanings and uses will be useful.

2.2 Properties of the Basic Vocabulary

The Basic Vocabulary is developed with a primarily monolingual oriented perspective: It focuses on the signs and aims at an adequate description of their variants and meanings. The meta-language of the dictionary is written German. Following this approach the dictionary will contain only one type of entries, that is, sign entries. Nearly all information included in the Basic Vocabulary will concern DGS signs while at the same time there is neither intention nor need to reproduce information on German words that can be obtained easily from existing dictionaries of German. Nevertheless, the actual product will exhibit some features of a bilingual dictionary: The signs' senses will be identified and expressed by German translation equivalents and disambiguating contexts. Furthermore there will be an alphabetical index of the German equivalents, making the sign entries accessible also via German (as source language) and thus providing also for a bilingual access.

The sign entries will be ordered by formational parameters of the signs.[6] However, since the Basic Vocabulary will be produced in an electronic form, the primary ordering of signs (macrostructure) is not crucial. The product will include a search function accessing the entries via sign form. For a search the user can select different combinations of manifestations of the sign form parameters as presented in a series of menus.[7]

2.3 Sign Collections: Sources for Lemma Selection and Spoken Language Equivalents

For the compilation of the Basic Vocabulary nine previously produced sign collections of DGS are taken as a basis for the selection and analysis of basic DGS signs.[8] The sign collections are used in two ways: for lemma selection and as a starting point for word sense disambiguation.[9]

6 Formational parameters are e.g. handshape, location, number of hands, movement types. There is no agreed upon order available and existing sign dictionaries have arrived at different solutions. For the current project the final order has not been decided yet.

7 This search function will be similar to the one implemented for the *Fachgebärdenlexikon Gesundheit und Pflege* ([5e]). To go to the sign form search function click on "Suche über Gebärdenform" at http://www.sign-lang.uni-hamburg.de/glex/.

8 Some of these collections consist of several publications that have been produced by the same team or institution and are follow ups of each other or a series of complementary works. For an overview see separate list in the reference section. The list of sign collections also include some teaching materials with vocabulary lists.

9 *Word sense* here in the case of sign language means "sign sense". In linguistics many technical terms exhibit a spoken language bias. Examples are: speaker, word, phoneme, phonology and oral. However, in a more general sense they can be used to describe the same abstract or corresponding phenomenon, category or role in both language modalities. Other technical terms with a spoken language bias are applied to sign language accordingly in this paper.

2.3.1 Lemma Selection

The Basic Vocabulary aims at including signs that are most basic with regard to everyday communication needs and that are frequently used and widely known. With frequency lists still unavailable for DGS the second best way to approximate this aim was to look for signs that had been listed in several previously published sign collections. Following this, we based the lemma selection process primarily on a comparison of these signs.

Without a written representation of the sign forms in the products the only way to identify and compare the signs was to look at each visual representation (photos with arrows or films) of the given sign, sign combination or sign sentence, break sign combinations and sentences down into single signs (tokenisation) and lemmatize them consistently. This was done in our working environment iLex following the annotation rules of the DGS Corpus Project. The nine sign collections contained about 40.800 tokens (single signs) that have been processed. Not counting uses of the manual alphabet, pointing signs, number signs and productive signs about 9350 different sign forms of 5440 basic conventional signs haven been documented. All basic signs – about 1050 – that were found in at least five of the nine sign collections have been selected for further analysis and inclusion into the Basic Vocabulary.[10]

2.3.2 Spoken Language Equivalents

All of the nine sign collections contain to some extent bilingual information matching signs and words as equivalents. They either have a primary make-up using German as source language showing signed equivalents (e.g. [6]) or they have sign entries and use German equivalents and contexts to list the signs' meaning(s) (e.g. [5c-5f]). The vocabulary lists of learning materials (e.g. [9a, 9b]) and the collection of phrases ([2c]) use glosses as written labels for signs where the gloss word is to be taken as an indication of the meaning of the sign.[11] In the children's visual dictionary ([8a, 8b, 8c]) pictures of objects and sign films are presented without any written words. However, the sign film contains an audio track where the corresponding German word is spoken aloud.[12] All available equivalents and meaning indications assigned to one sign form in different sign collections are taken as a starting point for identifying word senses for signs that are to be included in the Basic Dictionary.

In the Basic Vocabulary we aim at the most central meanings of selected signs. Not all of the equivalents listed in the sign collections are equally basic or central, and it is sensible and necessary to make a selection. This selection is operationalized in the following way: German equivalents are compared

10 Additional signs will be included in order to make sure all concepts of a specific learner's basic word list of German (Glaboniat et al. 2005) and some Deaf-specific concepts are covered.

11 In corpus annotation and lexicography glosses are used as a unique identifiers for signs: one gloss always stands for the same form and different forms receive different glosses. In learning materials glosses are often used as a means to represent the sign order of example sentences and at the same time to indicate the signs' meanings in this particular context. Depending on the context a sign might therefore be represented by different glosses and the same gloss might be used for different signs.

12 In the print version of the product German equivalents of signs are printed below the signs' drawings (comp. Kestner 2002).

to a combined list of German words that are considered relevant for a basic (German) vocabulary. This combined list is basically a merge of a foreign language learner's basic vocabulary list of 1071 words (adapted from Glaboniat et al. 2005, levels A1 and A2 (active vocabulary)) and a frequency list of the about 4000 most frequent German words (Jones & Tschirner 2006, 2011). Words that are both listed as equivalents in the sign collections and found in the combined German word list are selected for further processing.

2.3.3 Reversal of bilingual information

After the basic German equivalents of a sign have been determined, the next step is to identify individual senses that the words and the sign potentially share. This step can be understood in terms of the reversal of a mono-directional bilingual dictionary. It goes beyond the reversal on the form level and is aimed at the level of meaning, i.e. the level of lexical units (as outlined by Martin 2013: 1447-1448). However, in our case the equivalents do not stem from one uniform source but from different sources and are of different status, quality and granularity as far as meaning specification is concerned.

It is reasonable to assume that any German word listed as translational equivalent must share at least one sense with the corresponding sign. If the German word is polysemous each of its senses has to be considered separately, whether it could also be a potential sense of the sign. At this stage it is inevitable to do a considerable amount of sense splitting. For this we use German dictionaries. The approach we take is to identify relevant senses of the German word according to two criteria: First, the senses chosen are to be the most basic and central ones with regard to everyday communication. More marginal senses are omitted from further processing. Second, the senses selected for the feedback have to be certain or likely candidates for being senses of the sign as well. Whenever the context information provided in a sign collection clearly identifies a particular sense this information is taken into account and the identified sense goes into the feedback process for verification. The same is true for senses that are attested clearly in the data from previous projects or already transcribed corpus data in our database. In some cases it is very straightforward that a particular word sense is clearly not a sign sense and can be omitted, e.g. because it contradicts with the iconic properties of the sign. In many other cases a decision cannot be reached without further information of the sign's use in context. Uncertain basic senses are prepared for the feedback process to be verified or rejected by the language community (see 4.).

3 Sign Language Specific Issues concerning Word Sense Disambiguation

In the process of identifying and diambiguating word senses of signs two sign language specific properties have to be taken into consideration: iconicity and mouthings.

3.1 Iconicity

Many signs are iconic. The form of an iconic sign resembles something – the underlying image – which is directly or indirectly associated with one or several of the sign's meanings. The underlying image can influence the ways a sign's meaning may be restricted or open for extension. Therefore, when checking possible senses of a sign special attention has to be given to its underlying image.

3.1.1 Underlying Image and Meaning

The underlying image of a sign may restrict its use to certain specific senses and prohibit its use for related senses. Due to iconic aspects the granularity of meaning differentiation of a sign may be much finer than that of the corresponding word of the surrounding spoken language. When the underlying image of an iconic sign is very specific and the sign's meaning is restricted accordingly, related senses that are not compatible with the iconic properties of the sign in question are often covered by other signs.

Consider for example different DGS signs for *Pfeife*. The German word *Pfeife* can mean either "whistle (object)" or "smoking pipe (object)" while DGS has different signs for these concepts PIPE[13], WHISTLE1 and WHISTLE2 (see figure 1). The sign form of PIPE resembles a person holding a pipe at its bowl some small distance away from the mouth so that the end of its imagined shank touches the lips for smoking. In WHISTLE1 the fingers indicate a whistle (as used by referees) located with its mouthpiece at the mouth. WHISTLE2 shows how a person is holding a whistle to the mouth. The signs' meanings are restricted to the different objects in accordance with their underlying image.

13 Since there is no conventional writing system for sign languages it is a convention in sign language linguistics to use capital letter glosses with numberings to identify and represent different signs. The spoken language word used as gloss simply is an easily written and readable label or name for the a sign, but neither is the gloss' part of speech nor are its meanings to be interpreted as indicating the sign's part of speech or meanings.

Figure 1: DGS signs PIPE, WHISTLE1 (Source: [5e]), WHISTLE2 (Source: [6]), AND WHISTLE3 (Source: [5e]).

WHISTLE1 and WHISTLE2 can also mean "producing a whistling sound, whistling". Another DGS sign listed for this meaning is WHISTLE3 (see figure1). On the iconic level the three signs for *whistling* are specific as to how the whistling sound is produced: WHISTLE3 shows how thumb and index finger are rounded to an O and put to the mouth producing the sound without the help of a whistle, while for WHISTLE1 and WHISTLE2 a whistle is involved. The question arises whether these iconic differences are reflected in sign usage or not. In a specific context the sign used will usually be the one that iconically best matches the named or described incident.[14] Thus the "whistling" sense of the signs could be further specified as "producing a whistling sound by blowing into a whistle" for WHISTLE1 and WHISTLE2 and "producing a whistling sound by using your fingers in the indicated way" for WHISTLE3. Sign language users may find this trivial or logical or even unnecessary information since the form of the sign shows these aspects of meaning, but this only proves the point that iconicity plays a role in sign language use. Whenever the selection and use of a particular iconic sign takes its underlying image into account this kind of information should be made explicit in a dictionary and thus be accessible to the language learner. This can be done by listing the specific meaning as determined by the iconic value of the sign as one of the sign's senses.

However, iconic signs do not always exploit their more specific underlying images for subtle meaning differences. Often sign forms are conventionalized for a general meaning without paying attention to the more specific images their forms are derived from. One example is the sign TELEVISION (see figure 2). Its underlying image shows a person turning two knobs (the kind old television sets had for adjusting the frequency). This sign is used for all kind of television sets regardless of their technical type or outward appearance.

14 This is a hypothesis based on informal observations and signers' intuition. To what extend this can be confirmed by corpus data remains an open question.

Figure 2: TELEVISION and old television set with turning knobs.

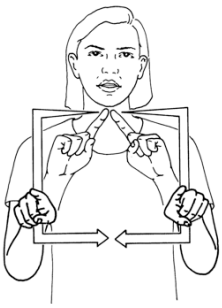


Figure 3: RECTANGLE.

Signs – just like words of spoken languages – can extend their meaning from the specific to the more general. With this we come back to the whistling examples: When the particular circumstances of sound production are not known to the signer or are not relevant or if one talks about whistling in general, is it then possible that any of the three signs may be used for the more general concept of "whistling" independent of how the sound is produced? And what about the whistling of a boiling water kettle or of an old steam engine train? Can the three signs for "whistling" also be used in a transferred sense for these kinds of whistling or are there other signs to be used instead? Do WHISTLE1, WHISTLE2 and WHISTLE3 additionally to their very specific sense matching their underlying image have a more general meaning "producing a whistling sound" (no matter how it is produced)? These are questions to be asked and hopefully answered in the feedback and – at a later stage of the project – by analysis of corpus data.

Whenever a sign's form resembles one particular example of a more general category its meaning can either be restricted to that specific kind resembled, or the sign can be used for the more general concept. There is no general rule as to what applies to a particular sign. For each sign this has to be determined separately. Thus, for iconic signs special attention has to be given underlying image when analysing their meanings.

3.1.2 Underlying Image as a Source for Polysemy

The iconicity of signs is an important source for polysemy and meaning extensions in DGS. This is especially true when the sign's form represents a very unspecified image that may stand for a number of objects, situations or actions. The underlying image can serve as a common core for rather diverse meanings which either share a common visual trait in their real-world manifestations or which utilize the visual metaphor provided. For example, among the translation equivalents listed in various sign collections as meanings for the sign RECTANGLE (see figure 3) quite a number refer to objects or convey meanings that either consist of or are associated with a piece of paper, such as: "paper", "piece of paper", "slip (paper)", "page", "form (to be filled in)", "map", "recipe (cooking)", "prescription", "sick note", "official notification", "certificate", "report card". Other meanings refer to rectangular and predominantly two-dimensional objects that are solid such as "window", "signboard", "mirror", "screen (computer, TV)", or that are soft such as "pillow", "towel", and "(cleaning) rag". On the semantic level a

"pillow" does not have much in common with a "computer screen" neither does an "official notification" have anything to with a "cleaning rag" – the characteristic all these share is their rectangular shape or their association with something of a rectangular shape. In cases like this the iconic potential depicted in the sign's form serves as the basis for meaning extension of the sign.[15]

This kind of extensive polysemy in DGS is facilitated by an interaction of iconicity and mouthings. The mouthing helps to contextualize and disambiguate the polysemous sign (see below 3.2.).[16]

In DGS there are many ways of how the meaning of a sign can be extended. Often the meaning is extended along the lines of regular polysemy. Bentele, Konrad & Langer (2000: 621) list some conventional and productive uses of the sign STIRRING that exemplify some of these patterns (see Table 1). The form of STIRRING (see figure 4) depicts someone holding a tool, such as a cooking spoon, and stirring something with it.

It is very common in DGS that in addition to a sign's core meaning (an action represented by the sign form as in STIRRING) it is also used in many other related senses – for example – for a profession typically associated with the activity, for a place or institution where the activity takes place, and for an object produced or manipulated by the activity.

Figure 4: STIRRING, WRITING, and KNEADING.

Figure 5: BEAK.

15 The meanings listed here are conventional uses of the sign RECTANGLE as they appear in sign collections. The list here is not complete. These conventional meanings are strongly and regularly associated with this sign. However, in addition the sign may also spontaneously be used to denote other rectangular two-dimensional objects in a specific context, often in combination with a corresponding mouthing. This is a common strategy to be used, especially when there is no other conventional sign or when the signer doesn't know such a sign. Such more spontaneous uses are backed by the sign's iconic potential and will be easily understood in context. In contrast to the more *conventional uses* these have been called *productive uses* of a sign (cf. König, Konrad & Langer 2008).

16 Specific mouth gestures (mouth movements not derived from spoken language words) and facial expressions can also contribute to the disambiguation of meanings. These are not discussed here.

Other examples of this type are WRITING ("writing", "author, writer", "school" and "literature"), and KNEADING ("kneading", "baker", "bakery", "dough"). Our impression is that in addition to well-known patterns of regular polysemy found in spoken languages (e.g. Atkins & Rundell 2008: 140) some DGS patterns may be specific to a visual language.[17]

German equivalents	English equivalents	meaning extension pattern
rühren, umrühren	to stir	core meaning: action directly represented by sign's form
kochen	cook (prepare food)	more comprehensive activity typically involving the represented action
Küche	kitchen	place or area where activity usually takes place
Koch	cook, chef	person who usually or professionally carries out the activity
(Koch-)Löffel	(cooking) spoon	tool used in the represented action (productive use)
Eintopf	stew	object or substance manipulated by action (productive use)
Hauswirtschaft	home economics	superordinate domain the activity belongs to (productive use)

Table 1: Some meaning extensions for STIRRING (adapted from Bentele, Konrad & Langer 2000: 621).

The conclusion that we draw from these observations is that the underlying image and image producing technique of an iconic sign can be and should be considered for decisions on lemmatisation and in the analysis of word senses. For us a common underlying image connecting diverse meanings is a valid reason to treat these meanings as different senses of a polysemous sign in one entry rather than treating them as homonyms in different entries (cf. also König, Konrad & Langer 2008).

3.2 Mouthings

Many manual signs are accompanied more or less regularly by mouthings. Mouthings are mouth movements that remind of the articulation of spoken language words. Some of these mouthings are quite conventionally attached to certain signs, while others appear to be more occasional, dynamic or even spontaneously combined (cf. Ebbinghaus & Heßmann 1996, 2001). While the linguistic status of these mouthings is not yet finally agreed upon in sign language linguistics, we consider mouthings to be some kind of lexical material from German integrated into DGS signing.

Most signers are to some degree bilingual and have at least some basic knowledge of German words and their core meanings. Mouthings in DGS originate from German words and bear the potential of

17 Further research might reveal that certain kinds of regular polysemy may be typical for or restricted to certain iconic sign types – whether the underlying image is a representation of an action (manipulative technique as in STIRRING), the outline of the form of an object (sketching technique as in VIERECK) or representing an object and its movement (substitutive technique as in BEAK see figure 5 above). For a description of image producing techniques in DGS cf. Langer (2005) and König, Konrad & Langer (2008: 388-389).

carrying and utilizing the semantic load of the mouthed word. This both supports and at the same time complicates the process of determining the various senses of a sign.

Mouthings can help to disambiguate polysemous signs. Often the mouthing is an important factor to determine the meaning of a polysemous sign in a specific context. It is not unusual to have one sign (for example BEAK see figure 5) to cover a basic category concept (such as "bird") and a set of subordinate concepts (such as "sparrow", "blackbird", "pigeon", "chicken" and a number of other birds with small narrow beaks) in everyday non-expert signing. One sign form can also cover complementary concepts such as "brother" and "sister" or antonymic concepts such as "dry" and "moist". The occurrences of such polysemous signs are usually accompanied by mouthings that have a disambiguating function and correspond to the meaning.

When a sign is conventionally associated with a particular mouthed word not all of the word's meanings automatically are also conventional meanings of that particular sign. Some of a word's meanings might not match the iconic value of the sign. However, in a specific context the sign form might just be used to contextualize that word, occasionally even for a meaning that is not lexicalized for the sign. Also, a polysemous German word might contribute some of its other or more marginal senses into sign utterances by way of mouthing and finally extend the sign's meaning in a way that might not be expected or not be in accordance with the sign's iconic core but parallels the meanings of the German word. One example for this is the sign CRAB. The underlying image of the sign shows the opening and closing of a pair of crab pincers. The German name for this animal is *Krebs*. But *Krebs* has also the meanings "cancer (disease)" and "Cancer (zodiac sign)". This is paralleled in DGS in that the DGS sign CRAB is also used in all three senses of the German word.

The interplay of a sign's iconic potential and its productive exploitation, the sign's lexicalized (conventional) meanings and the meanings of associated mouthed words make it difficult to determine the degree of conventionalization for a particular meaning. One major issue here is that a frequently used mouthing has the potential to bring additional meanings from another language (German) into the equation. This meaning potential of the word has to be considered especially when dealing with rather isolated signs as listed in sign collections and dictionaries that use spoken language equivalents to indicate the sign's meaning(s). Analyses of usage data from a large corpus can provide evidence for which senses are common usage and which are to be considered marginal or even occasional or creative occurrences. Until corpus data analyses of this kind will be possible we resort to the data as presented in the sign collections, documented in our internal database from previous projects and signer's intuition to identify and disambiguate possible senses to be validated in the feedback process.

4 Feedback from the Signing Community

In order to verify and complement our data, e.g. for regional variation and passive vocabulary, we rely on feedback from the signing community. This holds especially for the signs to be included in the Basic Vocabulary.

4.1 Preparation of Content

Each potential combination of sign form and meaning identified in the process described above has to be verified by the community. The crucial point is how to convey the intended meanings. Ideally context information in DGS would be given for each instance, however such an approach is cumbersome and not feasible within the project. Furthermore, in several cases senses seem clear enough from a short disambiguating written context such as "mouse (computer)" vs. "mouse (animal)". In cases like this we can spare the participants the effort of having to watch an additional film. For the feedback we chose a rather pragmatic approach: In order to make it as easy for the participants as possible, we present German equivalents of the signs, often along with German disambiguating contexts in brackets, and add signed contexts wherever we feel it is needed. German contexts can be a superordinate term, a domain, a relevant dimension (e.g. space, time), further disambiguating synonyms, a short explanation, collocations or other suitable context information. For DGS contexts we mainly use examples, but we also try short explanations, collocations and sign synonyms. For this we rely largely on the linguistic knowledge and intuition of a Deaf colleague. To a smaller extent, as far as it has been already made accessible through transcription, we also use data from our corpus as basis for the construction of examples.

4.2 The Feedback System

For the realisation of the feedback we decided for a crowd sourcing approach using an online feedback platform. Our aim is to get as many members of the signing community involved as possible, i.e. not only Deaf people but everybody using DGS.[18] Dealing with a relatively small language community spread across the country, the crucial point is not only how to attract enough participants, but also how to make them check back regularly. The newly developed online platform, running on desktop computers as well as mobile devices, allows for as much freedom for the participants as possible.[19] While the target community's pride of their own language and interest in supporting the project can be expected to be the main motivation for taking part in the feedback, computer game elements such

18 Metadata information will be used to weigh the participants' answers depending on their proficiency in DGS.

19 A more detailed description of the Feedback System in DGS and German can be found at http://feedback.dgs-korpus.de.

as levels and high score are used as extrinsic motivation. An important point for developing the system was to make it accessible by presenting all information in DGS and also to allow for answers and comments to be given in DGS (video upload via webcam). The system is optimised for our use, but the software is open source and can be adjusted for other purposes.

Participants create individual accounts where their answers and metadata are stored. Within the feedback system tasks are organised in categories and work packages. Due to the complexity of the tasks, users are asked to deal with only one type of task at a time. Categories therefore contain work packages of generally one type of task, and a new category is only released after a certain amount of work packages of the previous category have been completed. A so-called "golden trail" leads through the work packages. A help page contains explanations dealing with different aspects of the feedback system and with problems and questions that might arise. A "comment"-button on each page allows the participants to leave comments (in DGS or written German) that provide us with extra information or help us to spot problems regarding the tasks and the system.

4.3 Presentation of Content in the Feedback System

Work packages are organised in pages and rows. Task explanations and questions are all presented in DGS and written German (a button allows to switch between both languages), other content is provided as film, text or pictures depending on the purpose.

Figure 6: Example page of a work package "Form and Meaning"[20].

20 In this example the sign is made by two hands with "V" handshape moving up and apart. The meanings listed for verification are: "famous", "celebrity", and "public" (rows are moving upwards when answer button has been clicked).

Our first question type deals with sign forms and meanings as described above. Each sign form is first shown without mouth pattern in order to have the participants concentrate on the hands.[21] Participants are asked whether they know the sign form presented. Clicking "no" leads directly to the next page with a different sign form. When pressing "yes" further questions are revealed concerning the meanings of the respective sign form. In each row one meaning of the sign is presented by a film clip of the sign with a common mouth pattern and one or more German equivalents (often including sense indicating contexts in brackets). If needed an additional film clip provides context information in DGS also. Participants are asked whether they use the sign with the respective meaning themselves, don't use it but have others seen using it, or don't know this sign with this meaning at all. At the end of each page the system allows to add further meanings in written German or DGS (via webcam).
The completion time for each work package is planned to be approximately 20 minutes. On average a work package contains about eight different sign forms depending on the number of meanings to be verified (so far up to a maximum of 12 meanings with one sign form). When a work package is submitted the participant receives the number of points as defined for this package (i.e. independent of the number of "positive" or "negative" answers given) and is listed in the high score table. While this game-like approach primarily aims at motivating participants and boosting competition between them it also serves as a tool to stir which tasks are assigned to a participant (i.e. a certain number of points is needed to access the next category with different types of questions).

5 Conclusion

While the corpus collected within the project is still in the process of transcription and annotation we rely on other sources for our lexicographic work, especially for the production of our Basic Vocabulary. We have compared the signs of nine other previously produced sources (sign collections, learning materials) and based our lemma selection on the highest overlap in these products. The equivalents listed are taken as a starting point for a process of identifying different senses of the signs in a process of reversal of mono-directional bilingual information. In this process the iconicity of signs is also considered. The given meanings of signs are split into a finer and more sufficient granularity of senses, disambiguated through suitable contexts and presented via an online feedback system to the language community for verification or rejection.
The described procedure is not ideal to identify word senses of signs. It does things a little backwards, starting with the meanings of the German words given as equivalents instead of looking at signs and their contexts of actual use. However, this procedure takes the already published material containing

21 Considering mouthings as an essential part of signing raises the issue of how to present sign forms in the feedback as signers should not rate sign-mouthing but sign-meaning combinations. Our solution so far is to omit any mouth pattern for the more general question on sign form and show a separate film of the sign with the presumable most commonly used mouth pattern for each meaning. An accompanying explanation is included in the tasks explanation and pre-test have revealed no major problems.

information on signs' meanings seriously. Also, involving the language community in the feedback process has some advantages: While the corpus contains signing from only 330 informants we expect that through the feedback we can obtain data from even considerably more signers. This way we can obtain a more detailed account on regional variation concerning signs and sign senses. Direct questions also allow us to elicit data on passive as well as active vocabulary.

Nevertheless the result of such a feedback process can only be a first incomplete and approximated account of a sign's senses. This is the reason why we regard the Basic Vocabulary as preliminary. Once the corpus is available it will be the basis for further lexicographic work. Results from the feedback can be confirmed and complemented by corpus data, while the feedback process (crowd sourcing) may prove a valuable complementary method to obtain data for signs and questions the corpus does not cover.

6 References

Atkins, B.T.S, Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford Univ. Pr.

Bentele, S., Konrad, R. & Langer, G. (2000). Transkription und Analyse. In R. Konrad, T. Hanke, A. Schwarz, S. Prillwitz & S. Bentele. Fachgebärdenlexikon Hauswirtschaft. Vol. 2. Hamburg: Signum, pp. 617-632.

Brien, D. (ed.) (1992). *Dictionary of British Sign Language/English*. London: Faber and Faber.

Ebbinghaus, H., Heßmann, J. (1996). Signs and words. Accounting for spoken language elements in German Sign Language. In W. H. Edmondson, R.B. Wilbur (eds.) International Review of Sign Linguistics 1. Mahwah, N. J.: Erlbaum, pp. 23-56.

Ebbinghaus, H., Heßmann, J. (2001). Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different things. In P. Boyes Braem, R. Sutton-Spence (eds.) *The hands are the head of the mouth: The mouth as articulator in sign language*. Hamburg: Signum, pp. 133-151.

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile deutsch. Gemeinsamer europäischer Referenzrahmen*. Berlin, München: Langenscheidt. (Includes CD-ROM).

Hanke, T. (2004). HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In O. Streiter, C. Vettori (eds). *LREC 2004, Workshop proceedings: Representation and processing of sign languages.* Paris: ELRA, 2004, pp. 1-6.

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: ELRA, 2008, pp. 64-67.

Johnston, T. (1989). AUSLAN Dictionary. A dictionary of the sign language of the Australian deaf community. Victoria: Aust. Print Group.

Jones, R.L., Tschirner, E. (2006). Frequency Dictionary of German. Core Vocabulary for Learners. London: Routledge.

Jones, R.L., Tschirner, E. (2011). *Frequency Dictionary of German. Core Vocabulary for Learners.* (CD-ROM). London: Routledge.

Kennedy, G. (ed.) (2008). *A Concise Dictionary of New Zealand Sign Language.* Reprint. Wellington: Bridget Williams Books.

Kestner, K. (2002): Tommys Gebärdenwelt. Das Gebärdensprachbuch zur CD-ROM. Guxhagen: Manual Audio Devices.

König, S., Konrad, R., Langer, G. (2008). What's in a sign? Theoretical lessons from practical sign language lexicography. In J. Quer (ed.) *Signs of the Time. Selected Papers from TISLR 2004.* Hamburg: Signum, pp. 379-404.

König, S., Schmaling, C. (2012). Gebärdenschriften: Flüchtiges fixieren. In H. Eichmann, M. Hansen, J. Heßmann (eds.) *Handbuch Deutsche Gebärdensprache. Sprachwissenschaftliche und anwendungsbezogene Perspektiven.* Hamburg: Signum, pp. 341-356.

Kristoffersen, J.H., Troelsgård, T. (2010). Compiling a sign language dictionary. Some of the problems faced by the sign language lexicographer. In M. Mertzani (ed.) *Sign Language. Teaching and Learning. Papers from the 1st Symposium in Applied Sign Linguistics, Centre for Deaf Studies, University of Bristol, 24-26 September 2009.* Bristol: Centre for Deaf Studies, pp. 1-10.

Kristoffersen, J.H., Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford Univ. Pr., pp. 293-315.

Langer, G. (2005). Bilderzeugungstechniken in der Deutschen Gebärdensprache. In *Das Zeichen* 70, pp. 254-270.

Martin, W. (2013). Reversal of Bilingual Dictionaries. In R.H. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. (Vol. 5,4 of Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)). Berlin: De Gruyter, pp. 1445-1455.

Nishio, R., Hong, S., König, S., Konrad, R., Langer, G., Hanke, T., Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) Corpus Project. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, A. Schembri (eds.) *Corpora and Sign Language Technologies. 4th Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA, 2010, pp. 178-185.

Ordbog over Dansk Tegnsprog. Accessed at: http://www.tegnsprog.dk/ [20/03/2014].

Stokoe, W.C. (1993). Dictionary Making, Then and Now. In *Sign Language Studies*, 79 (22), pp. 127-146.

Stokoe, W.C., Casterline, D.C. & Croneberg, C.G. (1965). *A dictionary of American Sign Language on linguistic principles*. Washington, DC: Gallaudet College Press.

Sutton, V.J., Paul, F.A., Candelaria, I., Gunderson, J. (2009). *SignWriting Basics Instruction Manual*. SignWriting Pr.

Zwitserlood, I. (2010). Sign Language Lexicography in the Early 21st Century and a Recently Published Dictionary of Sign Language of the Netherlands. In *International Journal of Lexicography*, 23 (4), pp. 443-276.

Sign Collections for DGS

Sign collections for the Basic Vocabulary. Related or complementary works are grouped together.

[1] *777 Gebärden 1-3. Alle 3 Folgen auf einer DVD.* (2002). Version 2.0. Guxhagen: Manual Audio Devices. (DVD-ROM).

[2a] *DGS-Aufbau-Lexikon*. (1998). Aachen: Microbooks/Desire. (CD-ROM).

[2b] *DGS-Basis-Lexikon.* (1998). Aachen: Microbooks/Desire. (CD-ROM).

[2c] *DGS-Phrasensammlung.* (1998). Aachen: Microbooks/Desire. (CD-ROM).

[3a] Maisch, G., Wisch, F.-H. (1987). *Gebärdenlexikon. Band 1. Grundgebärden*. Hamburg: Verlag hörgeschädigte kinder.

[3b] Maisch, G., Wisch, F.-H. (1988). *Gebärdenlexikon. Band 2. Mensch*. Hamburg: Verlag hörgeschädigte kinder.

[3c] *Grundgebärden 1. Für Einsteiger.* (1999). Hamburg: Verlag hörgeschädigte kinder. (CD-ROM).

[3d] *Grundgebärden 2.* (2000). Hamburg: Verlag hörgeschädigte kinder. (CD-ROM).

[4a] Metzger, C., Schulmeister, R. & Zienert, H. (2000). *Die Firma. Deutsche Gebärdensprache do it yourself.* Hamburg: Signum. (CD-ROM).
[4b] Metzger, C., Schulmeister, R. & Zienert, H. (2003). Die Firma 2. Deutsche Gebärdensprache interaktiv. Aufbaukurs in Deutscher Gebärdensprache - Schwerpunkt Raumnutzung. Hamburg: Signum. (CD-ROM).
[5a] Arbeitsgruppe Fachgebärdenlexika (ed.). (1994). *Fachgebärdenlexikon Computer*. 2 Vol. Hamburg: Signum.
[5b] Arbeitsgruppe Fachgebärdenlexika (ed.). (1996). *Fachgebärdenlexikon Psychologie*. 2 Vol. Hamburg: Signum. Online-Version can be accessed at: http://www.sign-lang.uni-hamburg.de/plex/ [20/03/2014].
[5c] Konrad, R., Hanke, T., Schwarz, A., Prillwitz, S. & Bentele, S. (2000). *Fachgebärdenlexikon Hauswirtschaft*. 2 Vol. Hamburg: Signum. Online-Version can be accessed at: http://www.sign-lang.uni-hamburg.de/hlex/ [20/03/2014].
[5d] Konrad, R., Schwarz, A., König, S., Langer, G., Hanke, T. & Prillwitz, S. (2003). *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Hamburg: Signum. Online-Version can be accessed at: http://www.sign-lang.uni-hamburg.de/slex/ [20/03/2014].
[5e] Konrad, R., Langer, G., König, S., Schwarz, A., Hanke, T. & Prillwitz, S. (eds.) (2007). *Fachgebärdenlexikon Gesundheit und Pflege*. 2 Vol. Seedorf: Signum. Online-Version can be accessed at: http://www.sign-lang.uni-hamburg.de/glex/ [20/03/2014].
[5f] Konrad, R., Langer, G., König, S., Hanke, T. & Rathmann, C. (eds.). (2010). *Fachgebärdenlexikon Gärtnerei und Landschaftsbau*. 2 Vol. Seedorf: Signum. Online-Version can be accessed at: http://www.sign-lang.uni-hamburg.de/galex/ [20/03/2014].
[6] Kestner, K., Hollmann, T. (2009). Das große Wörterbuch der deutschen Gebärdensprache. Der erste umfassende Gebrauchswortschatz in DGS als elektronisches Wörterbuch. 18000 Begriffe in Schrift und Video von A wie Aachen bis Z wie Zypresse. Deutsch – DGS, DGS – Deutsch. Bundeselternverband Gehörloser Kinder (ed.). Guxhagen: Kestner. [DVD-ROM with booklet.]
[7] Arbeitsgruppe ProViL. (2006). *DGS 3 und DGS 4. Aufbaukurse in Deutscher Gebärdensprache*. [Located at the eLearning platform WebCT/Blackboard; for members of the University of Hamburg accessible via OLAT - Online Learning And Training. Accessed at: *www.olat.uni-hamburg.de* [20/03/2014].
[8a] *Tommys Gebärdenwelt*. Version 3.0. (2007). Guxhagen: Kestner. CD-ROM.
[8b] *Tommys Gebärdenwelt*. 2. Tommy und Tina. Version 3.0. (2008). Guxhagen: Kestner. CD-ROM.
[8c] *Tommys Gebärdenwelt*. 3. Mit Tommys erstem Lexikon. Version 3.0. (2009). Guxhagen: Kestner. DVD-ROM.
[9a] Keller, J., Zienert, H. (2000). *Grundkurs Deutsche Gebärdensprache. Stufe I. Vokabel CD-ROM*. Hamburg: Signum. (CD-ROM).
[9b] Keller, J., Zienert, H. (2002). *Grundkurs Deutsche Gebärdensprache. Stufe II. Vokabel-Video*. Hamburg: Signum. (VHS-Video).

Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

Dearcadh na nDéise – Representations of Gaeltacht na nDéise in Dineen's Bilingual Irish-English Dictionary (1927)

Chris Mulhall, Seamus ó'Diollúin
Waterford Institute of Technology
cmulhall@wit.ie, sodiolluin@wit.ie

Abstract

Dictionaries are a written manifestation of the world, but the selection of words is a reflection of society at a particular time from one or more viewpoints. Irish dictionaries are largely unchartered territories in terms of lexicographic research and contain important regional and national narratives that reflect the ever-changing Irish cultural landscape in the late nineteenth and early twentieth century. The creation of the independent Irish state in the early part of the twentieth century and increasing demand for bilingual Irish-English dictionaries at this time heralded a beginning of a new area of activity in European lexicography. One of landmark editions from this period is *An Foclóir Gaedhilge agus Béarla* by Patrick S. Dineen. The aim of this paper is to explore a regional variety of Irish, Gaeltacht na nDéise, found in the South-East of Ireland and examine some of the socio-cultural narratives behind Dineen's choice of entries to depict this regional dialect. Based on a recategorisation of the empirical work of Nyhan (2006), a brief discussion of three prevalent socio-cultural themes, namely; anger, anxiety and poverty, recurring in some Gaeltacht na nDéise-specific entries found in Dineen's work is presented. These offer an interesting window into how the lexis of this particular region was represented and potentially contributed to its regional identity in a national context.
Keywords: Bilingual Irish-English Dictionaries; Socio-cultural narrative; Regional Dialect

1 An Foclóir Gaedhilge agus Béarla and Gaeltacht na nDéise

An *Foclóir Gaedhilge agus Béarla* by Reverend Patrick S. Dineen stands as the most iconic works in the comparatively short history of bilingual Irish-English lexicography. The early editions of this pioneering work came at a time of upheaval as Dineen sought to encapsulate the regional and national linguistic identities of the newly-liberated Irish State. A substantive part of Dineen's narrative of the Irish language engages in what Hartmann (2001) refers to as a 'personal-biographical' account of the

language[1], but this was tempered with a widely representative account of regional dialects and a conservative acceptance of borrowings. In the preface, Dineen stresses his efforts to flavour his dictionary with as many dialectal words as possible, most notably from Connaught and Aran, North and Mid-Ulster, West and North Kerry, South and South-West Cork, Meath, Omeath, Clare and West Limerick and the Comeragh District of Waterford. The regional specific lexis offers an interesting panorama into the linguistic sub-cultures in Ireland, but is also productive in highlighting particular values, identities and struggles pertinent to those areas at this point in Irish history. In the preface of *An Foclóir Gaedhilge agus Béarla* Dineen lists Fr. Mícheál McGrath and Riobaird Bheldon (1838-1914) as one of his lexicographical sources[2]. While not mentioned as a source in the first edition (1904) of the *Foclóir Gaedhilge agus Béarla* the 1927 edition contain 96 entries from this 'Cm.' source. The absence in the 1904 edition may be explained by the fact that Bheldon's poetry was only first published in December, 1903 - just months before the first edition of the dictionary was published. Dineen himself admits that the 1927 edition was "practically a new work" and "an effort has been made to secure much representative provincial Irish in word and phrase as possible" (Dineen, 1927: vii, xi). Bheldon and McGrath as immersed as they were in the language, literature and culture of the Déise Gaeltacht region of Waterford, and its surrounding hinterland, proved suitable sources of the living Irish language in south east Munster.
The words in *An Foclóir Gaedhilge agus Béarla* denoted by "Cm." are examples of the living language as spoken in rural county Waterford at the turn of the twentieth century. While the actual provenance of this collection of words might not be certain we are left in no doubt has to Dineen's reasons for including them in his *Foclóir Gaedhilge agus Béarla*:

> an effort has been made to net the chief living elements of the language while there was still time. The materials for the work have been drawn from the living language of Irish-speaking Ireland as well as from the written remains of the modern literature (Dineen, 1927: vii)

2 The Work of Riobaird Bheldon

Bheldon was a relatively little known poet until the publication of *Riobard Bheldon, Amhráin agus Dánta* (1995) by Pádraig Ó Macháin[3]. Ó Macháin (1995:125) mentions that Bheldon was an acquaintance of Fr.

1 The first edition was, to a considerable extent, compiled from memory. A large percentage of the illustrative phrases were taken from living expressions, conversations, etc, stored up in my childhood's memory; so that many of them have the vividness and directness characteristic of the spoken word. In this edition still further use has been made of this source; and the meanings and applications of key words, somatic terms and other important expressions have been considerably expanded (An Foclóir Gaedhilge agus Béarla [Second Edition], 1927:vii).

2 Cm.- a list of words from the Comeragh district of Waterford compiled by the late Fr. McGrath (P.P. of Ring) and the late Riobárd Ueldon, the poet (An Foclóir Gaedhilge agus Béarla [Second Edition], 1927:xxiv).

3 Pádraig Ó Dálaigh published *Riobard Bheldon: File an Chomaraigh* in 1925. Bheldon was also the subject of (or merely mentioned) in a small number of articles in newspapers and journals.

Michael McGrath and that, "le cabhair Riobaird dhein an sagart cnuasach de chaint an Chomaraigh a sholáthar don Athair Pádraig Ó Duinnín" (With the help of Riobard the priest, i.e. Fr. McGrath, supplied a collection of words from the spoken Irish of the Comeragh region to Fr. Dineen). Fr. McGrath was an Irish-speaking priest in west county Waterford at the time. It has been suggested that Bheldon may not have been able to read or write in the Irish language (Ó Macháin, 1995: 8-9). One could infer, then, that the task of integrating Bheldon's input to Dineen's dictionary may have fell on McGrath. While it cannot be determined with any level of certainty what was the actual origin of the list one can surmise that it not only contained a collection of words and phrases from the spoken language of the region, but also examples from Bheldon's poetry. It is interesting to note, however, that in his *Sean-chaint na nDéise*, Bishop Michael Sheehan seems to suggest that the list was primarily of McGrath's making:

> A valuable list of Déise words and phrases was supplied to Dr. Dineen by Fr. Michael McGrath which he took down while curate in Kilrossanty (Comeragh) parish. They are indicated in the dictionary by "Cm." Fr. McGrath later became P.P. of Ring, and died in 1919. Many of his notes, not used by Dr Dineen, are in my possession, and may appear, I hope, in a future edition of this book (Sheehan, 1944: 216).

3 A Recategorisation of Gaeltacht na nDéise Entries in Dineen's Dictionary

The following sections present a narrative account of socio-cultural aspects of the Comeragh dialect contain in Dineen's *Foclóir Gaedhilge agus Béarla*. This analysis is founded upon the work of Nyhan (2006), who identifies 355 entries in Dineen's work that have provenance in this particular region. A small number of these entries have been selected, which depict emotional states (anger and anxiety) and social class (poverty).

3.1 Anger

Dineen's 1927 dictionary includes 15 entries from the Comeragh specific dialect that have connotations of acts of anger, dispute or violence (see Table 1). There are no evident regional factors explaining the choice of these particular words, apart from the backdrop of a national unrest in Ireland against foreign occupation. However, a closer inspection of Dineen's entries in this particular lexical set communicate certain feelings of regional tension, for example, *reaping with a swing round killed the Munsterman* found in the entry **baic** or in the agricultural endeavours strongly associated with the region, *the dog made a drive at me* in **ablach**. Most interesting is the finality or permanency associated with the acts of violence that are narrated through certain listings, such as *I will give you a lasting wound* (**cio-na-sheicean**), *He injured him for life/He is permanently marked or injured* (**failc**) and *you have finished him*

with that blow (**mart**). The explicit description of physical aggression, particularly referencing males, appears to be a strong theme in anger-specific lexis and could be connected to the powerful autocratic role of the Catholic Church in Ireland during this period, particularly given the religious background of Dineen and his named contributors.

Headword	Source Language Entry	Target Language Translation
Baic	*buaint tar b. do mhairbh an Muimhneach*	Reaping around with the sword killed the Munsterman
Ablach [2]	*Thug an mada a. orm*	The dog made a 'drive' at me
Ciona-sheicean	*In phr. cuirfead c. ort*	I will give you a lasting wound
Cofach	*c. chum troda atá ort*	You are spoiling for a fight
Donagar	*A thuilleadh donagair chughat*	May more misfortune be thine
Failc	*Chuir se f. ann* *Ta. f. ann*	He injured him for life He is permanently marked or injured
Faimin	----	A blow
Fuarpadh	---	One in a rigid or unconscious state after a blow
Giolcadh	*Fuair sé g. maith ón máighistir*	The teacher gave him a good beating
Liath-shúil	*Thug sé l.orm*	He eyed me bitterly
Máiglid	*Ag m. Le n-a chéile*	Act of wrangling, disputing
Mart	*Tá sé 'na mb. agat leis an mbuille sin*	You have finished him with that blow
Pliastrail	*Ag p. ar fuaid an tighe*	Knocking things about the house
Rámhghail	----	Ranting, raving, medley
Sméideadh	*ní leomhfá s. air*	You dare not wink at him

Table 1: Entries denoting Anger, Dispute or Violence.

Fractured personal relationships are also communicated through Dineen's content with some of these relating to formal social structures, such as school, *the teacher gave him a good beating* (**giolcadh**) or towards individuals in the wider regional community, for example, *he eyed me bitterly* (**liath-shúil**) or *you dare not wink at him* (**sméideadh**). Collectively, these entries express a certain malaise experienced by Dineen either in childhood or on the behalf of those who contributed examples of the Comeragh dialect to his dictionary.

3.2 Anxiety

Another subset of Comeragh dialect words recorded by Dineen convey a measurable sense of anxiety (see Table 2), which appears to be connected to personal, relating to geographical surroundings or associated with a recent personal loss. A particularly strong emotive theme is the portrayal of being un-

settled in a particular place, which makes possible reference to the region. This can be found in the entry **connuighim**, which includes the listing translating as *he was sad enough until he became familiar with the place*.

Headword	Source Language Entry	Target Language Translation
Cásnach	----	Full of concern
Connuighim	*Bhi sé diachrach go leor gur chonnuigh sé léis an áit*	He was sad enough until he became familiar with the place
Deighreán	*Ag déanamh deighrein dó*	Giving him anxiety
Diadhánach	----	Lonesome, as a cow bereft of her calf
Diúdaireacht	*chuirfeadh sé d. ar mo chroidhe*	It would rejoice my heart
Dorn	*dhein an ghaoth d. dubh orn*	The wind did me a bad turn
Fógla	*f. chum imthighthe*	Anxiety to depart
Iolchaing	*In phr. Tàim ar i. chum*	I am anxious to get at or be at
Ionnas	----	Expectation
Sonntaighe	*thuig mé go raibh mo dheirbhshiúr atá curtha le bliadhain lem ais agus tháinig s. orm*	I understood my sister, a year buried, was beside me and I became unnerved
Stiugaighil	*Ag s. les an mbás*	In the throes of death

Table 2: Entries denoting Anxiety.

3.3 Poverty

Certain entries chosen by Dineen in his account of the Comeragh dialect make a discernible reference to poverty and lower social class as being associated with its regional identity (see Table 3). These typically centre on the quality of food available and the culinary expectations of Comeragh people. In the case of the former, the entry **gannaire** clearly communicates the state of the social class stratum of this region by their choice of food. Another example of a clear statement of poverty can be found in the entry **beaindin** containing the citation *the tap of the cream vessel through which impoverish milk is withdrawn*.

Headword	Source Language Entry	Target Language Translation
Beaindin	----	The tap of a cream vessel, etc, through which the impoverished milk is withdrawn
Cearbh [1]	*Cuirim c. mo shúl ann*	I regard it with a covetous eye
Cuthaigh	*biadh c.*	Stimulative food
Forthain	*cead f. den bhiadh a dh'ithe*	Permission to eat enough of the food
Gannaire	*íosfaimíd an g.*	We will eat the poor food
Lóta	*lótaí na phócaí*	Scraps in his pockets
Molaim-mo-làmhadas	*Is dóigh le muinntir an ch. gur m. arán agas gruth*	The Comeragh people think that bread and curds make the finest of food

Table 3: Entries denoting Poverty.

4 Conclusion

The socio-cultural narrative of the Comeragh region in Dineen's *Fócloir Gaedhilge-Béarla* (1927) depicts a regional existence characterised by feelings of anger, anxiety and an acute sense of poverty. Although the thematic areas under analysis represent only a small proportion of Dineen's account of this particular Irish dialect, they allow an exploratory insight into the lexicographical portrayal of this geographical area, contextualising its emotional disposition and social stratum. Together, these contributed to shaping its regional identity of this early twentieth century period within a larger national context.

5 References

5.1 Dictionaries

Dineen, P.S. (1927) *Foclóir Gaedhilge agus Béarla*. Second Edition. Dublin: The Educational Company of Ireland Ltd.

5.2 Other Publications

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. London: Pearson Education.

Nyhan, J. (2006). *Findfhocla an Chomaraigh: Cnusach Riobaird Bheldon*. An Linn Bhuí (10). Kilkenny: Leabhair na Linne.

Ó'Macháin, P. (1995). *Riobard Bheldon: amhráin agus dánta*. Dublin: Poddle Press.

The *eLexicon Mediae et Infimae Latinitatis Polonorum.* The Electronic Dictionary of Polish Medieval Latin

Krzysztof Nowak
Institute of the Polish Language, Polish Academy of Sciences
krzysztofn@ijp-pan.krakow.pl

Abstract

The paper presents goals, methods, and results of the project of the Electronic Dictionary of Polish Medieval Latin. First, a brief history of the paper dictionary, as well as an account of its main features are presented. Second, the main problems of the metalexicographic analysis and the subsequent XML encoding of the lexicographic content are discussed. The main purpose of both being a fine-grained description of linguistic resource, it was necessary to make explicit a fair amount of data which are coded only by means of convention. Third, the web interface of the dictionary is treated in more detail. Its most important of its design principles include separation of the expert and novice user perspective, system of aids and suggestions, integration with external sources.
Keywords: electronic lexicography; Medieval Latin; dictionary interface; TEI XML encoding; implicit information; lexicographic convention

1 Introduction

The *eLexicon Mediae et Infimae Latinitatis Polonorum* (henceforth referred as the *eLexicon*) is an electronic dictionary based on the first 7 volumes[1] of the paper *Lexicon Mediae et Infimae Polonorum* (henceforth the *Lexicon*) which has been published since 1953 under the auspices and with the financial support of the Polish Academy of Sciences (Plezia, Weyssenhoff-Brożkowa, Rzepiela 1953). The *Lexicon* was conceived by its first editor, the Polish eminent philologist Prof. Marian Plezia, as a work which would fully document the use of the Latin language on the Polish territory between the Xth and the mid-XVIth century (Plezia 1958). As such, it was meant to form a part of the European network of the national dictionaries of Medieval Latin which started to emerge at the same time in response to an appeal of the *Union Académique Internationale* (Bautier 1981: 433–436). Users to which the print *Lexicon* has been addressed are in particular members of a research community, which is the reason why so much emphasis has been put, among others, on the completeness of the source material included. The print dictionary provides, then, in-depth etymological, morphosyntactic and semantic description of each

1 They include entries from A to Q, which is ca. 6000 pages printed in two columns.

word attested in the Polish Latin during the Middle Ages. Sense definitions are formulated both in Polish and – with foreign readers in mind – in Latin, and are illustrated with appropriate source quotations, if the meaning was not known in the Antiquity. The audience of the *Lexicon* being scholarly community, it becomes partially clear why the *Lexicon* does not make much concessions as far as user friendliness is concerned, with its heavy use of the typographic conventions, tightly printed columns etc. The paper dictionary suffers also from the drawbacks symptomatic for every long-term academic publishing enterprise, and in particular from several inconsistencies of the editorial policy, which affect especially usage labelling system, semantic change description or sense nesting practice, to name only few.

The project of the electronic dictionary which would be based on the *Lexicon* was conceived by the author of the paper and has been carried out between mid-2010 and mid-2014 by the team of the Department of the Medieval Latin of the Institute of the Polish Language (Polish Academy of Sciences) in Kraków.[2] Regardless of its roots, from the beginning the *eLexicon* was expected to become a research tool on its own and not merely a digitized version of the paper work. Firstly, its content was to differ to various extent from what can be found in the print volumes. One source of substantial modifications was the incorporation into the main text of the *addenda et corrigenda*, 'supplements and corrections', printed at the end of each of the 7 volumes of paper dictionary. Another one was both manual and automatic update of the lexicographic content. The members of the team (and, in the same time, current authors of the paper dictionary) had to eliminate most obvious errors and, where only it was necessary, to adjust the text to the modern editorial rules.

Secondly, the *eLexicon* was expected to provide research community with capabilities that the print dictionary could not offer. Apart from the simple search and browse features, the on-line dictionary was meant to offer access to the wealth of information encapsulated either explicitly or implicitly in the dictionary entries. At the same time, the *eLexicon* was conceived as a constituent of a larger text analysis framework. It had not only to be integrated with the digital library of the scanned images of paper slips, but also to be actively linked to the bibliography list of the medieval sources and to constitute a *sui generis* wrapper around the Medieval Latin corpus.[3] Moreover, the on-line dictionary was to incorporate a fair amount of external resources, whether it be through locally triggered queries or by means of outward linking.

Finally, the *eLexicon* has been planned as an open-access and open-source project. From the beginning the access to the web service was meant to be free and unlimited, as was also the case of the XML annotated dictionary files, which are to be distributed under liberal licenses. Although there were many reasons to do so, the main of them was assuring the longevity of the project, a major challenge in academic projects with time-limited funding. The other factor expected to contribute to project's longev-

2 Its funding was provided by a grant of the Polish National Science Centre awarded to the chief-editor of the paper dictionary, Prof. Michał Rzepiela.

3 A 5 million words, balanced and representative corpus of the Polish Medieval Latin is now being developed by the same team and is due to be delivered by the end of the 2016.

ity was compliance with standards. Developed firstly as a set of the TEI-conformant files (TEI Consortium 2013), the electronic dictionary allows platform-independent implementations, the fact which implies two major consequences. On the one hand, one can benefit from the open-source technologies and existing text or data retrieval frameworks. On the other, one may hope that the available lexicographic data will be incorporated in other research contexts, integrated with NLP infrastructures, and, consequently, they will be steadily ameliorated and refined, even when the project itself comes to an end.

2 Methods

No phase of the e-dictionary creating was outsourced. After the volumes of the print dictionary had been scanned, the image pre-processing and OCR process began as a result of which machine-readable text was, firstly, obtained and, then, carefully proofread. After that metalexicographic analysis followed, its aim being twofold. First of all, it was expected to reveal the features of the print dictionary macro- and micro-structure to be retained in the *eLexicon*, but also to select lexicographic information worth retrieving by means of the on-line search interface. Contrary to what one might believe, first part of the analysis was far from trivial, since it was often equivalent to questioning the very foundation of the paper dictionary methodology and, at the same time, to designing principles of the future on-line dictionary. The main issues addressed included internal reference system, approach to the entries with deeply nested structure, status of idioms and multi-word expressions as lexical units etc. In what concerns lexicographic data, the guiding principle was to retrieve and make explicit as much linguistic and non-linguistic information as possible, since from the very beginning it was clear that the on-line dictionary should serve researchers of various expertise in medieval studies, from the historians working on the Medieval Latin sources, to the Latin and Polish linguists, to the historians of literature, art, philosophy and science. What is more, one of the goals of the *eLexicon* was also expanding the audience of its paper predecessor beyond the scholarly world to embrace students and teachers of Latin.[4] In order, then, to satisfy needs of the academic users[5], on the one hand, and to effectively distinguish between expert and lay users on the level of the web interface, on the other hand, a highly structured resource had to be created.

Secondly, the lexicographic analysis served also two other purposes, the first of them being to estimate the feasibility of the data annotation within project's time limits, that is without resorting to advanced NLP methods, the second – to conceptualize the dictionary macro- and microstructure by means of the TEI XML tagset. Although encoding standards in linguistic annotation constitute nowa-

4 Not only Medieval, but also Classical Latin, since there does not exist as yet any on-line Polish dictionary of Classical Latin, at least academic one.

5 Or what was believed to be their needs, since to my knowledge there do not exist any empirical studies of the needs of the users of (academic) Latin dictionaries.

days a topic on their own (Garside, Leech, McEnery 1997; Pustejovsky, Stubbs 2013), I will limit myself to indicating three main reasons why the TEI XML has been chosen as an output format of the dictionary files. First of them has been already mentioned: storing lexicographic data in text files (contrary to binary ones) makes them at least partially immune to platform or software-related issues. XML encoded resources are human-readable, so they may be easily subject to modifications, adaptations and further refinement even by less technical-oriented users. Secondly, the TEI XML encoding serves well the purpose of storing highly structured, paper-born documents. Since the print dictionary being a starting point of the *eLexicon* is a result of 60 years' work, not only is it far from unified, but it also makes heavy use of sense nesting, *ad hoc* usage hints etc., all of which makes putting it into database format a non-trivial task. Thirdly, the use of widely supported formats and standards becomes essential, if one wishes to benefit from the already existing software solutions, on the one hand, and, on the other, to make one's data useful in yet unpredicted research environments. As to the former, there was no intention to create from scratch a proprietary interface to serve the dictionary content. In fact, the *eLexicon*, rather than being a closed interface solution, attempts to initiate discussion about what tool do the medievalists need and to dynamically change as the community will express its expectations. In that it differs significantly from now outdated in their design, closed-source and paid resources, such as the *Database of Latin Dictionaries* published by Brepols[6]. The availability of the XML annotated files, in turn, should encourage dictionary content reuse, whereas applying the TEI recommendations should facilitate data exchange, as, despite their drawbacks, they were generally adapted in other open Medieval Latin dictionary projects, such as precursory digitisations of Lewis and Short's *A Latin-English Lexicon* (1879)[7] by the Perseus Project team (Crane, Seales, Terras 2009; Bamman, Crane 2009), DuCange's *Glossarium* by Ecole Nationale des Chartes (Glorieux, Thuillier 2010),[8] *Novum Glossarium* in frame of the project Omnia (Bon 2009; Bon 2010; Bon 2011).[9]

Once the analysis had come to an end, the annotation guide was created and the annotation itself started. After the XML files had been generated through PERL and XSLT processing of the OCR output, they were next distributed among project's team members who diligently proofread them, modified when necessary the dictionary content and adjusted automatic encoding. Verified for their well-formedness, the files were next validated with a previously generated *Document Type Definition* (DTD). The web interface of the *eLexicon*, which will be treated in more detail below, was built around the eXist-db, a free and open-source, XML native, no-SQL database running in the back-end. Programmed as a set of XQuery scripts, it produces on the front-end a light-weight HTML5+CSS web ap-

6 http://www.brepolis.net/.

7 http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0059, now consultable also within Perseus under Philologic Project (http://perseus.uchicago.edu/Reference/lewisandshort.html) or, in a more convenient way, within Logeion project (http://logeion.uchicago.edu/).

8 http://ducange.enc.sorbonne.fr/.

9 The TEI was also employed in other historical lexicography projects, such as the Anglo-Norman Hub (http://www.anglo-norman.net/,). For details, see (Trotter 2011).

plication. In order to introduce some elements of interactivity, as well as to provide users with instant headword suggestions and similar features, a moderate amount of jQuery scripting has been added.

3 Results

3.1 The XML Annotation

The principal rule applied in the process of the XML encoding was, as was already mentioned, to make as much lexicographic information explicit as possible without caring so much about the typographic peculiarities of the print dictionary. Formatting information was generally "translated" into appropriate "semantic" annotation and retained only if it could add to the on-line dictionary features. No special effort, then, was made to save indentation or font, since they should be otherwise easily deducible from the semantic encoding. Page and line numbers have been preserved in order to ensure correct resolution of the cross-references and, thus, successful intra-linking[10].

Each of the output files corresponds to one print volume and is preceded by a metadata header in which basic bibliographic information was recorded. There is, however, nothing that could prevent prospective users from decomposing original files according to their specific needs. As far as the annotation design is concerned, dictionary entries have been first separated from each other[11]. For the purpose of subsequent processing, unambiguous identification of the entries had to be assured by means of the automatically generated identifiers, however, the lemma-based identification has been also retained[12]. The entries were next classified[13], so as to distinguish between standard and reference entries of the type:

- "LETANIA *cf.* LITANIA", where one of the orthographic variants points to the canonical word form;
- "LETARG ... *cf.* LETHARG ...", where a word fragment (most frequently word prefix) points to the position in the dictionary text rather than to a precise headword.

Since the entry access in the *eLexicon* was meant to be subject to major redefinition, further refinements needed to be applied to the selected "secondary headwords" (Atkins, Rundell 2008: 235-236), with the most significant example being derived forms. Although such forms as n. *laureus* 'laurel' which is to be found in the paper dictionary as a sub-entry of the adj. *laureus* 'laurel', remain embedded in their respective superordinate entries, they will also function as separate lexical units during

10 In the *Lexicon* cross-references come generally in two forms and may point either 1) to a precise entry or one of senses (e.g. *„Cf.* LATIO *II*" under LEGISLATIO), or 2) to a volume, page and line(s) of the dictionary (e.g. *„cf. supra I 1076,49 sqq.*" under LEX which may be rendered as 'cp. above, [volume] I, [page] 1076, [line] 49 and foll[owing]').

11 For this purpose <entryFree> tag („unstructured entry") was used which allows for a more liberal encoding of the paper-born and, thus, text-oriented dictionaries.

12 Here, the attributes @xml:id ("identifier") and @n ("number") were used.

13 By means of the @type attribute, with e.g. homonyms labelled as @type="hom".

alphabetical browsing or when listed in results lists.[14] The same can be said about other instances of secondary entries, such as multi-word or idiomatic expressions.

Orthographic, etymological and morphosyntactic information was subject to diligent, fine-grained encoding. In spite of the privileged position that a headword occupies in traditional lexicography, careful annotation of the variant orthographic forms is essential for a Medieval Latin dictionary to be a serious tool of research, and that from many reasons. With orthography changing often within one manuscript from scribe to scribe, a lexicographer can never know which word form dictionary user may be looking after, which makes selecting unique, "canonical" form somewhat anachronistic, if one takes into account the medieval sense of language correctness.[15] What is more, in the print *Lexicon* one can find headwords which serve only a purpose of identifying entries and they have been never attested in medieval texts. This is the case of such entries as "[LAVANDA] *s.* LAVENDA" 'lavender' , where the square bracket is employed to indicate that, unlike *lavenda*, the form *lavanda* does not occur in sources, but, in turn, was used by the lexicographer only as a conventional representation of the entry as an ideal, in his or her opinion, reconstruction of the Italian *lavanda*.[16] In a case like the one just mentioned, separation of the genuine linguistic material from what is only a pure convention becomes crucial.

Without going into unnecessary detail, let it suffice to say that the other elements of the grammatical description of the headwords were subject to likely minute encoding, which, apart from its direct and obvious goal, namely that of description of linguistic resource, had two secondary objectives. First of them was making implicitly coded lexicographic data fully explicit. The variety of the information which is tacitly conveyed in the *Lexicon* is striking, however, what may be only challenging for a human reader, if she is not accustomed enough to dictionary convention, makes a good deal of data inaccessible for automatic processing. Among those pieces of information which could be lost, if they were not scrupulously deduced from sometimes cryptic metalanguage and, then, redundantly added to the original files, one can mention part-of-speech labelling, which is explicit (that is, expressed with appropriate labels) for adverbs or sub-headwords,[17] but for verbs, nouns and adjectives is to be inferred from the inflectional information.[18] The same is true about the language from which the headword originated, since appropriate labels are in the print dictionary employed uniquely for languages other than Latin, so, for instance, while the entry LEXICON 'a dictionary' includes a self-explanatory etymology "*Gr. λεξικόν*" (where *Gr.* stands for "Greek"), in the entry LICENTIO 'to give a license' one

14 Here <re>, ie. „related entry", tag was used.

15 The problem of the abundance of word forms is even more striking for Medieval Latin as was used in France or Spain, where it experienced substantial assimilation to a vernacular language.

16 *S.* is here abbreviation for the Lat. *siue* 'or'. Such notation can be found on a regular basis when hypothetical Classical Latin form of the word is reconstructed, see, for example, „[RHINOCERON] *s.* RINOCERON" 'rhinoceros' .

17 See, for example, LICENTIOSE *adv.* 'violently'.

18 See, for example, entries for a verb LICENTIO, -are, -avi, -atum 'to give a licence', a noun LICENTIA, -ae *f.* 'licence' or an adjective LICENTIOSUS, -a, -um 'licentious', for which PoS information should be determined from inflectional description.

finds notation "licentia", from which one should infer that the word was coined during the Middle Ages from a Classical or Medieval Latin term. To make things even less transparent, the entries like LEX have no etymology at all, which, in turn, means they were inherited from the Classical Latin.

The list of information types which are encoded only indirectly is, naturally, far from complete and should also include such important features of historical lexicography as time and place of word's attestation. While geographical information is never explicitly given in the paper *Lexicon*, chronological data are provided for the sake of precision, that is only if the source quotation comes from a work which includes multifarious, chronologically diverse material.[19] Otherwise, spatio-temporal characteristics of the quotation should be deduced from the alphabetical list of the dictionary sources. Yet, there are many reasons why information of this sort should be accessible in the on-line dictionary and, thus, why it should be also explicitly declared in the XML files. The reason that comes first to mind is, obviously, more efficient and straightforward retrieval of these data within the search interface of the on-line dictionary. The other reason why aforementioned, but also e.g. genological properties of source quotations should be explicitly encoded is that it could greatly facilitate its interactive representation in form of maps, timelines or charts, as the example of the Wiki Lexicographica (Bon, Nowak 2013) demonstrated.

The other of the secondary goals of dictionary encoding was standardisation of the lexicographic description. This included, for instance, eliminating domain or usage labels which are now obsolete or were coined *ad hoc* at some point of the dictionary writing process and shortly after fell from use.[20] On the contrary, some of the subtle or nowadays less useful distinctions were subsumed on the encoding level under general or more frequently used ones. This was the case of the labels indicating direction of the semantic change. Thanks to their unification, one will get an access to words which experienced metaphorical extension, although they were originally marked in the paper dictionary either with the standard label *metaph.* standing for *metaphorice* 'metaphorically', or with a more verbose label, *in imagine* 'in the image of'.

It should be also added that the XML encoding, apart from its obvious, data-oriented objectives, has many practical, user-oriented ramifications. In the print dictionary, for instance, sense definitions are given, as was already noticed, both in Polish and Latin. Clear separation of the definition strings not only allows their subsequent retrieval and reuse, but also, on more practical level, allows Polish users to consult the on-line dictionary in their mother tongue, while serving foreign researcher with a Latin version of the entry. Fine-grained linguistic data encoding, in turn, facilitates differentiating basic

19 For example, the only quotation which can be found under the headword LEXICON is labelled as „AKap p. 61 (a. 1540)", where „AKap" is a source identifier (pointing to a multifarious collection of the chapter tribunal), „p." stands for *pagina* 'page' and „a. 1540" is a chronological hint which should be resolved as „anno 1540", ie. 'in the year 1540'.

20 Naturally, XML encoding allowed for more obvious ameliorations as well. It was possible, for example, to introduce explicit distinction between domain (e.g. *astr.* for 'astrology', *eccl.* 'ecclesiastical term') and attitude (eg. *in malam partem* 'pejorative') labels on the one hand and the syntax markers (such as *intrans.* for 'intransitive' or *refl.* for 'reflexive') on the other.

and advanced user scenarios, and enables adapting lexicographic content perspective to the varying user needs.

3.2 The Web Interface

Apart from the obvious goal of overcoming the well-known drawbacks of paper dictionaries, the web interface of the *eLexicon* was created in order to facilitate advanced retrieval of the data obtained in the process described above. Thus, expected to constitute the main entry point to the electronic dictionary and other tools of textual studies, it was meant to provide professional users with a fully-fledged research platform. At the same time, however paradoxical it may appear, it had to satisfy the needs of less-advanced users, students and language teachers, by clearly separating basic and advanced perspective on lexicographic content. In order to serve well both groups, namely that of expert, as well as that of novice users, the guiding principle of the web interface creation became to help users better understand what they are looking for and to produce meaningful output, even if the phrase the user looked for, was not found in the dictionary.

When visiting the *eLexicon* page for the first time, users are proposed a quick tour of the search and browse features the dictionary offers. The main page is not meant, however, to overwhelm a visitor with a plethora of options (Figure 1). Rather the contrary is true, since apart from the simple menu, which gives direct access to the search and browse interface, it does not display anything but a simple search form which is, though, an actual entry point to the dictionary content.

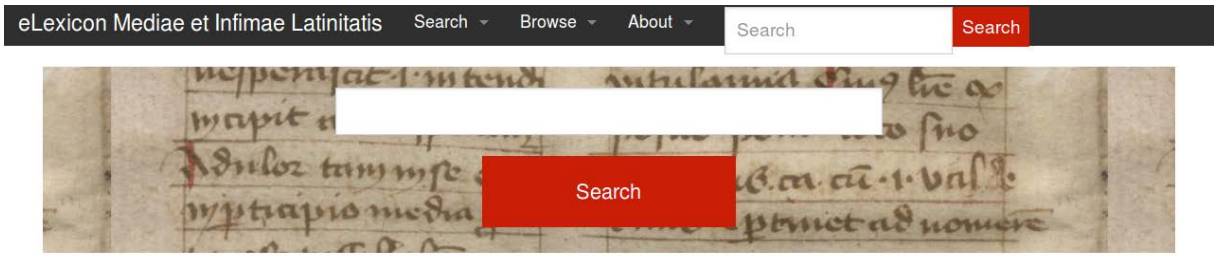


Figure 1: The on-line dictionary: the main page.

Its underlying logic is to support two expected use scenarios:

- a user is looking for a lemmatised word form, for which there exists a corresponding main or secondary headword;
- a user is looking either a) for an inflected form, a Polish or other non-Latin term, a Latin word which is not attested in Polish sources, or b) for an incorrect word form.

The first scenario, i.e. successful lookup of a headword included in the *eLexicon*, is promoted by means of the Ajax-based suggestion list which appears once the user types in three first letters of the phrase

she is looking for.[21] The list of the words suggested consists of all the orthographic variants of the headwords included in the dictionary, as well as of the multi-word expressions and idioms, which are, however, still distinguishable from the former thanks to their different formatting. The suggestion list not only should significantly speed up the lookup process, but also may handle potential typing errors and, since Latin is an inflected language, point user to a correct lemma of the word she is querying. Once suggested option is selected, the user is redirected to the appropriate entry.

Here, two perspectives on the dictionary content are provided as separate tabs which, once clicked, reveal, respectively, a basic and a full content view. First of them (Figure 2), under clearly separated headings presents selected morpho-syntactical properties of the word, a brief overview of its meaning, as well as various summaries of its use.[22] As such, it should aid the novice, as well as expert users to get the general idea of the word they are looking for, without necessarily overwhelming them with the full apparatus of the academic lexicography.

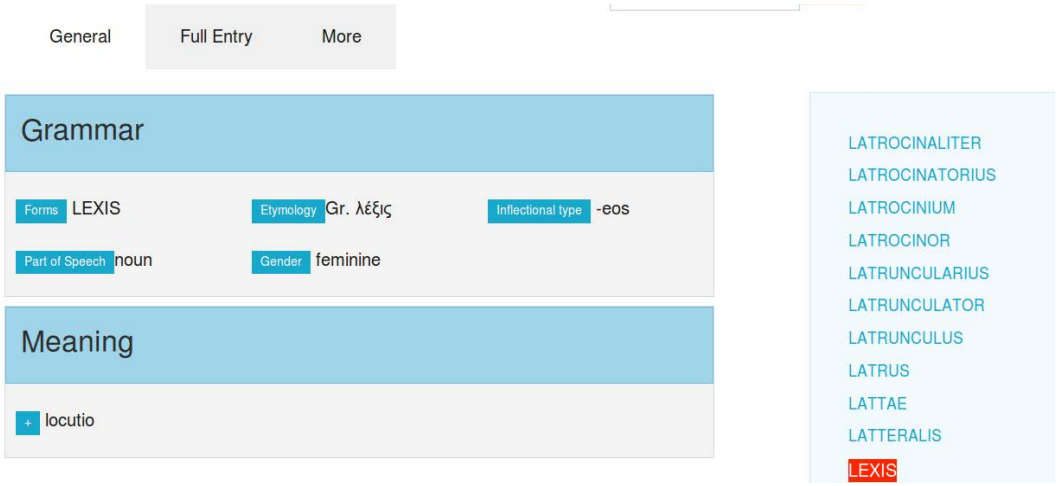


Figure 2: The on-line dictionary: a single entry ("Basic View" tab).

For both user groups the basic view may also be a convenient access point to the full version of the entry. The latter, in turn, makes heavy use of the CSS and JavaScript styling in order to improve the readability of the paper-born entry and facilitate information retrieval.

If the user does not decide to follow the suggestions and her query does not correspond directly to one of the entries, she is taken to the disambiguation page, where the second of the aforementioned scenarios is handled. User's input is being processed and searched for in the sections of the *eLexicon* different than headwords.

21 The threshold was selected as a compromise between acceptable server load and usefulness. It is certain that it will be adjusted, once the user search logs are collected and analysed.

22 It is strongly inspired by the basic view previously implemented in the WikiLexicographica (Bon, Nowak 2013).

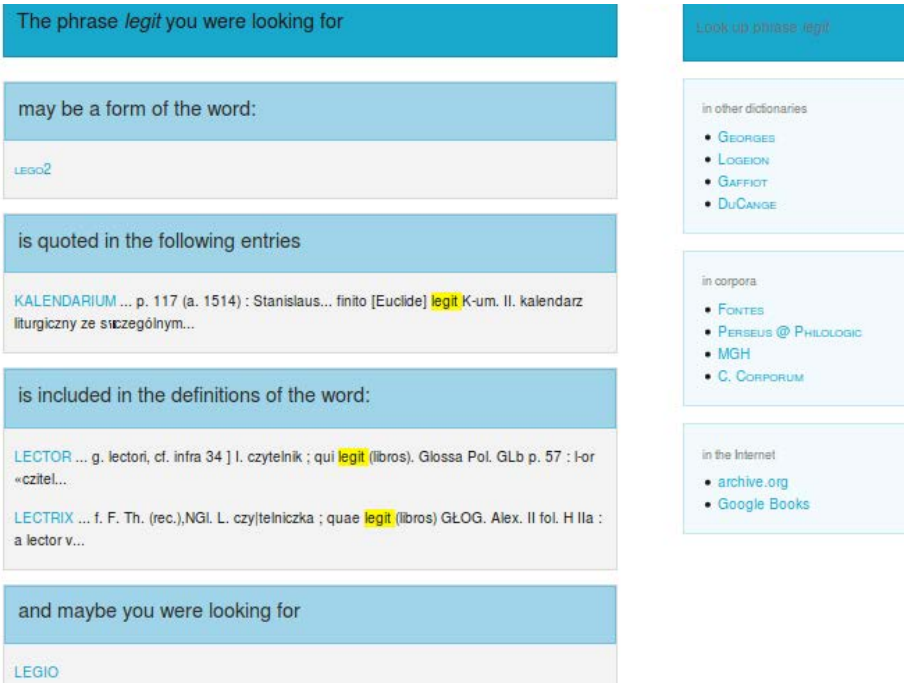


Figure 3: The on-line dictionary: disambiguation page.

The user is next presented with a result list (Figure 3) which, depending on the case, includes all or only some of the following parts:

- lemmatised form, in case if the user's input was an inflected form of a word;
- results of the word lookup in source quotations and/or definitions, if the user was using the dictionary as a source of attestations or synonyms, if she was after translation of a Polish term or if she was mimicking onomasiological search;[23]
- suggestions of similar words, if the input does not yield any meaningful result, so instead a correct word form should be suggested.[24]

The *eLexicon* content, however, may be accessed not only from the simplified main page search form, but also from the browse and expert search pages. The former functions as an equivalent of turning pages of the paper dictionary. Entries may be, then, selected by specifying the respective volume, page and line of the print edition. What is, however, more important, the browse interface offers dynamic (changing as the user is typing), synchronous lookup of the user's input at the beginning, in the middle and at the end of the dictionary headwords. This seemingly trivial feature was included to serve especially the Medieval Latin paleographers and manuscript readers in general who used to consult dictionaries looking for a reading suggestion of a hardly legible characters rather than for precise sense explanation.

23 By looking, for example, for all words of which the Latin definitions employ word *color* 'colour'.

24 Since the eXist-db makes use of the Lucene engine for text search, this feature is implemented as a fuzzy search of the user's phrase in the dictionary headwords. As the default Levenshtein distance value (Jurafsky, Martin 2009: 152) looks to be too liberal to produce helpful output, it will be certainly adjusted, once the data about the actual queries are collected.

The third access point is constituted by the advanced search facility, which takes full advantage of the scrupulous XML encoding of the lexicographic information (Figure 4)

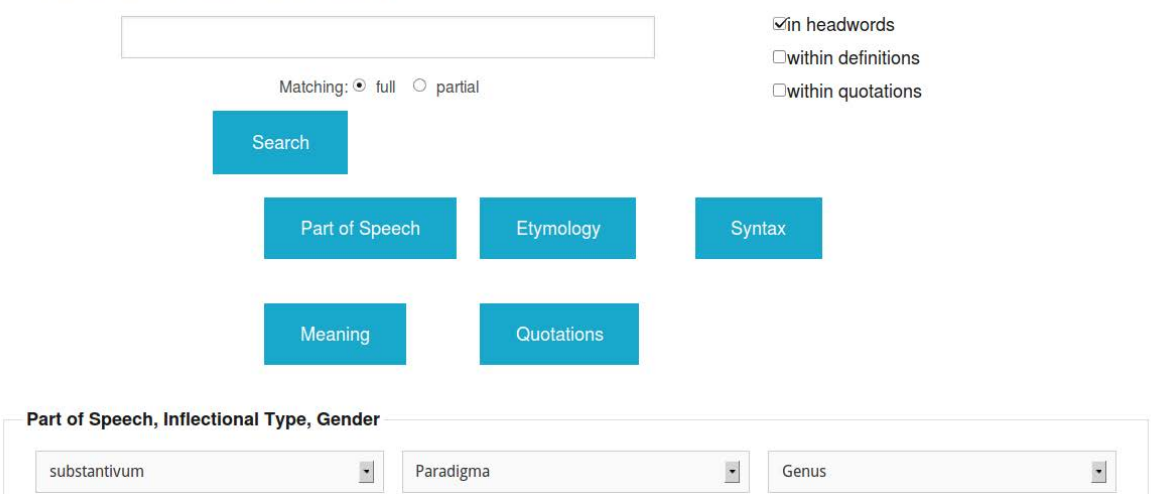


Figure 4: The on-line dictionary: an advanced search page.

The user is here provided with a search form consisting of two main parts:

- the text input field, in which query string should be typed. The user may further specify scope of her search,[25] as well as matching strategy of her choice.[26]
- the list of additional restrictions to apply to search results. The user is free to refine her query and limit results by means of the morphosyntactic, etymological, semantic and chronological criteria.[27] In case the query string is not specified, selected criteria are applied to all dictionary entries and the interface functions as a tool of the exploratory analysis of the Medieval Latin lexicon. The results page allows for further refinement, since it contains a filtering list of linguistic properties of the previously queried headwords, which behaves in a manner similar to faceted browsing widgets.

As was already mentioned above, the *eLexicon* is expected to become a centre of a fully-fledged research platform. For that purpose, it is meant to be closely bound with the corpus of the Polish Medieval Latin. Although there still remains much work to be done, even now, whenever possible, use is made of the already existing external resources that are expected to be of help for the *eLexicon's* expert users. Since, as for now, resources in question are stored externally, only appropriate links to the freely available corpora, dictionaries or on-line encylopedias may be provided to the users. In not so distant future, however, external resources are planned to be exploited locally and the content of at least some

25 That is, specific section of the dictionary entry within which the phrase should be looked for (currently options are limited to headwords, quotations and definitions).
26 That is, whether exact or approximate matching search should be applied.
27 Therefore, each query may be restricted, for instance, only to entries belonging to a specified inflectional class, originating from certain language, labelled as technical terms of a given domain or attested only in certain period.

of them will be directly embedded in the *eLexicon* search results.[28] In the current state of the interface, external resources are displayed:

- as a sidebar on the disambiguation page, as a way to suggest to the user other localisations in which she may find word absent from the *eLexicon* (see Figure 3 above);
- as a separate tab under a single entry view, so as to extend lexicographic perspective with corpus and knowledge base extracted data (Figure 5).[29]

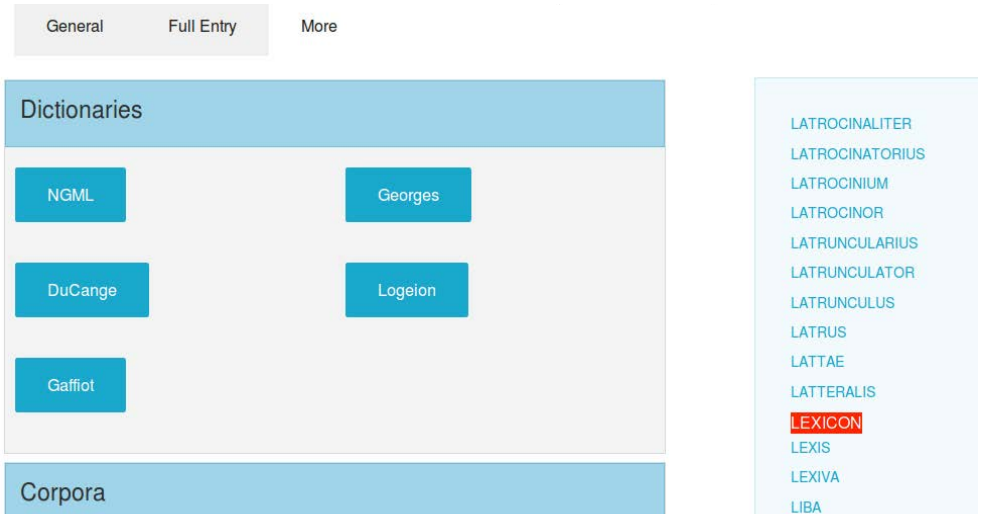


Figure 5: The on-line dictionary: a single entry ("More" tab).

4 Conclusions

The electronic dictionary of Polish Medieval Latin may become an important tool of the medieval studies, and this for many reasons. It will be freely accessible both in form of diligently encoded XML files and as a research-driven web application. To provide user with better insight into the medieval lexicon, the internet dictionary employs external sources, either by means of dynamically linking or direct embedding. Advanced users not only should find its single entry more readable, but they also will benefit from a configurable expert search and browsing interface, which provides the *a tergo*-like lookup. In turn, clear separation of a basic and advanced perspective on lexicographic content, as well as the use of suggestion lists and disambiguation pages may contribute to its becoming an effective tool for the Latin language students and teachers.

At the same time, there, naturaly, still remains much room for improvement. As far as data presentation layer is concerned, maps, timelines, charts and other alternative displays need to be implement-

28 This is the case of the freely available volumes of the *Novum Glossarium Mediae Latinitatis* or the *Glossarium* of DuCange, but also of the texts collected in the Perseus Library. The similar approach has been already applied in such inspiring tools as *Logeion* (http://logeion.uchicago.edu) or *Le Dictionnaire vivant de la langue française* (http://dvlf.uchicago.edu/).

29 In its current form, it is clearly still very far from being fully implemented.

ed. There is also a serious NLP work which has to be done, since the *eLexicon* is expected to provide conceptual search interface and to better integrate with knowledge bases.

5 References

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

Bamman, D., Crane, G. (2009). Computational Linguistics and Classical Lexicography. In *Digital Humanities Quarterly*, 3(1). Accessed at: http://www.digitalhumanities.org/dhq/vol/3/1/000033/000033.html [10/04/2014].

Bautier, A.-M. (1981). La lexicographie du latin médiéval. Bilan international des travaux. In La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du Moyen Age: Paris, 18-21 octobre 1978, Paris: CNRS, pp. 433–53.

Bon, B. (2009). OMNIA – Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins. In *BUCEMA. Bulletin du centre d'études médiévales d'Auxerre*, 13, pp. 291–92. Accessed at: http://cem.revues.org/11086 [10/04/2014].

Bon, B. (2010). OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2). In *BUCEMA. Bulletin du centre d'études médiévales d'Auxerre*, 14, pp. 251–52. Accessed at: http://cem.revues.org/11566 [10/04/2014].

Bon, B. (2011). OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3). In *BUCEMA. Bulletin du centre d'études médiévales d'Auxerre*, 15. Accessed at: http://cem.revues.org/12015 [10/04/2014].

Bon, B., Nowak, K. (2013). Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langements, M. Tuulik (eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013. Tallinn - Ljubljana: Trojina, Institute for Applied Slovene Studies, Eesti Keele Instituut, pp. 407-420. Accessed at: http://eki.ee/elex2013/proceedings/eLex2013_28_Bon+Nowak.pdf [10/04/2014].

Crane, G., Seales, B., Terras M. (2009). Cyberinfrastructure for Classical Philology. In *Digital Humanities Quarterly*, 3 (1). Accessed at: http://www.digitalhumanities.org/dhq/vol/003/1/000023/000023.html [10/04/2014].

Garside, R., Leech, G. N., McEnery, T. eds. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London-New York: Longman.

Glorieux, F., Thuillier, S. (2010). Grec ancien, latin médiéval, balisage comparé de deux dictionnaires, vers des ressources linguistiques. In *ALMA. Archivum Latinitatis Medii Aevi*, 68, pp. 161–81.

Jurafsky, D., Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River: Pearson Prentice Hall.

Lewis, Ch. T., Short, Ch. (1879). *A Latin Dictionary*. Oxford: Clarendon Press.

Plezia, M. (1958). Lexicon mediae et infimae latinitatis Polonorum. In *ALMA. Archivum Latinitatis Medii Aevi*, 28, pp. 271–84.

Plezia, M., Weyssenhoff-Brożkowa, K., Rzepiela, M. eds. (1953). *Słownik łaciny średniowiecznej w Polsce. Lexicon mediae et infimae Latinitatis Polonorum*. Vols. 1–8. Kraków: Wydawnictwo IJP PAN.

Pustejovsky, J., Stubbs, A. (2013). *Natural language annotation for machine learning*. Sebastopol, CA: O'Reilly Media.

TEI Consortium. (2013). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version: 2.5.0. Accessed at: http://www.tei-c.org/Guidelines/P5/ [10/08/2013].

Trotter, D. A. (2011). Bytes, Words, Texts: The Anglo-Norman Dictionary and Its Text-Base. In *Digital Medievalist*, 7. Accessed at: http://www.digitalmedievalist.org/journal/7/trotter/ [10/04/2014].

Acknowledgments

The *eLexicon Mediae et Infimae Latinitatis Polonorum* as well as the work on the present paper were supported by a research grant of the Polish National Science Centre (*eLexicon Mediae et Infimae Latinitatis Polonorum (A-Q)*, nr 3736/B/H03/2011/40).

From DANTE to Dictionary: The New English-Irish Dictionary

Pádraig Ó Mianáin, Cathal Convery
Foras na Gaeilge, Dublin
pomianain@forasnagaeilge.ie, cconvery@forasnagaeilge.ie

Abstract

Most major bilingual dictionary projects tend to largely involve adapting an existing bilingual dictionary, by adopting or amending the existing source language material as required and then supplying the necessary target language material. This paper describes the innovative approach followed in the production of the New English-Irish Dictionary (NEID) project. The online version of NEID was launched in 2013 (www.focloir.ie) and the project is to be completed when a printed version is produced in 2016. The English language content of NEID is based on the Database of ANalysed Texts of English (the DANTE database), a ground-breaking corpus-based lexical database developed specifically for the project. Attention is drawn to how the DANTE entry frameworks evolved through the various translation and editing processes to the final entries now available in the online. The paper also discusses the development of the Irish-language material and details some of the challenges, practical and editorial, encountered in the course of the production of the dictionary, and the working solutions that were developed.
Keywords: DANTE; bilingual; Irish-language

1 The New English-Irish Dictionary

The New English-Irish Dictionary (NEID) is being produced and funded by Foras na Gaeilge, the inter-governmental body with responsibility for the promotion of the Irish language, with a project budget of €6m. It is the first major English-Irish dictionary since de Bhaldraithe's English-Irish Dictionary (1959) and the first major bilingual dictionary in Ireland since Ó Dónaill's Foclóir Gaeilge-Béarla [Irish-English Dictionary] (1977). The first version of NEID was launched online in January 2013 with new material and revisions being uploaded at regular intervals until the envisaged completion of the online edition in 2015. The NEID will then be published in printed format in 2016.

2 Project Timeline and Phases

The NEID project started in 2000 and is on target to be completed at the end of 2016. The project is supported by IDM's DPS platform along with the Entry Editor interface. The online version of NEID

will eventually contain c. 130,000 sense units (c. 40,000 headwords) by December 2015, with a print version to follow in 2016. The project is divided into three major phases:

- **Phase 1: Planning and design** (started 2000, completed 2006). The planning and design phase was carried out by Lexicography Masterclass and delivered key elements such as an overall project plan, the English-language and Irish-language corpora which underpin the entire project, as well as sample entries, headword lists, draft style guides for each phase etc.
- **Phase 2: Compilation, Writing and Editing** (started 2008, to be completed 2016). This phase concerns the actual writing of the dictionary, and is discussed in detail below.
- **Phase 3: Publication** (started 2012, to be completed 2016). The online and mobile platforms were launched in January 2013 (www.focloir.ie) with about 30% of the final content of the dictionary; the online content is being added to on an incremental basis. The dictionary will also be made available as an app in 2014.

A number of separate databases were created to facilitate the progress of entries from the DANTE database through the various translation and editing stages to the online entry. The main databases are shown in Figure 1 below along with their purpose and some of the more significant changes between databases:

DATABASE	DESCRIPTION	PROCESSED BY
NEID 2A	**The source-language phase**	English-language lexicographers
NEID 2B	**The target-language phase** Changes automatically implemented: • Compounds extracted as full HWDs • FREQUENCY field inserted • TRID (translator ID) field inserted • TrCnt (translation container) inserted at set positions within entries	Translators
NEID 2C	**The editing phase (at sense unit level)** • Initial editing • Senior editing Changes automatically implemented: • EDID and SENED (editor & senior editor ID) fields inserted • reporting of style guide non-compliance	Editors
NEID 2D	**The final entry phase** • Sense units merged where possible • Disambiguators inserted • Streamlining of DOMAINs etc Changes automatically implemented: • Sense units grouped in blocks by POS • Links to grammar and sound files automatically inserted	Senior Editors
eNEID	**Content uploaded to website** Changes automatically implemented: • All fields not to be displayed are deleted (MEANING, FREQ, EDID etc)	IDM

Figure 1: The main databases involved in the phased transition between the DANTE database (designated NEID 2A within the context of the NEID project) and the final online entries.

3 The Source-Language phase: the DANTE database (began 2008, completed 2010)

Lexicography Masterclass also supplied the English-language lexical database on which the source language content of the dictionary is based. It has subsequently been made available as a resource for various other lexicographical and linguistic projects as the Database of Analysed Texts of English or the DANTE database. The entire database can be viewed and explored at www.webdante.com.

DANTE is based on systematic analysis of a 1.7-billion-word corpus of English supported by the Sketch Engine corpus query package and the GDEX feature for optimising the corpus example selection. The final package is a highly-structured 16-million-word lexical database of English that can be used not only by lexicographers but also by linguists, researchers and teachers. From a purely lexicographical point of view, DANTE is target-language-neutral in the sense that, despite its origins as the foundation for a bilingual dictionary, its sole focus is on describing and evidencing the English language without regard for any potential companion language (the Irish language specifically in the NEID project). As such, DANTE is equally a primary resource for monolingual and bilingual English dictionaries.

The headline statistics of the DANTE database give an indication of its coverage: 42,000 headwords (62,000 if compounds are counted as separate headwords), 12,000 phrases and 3,000 phrasal verbs, 149,000 sense units in total, over 600,000 relevant corpus examples and 16 million words in all. However the real value of the database is seen in the fine detail of its coverage within the entries, for instance:

- 42 different grammatical structures for verbs, 16 for nouns and 15 for adjectives (see Table 1 below);
- all headwords classified by 12 levels of complexity;
- entry templates;
- over 150 domain categories;
- regular proformas for closed set entries; etc.

From a purely NEID perspective, the DANTE database more than adequately served its purpose of providing the translation team with a comprehensive and detailed lexical profile of each English-language headword which was a candidate for inclusion in the final dictionary.

4 The target-language phase: adding the Irish-language content (began 2009, completed 2012)

The aim of the translation phase of the NEID project was to maintain the structure and detail of the English-language entry frameworks while providing the editing team in the next phase with a comprehensive and detailed translation database, by adding as many relevant Irish-language equivalents

as possible to the English-language frameworks. It was at this stage that Foras na Gaeilge's own lexicography team took over the writing of the dictionary, as the translation and editing phases would be entirely driven by the Irish-language requirements.

The most efficient way of implementing this phase was to clone the DANTE database as a separate translation database and, as part of the cloning process, to automatically insert translation-specific fields in specified positions within the entry frameworks. Also, all compound entries were extracted from their mother entries (as they are in DANTE) and were promoted to full headword status, not least to facilitate the distribution of work batches in a more efficient manner. The end product of the translation phase is a rich database of bilingual entry frameworks consisting of the unabridged DANTE, plus 4 million words in Irish in over 600,000 translation fields.

Given the size and nature of the DANTE database, the translation task posed significant challenges at a number of levels, particularly given the lack of experienced Irish-language lexicographers available to the project. There were three main challenges:

- **The size of the translation task**. It would have been impractical and quite pointless to translate the entire English-language database, particularly as the specific remit of the translators was to provide as many relevant translations as possible in Irish. The translators were directed to concentrate on the key element or node of the word or phrase in question, and not to translate any of the surrounding text in the supporting examples unless it impacted on the translation; this was facilitated by automatically highlighting the node as part of the cloning procedure. Also, as each lexicographically significant fact in DANTE is contained in a specific structure container with corpus examples to match, there was no need to add translations to each individual example; to minimise such duplication, translation fields were automatically inserted and highlighted at the head of each structure container as a guide to the translators:

Figure 2: An example of how translation fields (highlighted in green) were automatically inserted at specified positions within the DANTE entries for the translation phase: the first sense unit of *expect*.

- **The nature of the translation task**. Given the fifty-year gap since the previous English-Irish dictionary, a significant proportion of the English content of the NEID (approximately 30%) had to be translated into Irish from scratch, and a lot of the content that could be sourced in pre-existing dictionaries was dated or even obsolete in both English and Irish in terms of language, register etc. It was consequently decided from the outset to instruct the translation team to work ex tempore and NOT to consult published sources, except for technical terms. This self-reliance allied to the requirement to record as many Irish-language equivalents as they could think of came as a culture shock to translators who normally work from reference sources using one Irish-language word or phrase to translate a given English-language word or phrase. In addition to the translation challenges, the most frequent 1,000 headwords in English (20% of the sense units in DANTE) were each translated by three translators, one from each of the main dialects, who also recorded the relevance of each translation to their own dialect. Only technical terminology was translated from existing sources, with the translators recording those sources to facilitate the decision-making process at the editing stage.
- **The layout and structure of DANTE**. To the uninitiated – and bearing in mind that the translators were primarily Irish-language specialists with no experience of lexical databases – DANTE is a

daunting beast. The hierarchical structure, consisting as it does of headwords, multiword entries, senses, structures and patterns, took some getting used to, and some of the grammar labels (or underlying structures) are less than transparent to all but the linguistically trained. The early solution here was to instruct the translators to ignore the grammar or structural information and concentrate on the corpus examples as a prompt for their translations in those cases.

5 The editing phase (began 2011, ongoing)

The initial step in the editing phase is to clone the completed translation frameworks to provide the editors with the full English-language profile plus the accompanying translations. This also ensures that the macrostructure of the edited content remains secure while the editors are drafting the final entry. Working on a sense-unit by sense-unit basis, the editors first decide on the English-language content of the entry, then the matching Irish-language content. Examples in the edited entries are included according to strict guidelines depending on the type and level of entry or sense involved, and the editors can either amend one of the corpus examples from DANTE or compose a new example entirely. Finally, the remainder of the framework is discarded once the English and Irish content of the entry has been decided.

It is also during the editing phase that new sense units are added to the dictionary database for senses not present in the original DANTE database. This may occur

- if a single sense unit in DANTE requires significantly different translation solutions in Irish (for instance *bassist* as a single sense in DANTE covers the 'double base' and 'bass guitar' but each instrument requires a different translation in Irish):

Figure 3: The entry *bassist* as one sense unit in DANTE (left) and how it was split in the editing database (right).

- if a word sense in usage in Ireland is not covered in DANTE (for instance *hurl* as verb = 'to play hurling' and as noun = 'hurling stick');
- if senses have come to the fore in the intervening years (for instance *to friend sb* in a social network context).

Also as part of the final streamlining of metalanguage etc, most of the grammar and structure labels are either removed entirely or converted to the smaller subset in use in the published dictionary; for instance, the current dictionary entries show only three verb structure labels (intransitive, transitive and modal) where the underlying DANTE database has forty-two:

Field	Examples	Range of fields present in English frameworks (DANTE)	Range of fields present in final entry
DOMAIN labels	*agri, food, ornith*	156	98
GRAM (grammar labels)	*abbrev, c_u, proper*	28	3
STRADJ (adjective structures)	*AVP_premod, that_0*	17	0
STRN (noun structures)	*AJ_pert, PP_X*	18	1
STRV (verb structures)	*AJP, Part, that_0_cond*	42	3
STYLE labels	*child, euph, pc, tech*	21	9

Table 1: Examples of the reduction in the number of fields between DANTE and the final entry.

The editors can also recommend and mark entire entries or individual sense units for omission from the final dictionary, but as this is the prerogative of the senior editing team at the next phase no entries or sense units are discarded at this stage.
Though still not at a point where it could be published, the edited database is now a much leaner version of the translation database (for instance, the word count in the 19,000 frameworks edited by April 2014 was 2.2 million compared to 8.5 million in the same translation frameworks).

6 Publication phase (began 2012, ongoing)

In this phase the final dictionary entries are arranged and prepared for uploading to the website (and in 2016 for the print version). The most obvious change is the re-ordering of all sense units by part of speech: in the DANTE database, associated senses are clustered together regardless of part of speech, but NEID follows the traditional POS-based order. This re-ordering is done automatically, reflecting the order of precedence of the sense units in DANTE, and occasionally requires minor adjustment where the order of precedence of senses under one POS may not mirror another POS. In the case of *strip*, for instance, the first sense as a verb is 'to remove clothes', but as a noun the sense 'narrow band

of sth' would be much more common than 'an instance of removing clothes', and thus was brought to the beginning of the noun section:

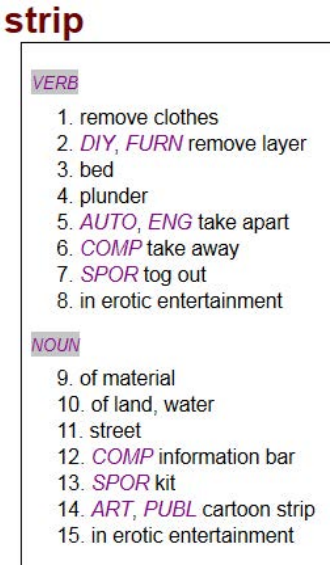


Figure 4: The manually adjusted sense order under *strip* where the noun sense associated with the first verb sense was manually transferred to a lower position in the noun order.

Another instance where the automatic re-ordering needed to be manually tweaked was *turf* as a noun, where the sense 'peat' is much more common in Ireland than the sense 'sod' and was brought to the top of the order. It is also at the publication phase that entries or senses are marked for exclusion from the published dictionary though they are not deleted from the dictionary database.

Sense units are then merged where possible in order to avoid the user having to scroll unnecessarily through numerous senses with similar translations. For instance, the Irish equivalent remains the same for seven DANTE senses of the word *studio* as 'work area', so those seven senses were amalgamated into one single sense in NEID with examples added to indicate the breadth of coverage. Similarly four sense units under *coffee* in DANTE became one in NEID with pertinent examples:

Figure 5: Multiple senses of *studio* and *coffee* merged into a single unit within the 2D database (left) and how the final entry appears in the online dictionary (right).

This approach, which would not be possible in a decoding dictionary, is facilitated by the fact that virtually all users of NEID are fully fluent in English and have an intuitive understanding of the English content.

To complete the editing process, sense disambiguators and domain labels are added or removed as required, and finally, when the text content of the entry is finalised, grammar and sound files are attached to the translation fields in preparation for publishing online.

7 Practical challenges

The project to produce a modern bilingual English-Irish dictionary faced a number of significant practical challenges. Some of these challenges may apply to similar projects in any language, some may apply in particular to other lesser-used languages, while others stemmed from the previous gap in bilingual lexicography in Irish and the consequent problems of finding suitably qualified and experienced staff at editorial and managerial level. The main such challenges were:

- **An innovative approach to dictionary compilation**. The approach followed in the NEID project is ground-breaking in that the final content of the dictionary is entirely derived from a lexical database based on systematic corpus analysis. This required all translators and editors to exercise a higher level of judgement throughout.
- **Technical challenges**. The technical working environment of modern lexicography posed a significant challenge, both at the organisational and at the individual level, where continuous training and monitoring was required.
- **Management of staff and processes**. The varied nature of the project and the overlapping of the various phases and sub-phases required careful organisation and management, particularly as a lot of the processes in the translation and editing stages were being developed from scratch.
- **Training and upskilling of staff**. The project required a significant and continuous programme of training and monitoring as the processes for each phase of the project were being tested and implemented, and this burden was exacerbated by the shortage of experienced people at all levels in the project.

8 Conclusion

Notwithstanding the challenges posed by undertaking such an innovative model for this dictionary project and the particular challenges arising in relation to Irish lexicography, the fundamental benefit to NEID is that DANTE enabled the project team to produce a dictionary which they can claim is uniquely Irish, in as much as every word of both English and Irish in the dictionary is there by the editors' choice. The customisations and additions to the DANTE database at the subsequent three stages of the compilation process of NEID show that DANTE is a very flexible resource which can be adapted to the requirements or wishes of any English-language dictionary project.

9 Further information

The Database of Analysed Text of English (DANTE): www.webdante.com
The New English-Irish Dictionary: www.focloir.ie

10 References

Atkins, B. T. S. (2010). The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In G.-M. de Schryver (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis: A Festschrift for Patrick Hanks*. Kampala: Menha Publishers (www.menhapublishers.com).

Atkins, B. T. S., Kilgarriff, A. & Rundell, M. (2010). Database of ANalysed Texts of English (DANTE): the NEID database project. In *Proceedings of the Fourteenth EURALEX International Congress.* Leeuwarden: Fryske Akademy, pp. 549-556.

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

Convery, C., Atkins, B. T. S., Kilgarriff, A., Rundell, M., Ó Mianáin, P., & Ó Raghallaigh, M. (2010). The DANTE Database (Database of ANalysed Texts of English). In *Proceedings of the Fourteenth EURALEX International Congress.* Leeuwarden: Fryske Akademy, pp. 293-295.

Convery, C., Ó Mianáin, P., & Ó Raghallaigh, M. (2010). Covering All Bases: Regional Marking of Material in the New English-Irish Dictionary. In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden, pp. 609-619.

Kilgarriff, A. (2010). DANTE: A Detailed, Accurate, Extensive, Available English Lexical Database. In *Proceedings of a meeting of the North American Association for Computational Linguistics* (NAACL-HLT), Los Angeles, June.

McCarthy, D. (2010). DANTE: a new resource for research at the syntax-semantics interface. In *Proceedings of Interdisciplinary Workshop on Verbs, Pisa*.

Ó Mianáin, P. (2013). The New English-Irish Dictionary. In Stickel, G. & Varadi, T. (eds.) *Lexical Challenges in a Multilingual Europe: Contributions to the Annual Conference 2012 of EFNIL in Budapest*. Frankfurt um Main: Peter Lang, pp. 111-114.

Rundell, M. & Atkins, B. T. S. (2011). The DANTE database: a User Guide. In *Proceedings of eLex 2011*. Trojina: Institute for Applied Slovene Studies, pp. 233-246.

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

User Support in e-Dictionaries for Complex Grammatical Structures in the Bantu Languages

Danie J. Prinsloo*, Theo J.D. Bothma*, Ulrich Heid**
*University of Pretoria, **University of Hildesheim
danie.prinsloo@up.ac.za, theo.bothma@up.ac.za, heid@uni-hildesheim.de

Abstract

This paper discusses direct user guidance as a mechanism in an e-dictionary to provide user support for complex grammatical structures in the Bantu languages. We present a design study to show that user support through direct user guidance can provide solutions in the case of complex concordial relationships between nouns and pronouns. The compilation of the complex relative construction is taken as a case in point. The concept of user support appropriately puts the user in focus. Our approach to user support also caters for the casual, on-the-fly user, who is not interested or in a position to devote time to the learning of a foreign language.
Keywords: e-Dictionaries; Bantu languages; complex grammatical structures; user support; the relative construction; Northern Sotho

1 Introduction

In the Bantu languages, there are many grammatical constructions that are insufficiently treated in current dictionaries because of the complexity of the constructions. Other solutions need to be designed as an integral part of a dictionary and additional levels of user support are required within the dictionary. Such support should be available to the dictionary user "on demand", and different options can be available for a specific information need. The nature of the support could typically also link to a user's level of knowledge of the grammatical system of the language. A user with a very limited knowledge of the language or a casual user, for example, may prefer a machine translation option in the dictionary, with links to the grammar rules which may be consulted on demand. On the other hand, a user who has a fair knowledge of the language may require a different type of support, e.g. through *inter alia* decision trees, structured paths or direct user guidance. Such technologies, integrated in the dictionary, may enable the user to find the correct information at exactly the right level of detail and complexity (s)he requires to solve his/her information need (cf. Bothma 2011).

The innovative use of decision trees and structured paths as tools to support users have been dealt with in some detail in Prinsloo et al. (2011, 2012) and it has been shown that these solutions can provide significant decision support to users for complex text production situations such as copulative constructions and kinship terminology in Northern Sotho. The purpose of such tools is to guide users

to the information they are looking for, i.e. without having to first study complicated grammatical structures in order to find the required information. This guidance process is done through decision trees (i.e. a series of basic choices made by the user) or through structured paths (e.g. visually linking kinship terms in a schematic illustration of a family tree) as discussed in Prinsloo et al. (2011, 2012) or through direct user guidance, as discussed in this paper.

Direct user guidance as an additional technique in the dictionary to provide user support for complex grammatical structures in the Bantu languages is not a solution for all user support. We regard it as a complementary technology that may be used in conjunction with other user support technologies for specific grammatical constructions, available to the user on demand, depending on the user's level of language knowledge, the nature of the information need and choice of support tool. We present a design study to show that user support through direct user guidance can provide solutions in the case of complex concordial relationships between nouns and pronouns. In terms of the Function Theory of Lexicography (Tarp 2008, Bothma and Tarp 2012) the design provides for text production, text reception and cognitive information needs. No case studies or user evaluation of these techniques has been done to date, as we feel that it is important to first define a range of techniques and the range of complex grammatical structures where such techniques could offer relevant user support before any serious implementation in actual real world scenarios would be warranted. This does not mean that small scale prototypes of individual techniques should not be developed to establish the technological feasibility of such techniques. However, to do proper usability studies on such prototypes that are not fully integrated into a full dictionary will have only limited value, as it will not be possible to determine whether (or to what extent) users would use such techniques in real world situations as an integral part of dictionary use. As will be clear from the discussion below and from Prinsloo et al. (2011, 2012), such techniques are made available "on demand", i.e., users are not forced to use them if they feel that their information needs have been solved by the "standard" dictionary article. In every case, the use of such a technique is therefore a conscious choice of the user to find more information or information that is easier to use / digest / apply than the information available in the dictionary, the outer text of the dictionary or other reference tools such as grammar books that the user may have available.

The importance of the user perspective as the main thrust in the compilation of modern dictionaries has been emphasized in numerous publications, e.g., Gouws and Prinsloo (2005), Tarp (2008, 2011, 2012). The concept of user support appropriately puts the user in focus. Compare Tarp's (2012:253) idea of individualization when he refers to "quicker, more accurate and personalized satisfaction of the corresponding user needs". Our approach to user support furthermore does not necessarily put the user into a specific category (e.g., as a learner of the language): it is not profile-based and does not assume that the user will be interested to study a complete grammatical paradigm before being able to produce (or understand) text and speech. We therefore also cater for the casual, on-the-fly user, who is not interested or in a position to devote time to the learning of a foreign language.

2 Phenomena and Data: Grammatical Distinctions as a Problem for Bantu Lexicographers

2.1 The Notion of Grammatical Distinctions

Due to the richness of grammatical distinctions, a given grammatical property may be expressed in many different forms. For example, there are different equivalents for a pronoun such as he in Bantu, determined by the grammatical class of the noun. Nouns in Bantu languages are subdivided into different noun classes and these classes have their own sets of, e.g., subject concords and object concords, as well as different sets of pronouns such as demonstrative, possessive, emphatic and quantitative. This means that in Northern Sotho a basic English pronoun such as he can be expressed by up to ten different subject concords, a form like him by ten object concords and more than 20 pronominal forms. Consider table 1 which distinguishes 15 different noun classes each having their own subject concords (Sc.); object concords (Oc.); demonstratives (Dem.); possessive concords (Poss.); emphatic pronouns (Ep.) and quantitative pronouns (Qp.).

Person or noun class	Example	Sc.	Oc.	Dem.	Poss.	Ep.	Qp.
1st Person singular	nna 'I'	ke	n-				
1st Person plural	rena 'we'	re	re				
2nd Person sing.	wena 'you' (singular)	o	go				
2nd Person plural	lena 'you' (plural)	le	le				
Class 1	monna 'man'	o/a	mo	yo	wa	yena	yohle
Class 2	banna 'men'	ba	ba	ba	ba	bona	bohle
Class 3	molato 'trouble, problem'	o	o	wo	wa	wona	wohle
Class 4	melato 'problems'	e	e	ye	ya	yona	yohle
Class 5	lesogana 'young man'	le	le	le	la	lona	lohle
Class 6	masogana 'young men'	a	a	a	a	ona	ohle
Class 7	selo 'object, thing'	se	se	se	sa	sona	sohle
Class 8	dilo 'objects, things'	di	di	tše	tša	tšona	tšohle
Class 9	ntlo 'hut'	e	e	ye	ya	yona	yohle
Class 10	dintlo 'huts'	di	di	tše	tša	tšona	tšohle
Class 14	bogobe 'porridge'	bo	bo	bjo	bja	bjona	bjohle
Class 15	go reka 'to buy'	go	go		ga		
Class 16	fase 'below'			fa			
Class 17	godimo 'above'	go	go		ga	gona	gohle
Class 18	morago 'behind'			mo			

Table 1: The noun class system of Northern Sotho with a few sets of concords and pronouns.

In table 1 the demonstrative 'this' varies depending on the class of the noun, e.g., class 1: *monna* ***yo*** 'this man' but class 14: *bogobe* ***bjo*** 'this porridge'. Likewise, the possessive 'of' differs for each class, e.g., class 1: *mosadi* ***wa*** *monna* 'wife of the man' but class 2: *basadi* ***ba*** *monna* 'wives of the man'. Concords and pronouns representing subjects and objects also vary according to the nominal class, e.g.:

(1) O e bone 'He saw it'

o (e.g. *monna* class 1)	*e* (e.g. *tau* class 9)	*bone*
he (the man)	*it* (the lion)	saw

2.2 Grammatical Distinctions in the Sentence Context

In table 1 the grammatical distinctions paradigm is mono-dimensional in the sense that it is always given for a single source language item which diverges into a single set of equivalents. More than one instance of grammatical distinction can, however, co-occur in a single construction or phrase:

A single occurrence

Example 1: *he*, as the subject of a sentence (subject concords):

(2) *O/a/le/se/e* thušitše mosadi.
He helped the woman

Example 2: how to express *all* (quantitative pronouns):

(3)Go bolaya *bohle/yohle/ohle/tšohle*
To kill *all*

Two occurrences: he as a subject and them as an object:

What is at stake here is direct guidance in terms of the simultaneous handling of subject and object concords:

(4) *O/a/le/se/e* tlo *ba/e/a/di* thuša
He will them help
He will help *them.*

Three occurrences (the verbal relative): he as a subject, and as a demonstrative and them as an object:

(5) *Yo/wo/le/se/ye* *a/wo/le/se/e* *ba/e/a/di* thušitšego
He he them helped
He who helped *them*.

3 Direct Guidance for Concords and Pronouns

Guidance is given by means of three possible access points depending on the user's need in terms of text production (access points 1 and 2) or text reception (access point 3), and his/her knowledge of the language:

- **Access point 1:** Step by step guidance: build your own Northern Sotho sentence/construction.
- **Access point 2:** The user enters an English phrase and the system then assists him/her in a step-by-step build-up process of the Northern Sotho construction.
- **Access point 3:** The user enters a Northern Sotho phrase and the software analyses it.

Utilising *Access point 1* simply requires the user to enter the Northern Sotho noun and the software will suggest the correct pronoun and subject/object concord from table 1. Where more than one option is applicable, the user has to select the correct one or utilise clickable help functions to guide him/her to the correct option or (s)he can revert to the *Access point 2* option.

Taking *Access point 2* as departure, the user can type in "the man is walking" and the software will return the noun (*monna*) + the subject concord (*o*) + the present tense marker (*a*) + the verb *sepela*, guiding the user to build *monna o a sepela*. Clickable help functions and 'more information' boxes are also provided.

Entering a Northern Sotho sentence from Access point 3 will result in the reverse process, e.g. *monna* 'man'(noun) + *o* 'he' (subject concord) + *a* (present tense marker) + *sepela* 'walk' (verb).

The full set of necessary data, for these cases, is thus as follows: In the case of single and two occurrences given in 2.2, subject and object pronouns/concords are independent from each other, and their choice is only conditioned by the noun class of their antecedents.

To be able to provide the above mentioned kinds of guidance for single and two-occurrence of grammatical distinctions a word (token) list tagged for part of speech for nouns, verbs, subject concords, object concords and pronouns, and a basic bilingual dictionary for word forms are required.

To be able to provide guidance for cases like example (5) more than just word lists are needed as an internal knowledge source for the guidance tool: the agreement between subject and relative demonstrative must be encoded as well.

To guide users in the creation of the relative construction, also the morpho-syntactic structure of this construction must be explained.

The verbal relative case (three occurrences) is different in so far as the relative demonstrative ("who") is in grammatical agreement with the subject. In addition to the requirements for single and two-occurrences cases, a basic five-element formation rule for the verbal relative: noun + demonstrative + subject concord + verb + relative suffix (-go) is required.

4 Example of Direct Guidance for the Verbal Relative

4.1 User Support for Text Production

Access point 1: This provides step by step guidance on how to build your own relative construction. The user with a basic knowledge of the grammatical system would like to express "the man who loves her". (S)he knows the different nouns and verbs in Northern Sotho but needs guidance in terms of the concordial system. In this case the user consults the article for "who" in an English to Northern Sotho dictionary, selects the button "Build your own relative construction" and types the Northern Sotho word for "man" = *monna*.

The system subsequently suggests the relevant concords from table 1. In the consultation, the relevant section for the appropriate noun class is highlighted while being placed in context within the grammar table. For cognitive use, clickable options to see larger portions of the table are provided; cf. table 2.

Person or noun class	Example	Dem.	Sc.		Oc.
		This	He/she		Him/her
Class 1	monna *'man'*	yo	o	a	mo
Class 2	banna 'men'	ba	ba	ba	ba
Class 3	molato 'trouble'	wo	o	o	o
Class 4	melato 'problems'	ye	e	e	e
Class 5	lesogana 'young man'	le	le	le	le
Class ...					
		Click to add	Click to add	Click to add	Click to add
		More information	More information	More information	More information
Click for full list	Click for full list	Click for full list	Click for full list		Click for full list

Table 2: The noun class 1 of Northern Sotho highlighted for selection of the correct concords.

Based on his/her knowledge of the relative construction (noun + demonstrative + subject concord + verb + relative suffix (-go)), the user can now build the full relative construction, to arrive at the full phrase *monna yo a mo ratago.* If the user is knowledgeable about the subject concords of class 1 in the relative, (s)he will click the subject concord *a* directly. If the user nevertheless needs more support, e.g., choosing between the subject concords *o* or *a*, the cursor could be momentarily rested on any of the "more information" boxes in the table triggering a pop-up box to support him or her in the selection task.

The user who needs full support can type a complete relative phrase in either English or Northern Sotho as portrayed in access points 2 and 3.

Access point 2: The user enters an English phrase: "The man who loves her", similarly to access point 1 in the dictionary. The system then assists the user in a step-by-step build-up process of the relative construction:

(i) the man: the tool provides the correct equivalent from the dictionary, i.e. monna tagged for part of speech as N01 (noun of class 1, cf. table 1);

(ii) who: keeping the agreement constraint from the sentence formation rule (noun + demonstrative + subject concord, + verb + relative suffix (-go) in 4.1), the tool extracts the demonstrative for class 1 from the closed-class list of demonstratives, i.e. yo;

(iii) (subject concord): The insertion of the SC is coded in the rule for relatives: it requires, in addition to the demonstrative in (ii), the subject concord for the noun in (i). As in (ii), the tool proposes a, i.e., the subject concord for class 1 which is appropriate, among others, for the relative as opposed to o.

(iv) her: unspecified, as there is no unique referent e.g., the woman. On the basis of corpus frequency the software suggests the top five ranked possible options for the object concord, i.e. classes 1, 9, 3, 7, 5 (calculated from the Pretoria Sepedi Corpus, PSC). For human nouns, class 1 stands out in terms of frequency, and the selection is therefore for class 1 = mo. If, however, the relevant word is, e.g., from class 7, the user can type the word or click on the full list of object concords and selects the concord se for class 7.

(v) loves: as for (i), the task is only to find the correct Northern Sotho equivalent: rata, plus adding the relative suffix -go which is, as for (iii), built into the relative construction rule.

In the consultation, the relevant section for the appropriate noun class is highlighted while being placed in context within the grammar table. For cognitive use clickable options to see larger portions of the table are provided, as shown in table 2. The construction rules ensure contextually appropriate highlighting, e.g., only of the subject concord, in step (iii) or of the object concord in step (iv).

4.2 User Support for Text Reception

Access point 3: A Northern Sotho phrase: *monna yo a mo ratago* (or part thereof, e.g., *ratago*, *mo ratago*, *yo a mo*, etc.). The software analyses the phrase in terms of the formation rule for the relative (automatically or user selected), in fact reversing the strategy explained with respect to access point 2 for "the man who loves her".

5 Conclusion and Future Work

User support through direct guidance (and other support mechanisms) for complex grammatical structures allows the user to navigate via the shortest route to the information (s)he is looking for in an dictionary without having to work through long and often complicated grammar-type representations of complex grammatical structures. Such guidance is always available on demand, i.e., the user is not forced to work through any such support mechanisms if (s)he finds that the "standard" data in the dictionary are sufficient to solve his/her information need in a given situation. However, if more information is needed or if the standard presentation of the information (be this in the dictionary, in outer texts or in reference tools) is too difficult or complex to be easily understood, the user would have an alternative mechanism (or alternative mechanisms) to obtain the relevant information. It also successfully combats information overload and fulfils the needs of not only the learner of the language but also of the casual on-the-fly-user of the language; its flexibility is intended to provide a step towards individualization.

Different access points are available to the user depending on his/her pre-existing knowledge. It is not a profile-based dictionary. We envisage that such mechanisms be implemented as "plug-in modules" in entries of specific lemmas of the dictionary, i.e., an additional link is shown to the user on screen which (s)he can follow on demand. Since such modules can exist independently from the dictionary database, it would be feasible to reuse them in other environments as well. It would therefore be feasible to use such tools as writing tools integrated in a word processor, again activated by the user on demand, if (s)he requires to check the correct formulation of a complex grammatical construction, similar to spelling and grammar checkers that currently occur in popular word processing software.

Future work includes the development of a working prototype and possibly the full-scale implementation of user support for complex structures proposed in this paper as a module of electronic dictionaries. Identifying and categorising additional support techniques and developing prototypes and the full-scale implementation of such additional support techniques are also envisaged, as well as identifying further complex grammatical structures for which additional user support techniques may need to be developed. We will also investigate the possibility of the reuse of all such modules in writing tools for user support.

6 References

Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In: Fuertes-Olivera, PA and Bergenholtz, H (Eds.). 2011. *e- Lexicography: The Internet, Digital Initiatives and Lexicography.* London & New York: Continuum. Pp. 71-102.

Bothma, T.J.D. & Tarp, S. (2012). Lexicography and the relevance criterion. *Lexikos* 22: 86–108.

Gouws, R.H. and Prinsloo, D.J. (2005). *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.

Prinsloo, D.J., Heid, U., Bothma, T.J.D. & Faaß, G. (2011). Interactive, dynamic electronic dictionaries for text production. In Kosem, I. and Kosem, K. 2011. *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011*. Bled, 10-12 November 2011. Bled: Trojína. Pp. 215-220. http://www.trojina.si/elex2011/elex2011_proceedings.pdf.

Prinsloo, D.J., Heid, U., Bothma, T.J.D. & Faaß, G. (2012). Interactive, dynamic electronic dictionaries for text production. *Lexikos* 22: 290-320. http://lexikos.journals.ac.za/pub/article/view/1009/526.

PSC, Pretoria Sepedi Corpus. University of Pretoria.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learner's lexicography.* (Lexicographica Series Maior 134). Tübingen: Niemeyer.

Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In: Fuertes-Olivera, P.A. & Bergenholtz, H. (Eds.) 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography.* London & New York: Continuum. Pp. 54-70.

Tarp, S. (2012). Online dictionaries: today and tomorrow. In: *Lexicographica 28.* 253-267.

Acknowledgement

This research is conducted within the SeLA project (Scientific e-Lexicography for Africa), supported by a grant from the German Ministry for Education and Research, BMBF, administered by the DAAD.

Concerning the Treatment of Co-existent Synonyms in Estonian Monolingual and Bilingual Dictionaries

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

Abstract

Estonian is a lesser used language of Europe (with about one million users) that belongs to the Finnic group of Finno-Ugric language. It has a large proportion of international words, which has resulted in extensive co-existent synonymy between native and international words. However, because of linguistic purism Estonian monolingual dictionaries do not treat the members of such synonym pairs on an equal footing. Usually, they give preference to native words, which in practice implies that international words are provided together with their native-language equivalents, but native words are without any reference to international words. Unfortunately, this kind of asymmetrical treatment is not helpful for language users and compilers of bilingual dictionaries; they need to develop synonym competence, which covers both international and native synonyms. Probably the best solution is symmetrical treatment, that is, both synonyms should be provided side by side on a regular basis, which contributes to their better knowledge. It is argued that the quality of bilingual dictionaries could be significantly improved by providing the synonym pairs together on a systematic basis.
Keywords: synonyms; monolingual dictionaries; bilingual dictionaries; language planning; English; Estonian

1 Introduction

Synonymy can be regarded as "a relation of similarity / identity of meaning between senses associated with two (or more) lexical forms" (Cruse 2002: 486). For practical purposes, it is useful to follow the "synonymy rule of thumb: X is Y and Y is X" (Atkins and Rundell 2008: 135). It is also important to bear in mind that synonymy is a language-specific phenomenon (Gouws 2013: 349).

The purpose of the present study is to examine the lexicographic practice of treating co-existent intralingual synonyms in Estonian monolingual dictionaries from the perspective of its usefulness for lexicographers who compile English-Estonian and Estonian-English bilingual dictionaries. Estonian is a lesser used language of Europe (with about one million users) that belongs to the Finnic group of Finno-Ugric languages. Estonian has a large proportion of international words; therefore, it is not surprising that the size of the latest edition of the dictionary of foreign words (VSL 2012) almost equals the size of the single-volume dictionary of Standard Estonian (ÕS 2013). In many cases synonymy

concerns pairs of synonyms between native (Estonian) and foreign (international) words. However, sometimes both synonyms can be foreign words (e.g. *kabatšokk, tsukiini* 'courgette Br, zucchini Am).
Extensive co-existent intralingual synonymy in Estonian is the result of conscious efforts of language planners to increase the proportion of native vocabulary. It is typical of Estonian specialized vocabulary but can be found in general language too. The following examples come from linguistic terminology: *nimetav, nominatiiv* 'nominative (case)', *omastav, genitiiv* 'genitive case'; *osastav, partitiiv* 'partitive (case)'; *täishäälik, vokaal* 'vowel'; *kaashäälik, konsonant* 'consonant'; *vormiõpetus, morfoloogia* 'morphology'; *lauseõpetus, süntaks* 'syntax'.
Stephen Ullmann discussed similar synonymy in German:
The German linguist, for instance, can choose between *Lautlehre* and *Phonetik*, *Formenlehre* and *Mophologie*, *Bedeutungslehre* and *Semantik* (or *Semasiologie*), and as these synonyms are used in the same contexts, and sometimes even in the title of the same book, one can hardly speak even of stylistic differences between them. (Ullmann 1962: 142)

> It could well be that Estonian may have followed the German example. Historically one could explain the development of extensive native-language terminology with national pride and identity building. In the case of Dutch, for example, Roel Vismans claims that "in the early Republic pride in the Dutch language also led to the introduction of many newly coined words for concepts expressed by Latinisms in other languages" (Vismans 1998: 133).

Estonian language planners have regarded the introduction of co-existent synonymy between international and native words as an important method of establishing native-language identity and increasing the proportion of native vocabulary. In 1980 Henn Saari, an Estonian linguist, published a seminal paper where he discussed at length the importance of such word pairs in Estonian terminology (Saari 1980)
Uno Mereste, another Estonian academic whose views have been influential in language planning, explained the need for intralingual synonymy as follows:

> There is no such thing as positive polysemy. However, in addition to undesirable synonymy, there is also positive sameness of meaning or synonymy, which one should try to preserve and develop further – it is synonymy between pairs of native and international terms, for example, *nimetav* or *nominative* 'nominative (case)', *ainsus* or *singular* 'the singular', *tasuvus* or *rentaablus* 'profitability', *tehis-* or *sünteetiline aine* 'synthetic substance', etc.. (Mereste 2000: 87)

In Estonia, one of the purposes of national word coining competitions is to seek native synonyms for the existing international words. A recent example is the winning entry *taristu* of the 2010 competition, which means the same as *infrastruktuur* 'infrastructure'. The word was immediately adopted by Estonians and is now widely used.
On the one hand, there is good reason to regard the co-existent intralingual synonymy in Estonian as a luxury because it contradicts the principle economy in language. D. Alan Cruse has claimed that

"there is very little semiotic motivation for such a state of affairs: the only possible utility for absolute synonyms is aesthetic, to avoid repetition of forms (2002: 488).

2 Treatment of Intralingual Synonyms in Monolingual Dictionaries

In order to understand the treatment of intralingual synonyms in Estonian bilingual dictionaries, one has to examine the practice of presenting intralingual synonyms in monolingual dictionaries. It appears that intralingual synonyms do not receive equal treatment in monolingual dictionaries. They tend to prefer native-language synonyms and provide the native-language synonym in the entry for the international word but not the other way round (see the treatment of the synonyms *regionaalne, piirkondlik* 'regional' below; EKSS stands for *Eesti keele seletav sõnaraamat* 'Explanatory Dictionary of Estonian').

(1) **regionaalne** adj mingit regiooni hõlmav, selles esinev v sellele isloomulik, piirkondlik (EKSS)
(2) **piirkondlik** adj piirkonda hõlmav, selles toimuv v sellele iseloomulik, piirkonna- (EKSS)

This practice of asymmetrical treatment is intended to discourage the use of international words. It is difficult to say whether it serves its purpose, but one could likewise regard it as not user-friendly. Ordinary dictionary users, as well as compilers of bilingual dictionaries, need to develop competence with regard to the knowledge of both synonyms. In some cases, a prescriptive monolingual dictionary, such as ÕS 2013 (*Õigekeelsussõnaraamat 2013* 'Dictionary of Correct Usage 2013'), may even indicate that the use of a foreign synonym is undesirable. Such is the case with *multikultuurne* and *mitmekultuuriline* 'multicultural'.

(3) [**multikultuurne] →** mitmekultuuriline (ÕS 2013)
(4) **mitmekultuuriline**. Kanada on mitmekultuuriline maa (ÕS 2013)

This dictionary uses brackets for undesirable lexemes; in this case language planners regard the use of the combining form *multi-* undesirable and recommend its native equivalent *mitme-* 'several' instead. On the other hand, the following example shows that there can be inconsistencies even in dictionaries published by the same institutions. The new edition of the dictionary of foreign words VSL (*Võõrsõnade leksikon* 'Dictionary of Foreign Words') lists two forms of the 'undesirable' item, fails to provide the native synonym (*mitmekultuuriline*) recommended by the dictionary of correct usage, and suggests a different synonym (*paljukultuuriline*).

(5) **multikultuurne**, **multikultuuriline** mitut kultuuri sisaldav, paljukultuuriline (VSL)

The next example shows that a dictionary user needs to develop a critical attitude towards the data provided in monolingual dictionaries and should not take everything at face value. The example concerns the notion of *kabatšokk, tsukiini* 'courgette Br, zucchini Am'. A similar example of *aubergine* Br, *eggplant* Am and its Polish equivalents was discussed by Adamska-Sałaciak (2013: 338–339). Such lexical differences between British and American English are regarded as cross-varietal synonyms. The present co-existence of *kabatšokk* and *tsukiini* in Estonian can be explained by the fact that before 1991 the vegetable was known in Estonia by its Russian name (which ultimately has a Tatar origin). In recent years, however, *tsukiini* has been adopted as well, and at present both terms are used. It remains to be seen whether the younger generation will continue to use *kabatšokk*, or it will gradually be replaced by *tsukiini*.

However, the analysis of the treatment of *kabatšokk* and *tsukiini* in dictionaries revealed several problems.

(6) **kabatšokk** *s* AIAND kõrvitsa pikerguste viljadega teisend, suvikõrvits, tsukiini (EKSS)

(7) **tsukiini** *s* kabatšokk, suvikõrvits (EKSS)

(8) **suvikõrvits** suvel v varasügisel kasutatavate kõrvitsate üldnimetus. *Kabatšokk, taldrikkõrvits ja spagetikõrvits on suvikõrvitsad* ‖ kabatšokk, tsukiini. *Täidetud suvikõrvitsad* (EKSS)

The problem with the treatment of *kabatšokk*, *tsukiini*, and *suvikõrvits* 'lit. summer pumpkin' in EKSS is that *suvikõrvits* is a broader term and should be regarded as a superordinate of the first two terms. However, it is now time to turn to examine how the data provided in monolingual dictionaries tends influences the treatment of co-existent synonyms in bilingual dictionaries.

The treatment of the synonyms *kabatšokk* and *tsukiini* in the dictionary of foreign words showed remarkable differences.

(9) **kabatšokk** *s* BOT suvikõrvits, melonkõrvits, puhmikkõrvits, hariliku kõrvitsa piklike viljadega teisend (VSL)

(10) **tsukiini** *s* BOT kabatšokk, hariliku kõrvitsa teisend (*Cucurbita pepo)* (VSL)

While the entry *tsukiini* claims that *tsukiini* is *kabatšokk*, the same is not true of the entry *kabatšokk*, which provided a number of other intralingual equivalents. It seems that different lexicographers were responsible for contributing material to different letters and no systematic harmonization was carried out.

3 Treatment of intralingual Synonyms in Bilingual Dictionaries

Analysis of the material of the Estonian bilingual dictionaries shows that more accuracy and systematicity is needed in the treatment of co-existent intralingual synonyms. At first the coverage of *courgette* and *zucchini* will be examined in three dictionaries. The English-Estonian Dictionary by Johannes Silvet (Silvet 4) is the best-known dictionary of this category; its first edition was published in the

late 1930s; at present its fourth enlarged edition of 2002 is widely used. The Estonian-English Dictionary by the TEA publishers (TEA 2005) is a recent Estonian-English dictionary. The Contemporary Estonian-English Dictionary by Andres Aule (see Aule 2003) is a large-scale Estonian-English dictionary; so far two volumes have been published (the letters A–J and K). So far the Estonian-English dictionary by Paul Saagpakk is the largest as to its coverage; unfortunately, it includes a considerable amount of outdated material while many more recent words are absent.

(11) **courgette** (no entry in Silvet 4)

(12) **zucchini** *s bot* kabatšokk, suvikõrvits, tsukiini (Silvet 4)

(13) **tsukiini** (no entry in TEA 2005)

(14) **kabatšokk** zucchini <pl zucchini, zucchinis>, (it), squash <pl squash, squashes>, (vegetable) marrow (TEA 2005)

(15) **kabatšokk** [melonkõrvits] vegetable marrow, marrow (*üldk*) (Aule)

(16) **kabatšokk** (no entry in Saagpakk)

(17) **tsukiini** (no entry in Saagpakk)

The presented examples reveal a considerable degree of inconsistency in the treatment of co-existent synonyms; nor are the suggested translation equivalents always accurate. In fact, the analysis shows that a bilingual lexicographer needs to critically study the material provided in monolingual dictionaries. However, in addition, primary corpus data is badly needed during the preparation of bilingual dictionaries.

Another area of intralingual synonymy concerns the situation where the international word belongs to a specific domain and the native synonym can be regarded as belonging to general language. In such cases the synonym that belongs to a specific domain should be labelled accordingly (e.g. *kardioloog med* 'cardiologist' and *südamearst* 'heart doctor').

(18) **kardioloog** *med* kardioloogia eriteadlane, südamearst (VSL)

However, analysis of a number of monolingual and bilingual dictionaries showed surprisingly that VSL was the only dictionary that provided the native synonym *südamearst*, which, in fact, is an everyday word in Estonian. It was also surprising that none of the other monolingual dictionaries (EKSS and ÕS 2013) listed the entry *südamearst*. Nor could it be found in the studied bilingual dictionaries. This finding suggests once again that a bilingual lexicographer should not rely too much on the existing monolingual dictionaries and has to use primary data as well.

What could be a solution to this problem? The answer is consistent symmetrical treatment of co-existent synonyms. Are there any Estonian dictionaries that treat the co-existent native and foreign synonyms on an equal footing? One such dictionary is the Estonian-English Dictionary of Linguistics by Erelt et al. (2012) (*Eesti-inglise keeleteaduse sõnastik*, EIKS). In this dictionary, synonyms are provided after the sense, separated by a comma. The treatment is consistent in that the same

information is provided in two places. This approach is also helpful for students who can see co-existent native and international terms side by side, which contributes to their synonym competence.

(19) **kaasa võnkuma, resoneerima** resonate (EIKS)

(20) **resoneerima, kaasa võnkuma** resonate (EIKS)

(21) **parasiitsõna, nugisõna** parasitic word, overused word (EIKS)

(22) **nugisõna, parasiitsõna** parasitic word, overused word (EIKS)

4 Conclusion

The study of co-existent intralingual synonyms in Estonian showed that monolingual dictionaries tend to promote native synonyms and do not cross-reference international words. While the rationale of the language planners is understandable, the disadvantage of this approach is asymmetry in the treatment of native and international synonyms. However, language users, as well as compilers of bilingual dictionaries, need to be familiar with the entire repertoire of such synonyms. For this reason, the symmetrical approach where the co-existent synonyms are provided side by side on an equal footing is fully justified. It is argued that the quality of bilingual dictionaries could be significantly improved by providing synonym pairs together on a systematic basis.

5 References

Adamska-Sałaciak, A. (2013). Equivalence, synonymy, and sameness of meaning in a bilingual dictionary. In *International Journal of Lexicography*, 26(3), pp. 329–345.

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

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

Cruse, D.A. (2002). Synonymy. In D.A.Cruse et al. (eds.) Lexikologie. Lexicology. Berlin: Walter de Gruyter, pp. 485–497.

EIKS = Erelt, M., Erelt, T., Veldi, E. (2012). *Eesti-inglise keeleteaduse sõnastik. Teine, täiendatud ja parandatud trükk*. Tallinn: Eesti Keele Sihtasutus [Estonian-English Dictionary of Linguistics, second enlarged and revised edition].

EKSS = *Eesti keele seletav sõnaraamat*. Accessed at: http://www.eki.ee/dict/ekss [10/11/2013] [Explanatory Dictionary of Estonian].

Gouws, R.H. (2013). Contextual and co-textual guidance regarding synonyms in general bilingual dictionaries. In *International Journal of Lexicography,* 26(3), pp. 346–361.

Mereste, U. (2000). Oskuskeele korraldamise mõningaid põhimõtteid (Teesid). In U. Mereste. Oskuskeel ja seaduse keeleline rüü. Tallinn: Eesti Keele Sihtasutus, pp. 84–88 [Some principles of LSP planning, Abstract].

Saagpakk = Saagpakk, P. (2000). *Estonian-English Dictionary*. Tallinn: Koolibri.

Saari, H. (1980). Omasõna ja võõrsõna paarid eesti oskussõnavaras. In *Keel ja Kirjandus*, 11, pp. 654–666, 12, pp. 737–743 [Pairs of native and foreign words in Estonian specialized vocabulary].

Silvet 4 = Silvet, J. (2002). *English-Estonian Dictionary*. 4th enlarged and revised edition. Tallinn: TEA.
TEA 2005 = Mägi, R. (ed.) (2005). *Estonian-English Dictionary*. First edition. Tallinn: TEA.
Ullmann, S. (1962). Semantics. An Introduction to the Science of Meaning. New York: Barnes & Noble.
Vismans, R. (1998). Dutch. In G.. Price (ed.) *Encyclopedia of the Languages of Europe*. Oxford: Blackwell, pp. 129–136.
VSL = Paet, T. (ed.) (2012). *Võõrsõnade leksikon*. 8., põhjalikult umber töötatud trükk. Tallinn: Eesti Keele Instituut & Valgus [Dictionary of Foreign Words, 8th thoroughly revised edition].
ÕS 2013 = Raadik, M. (ed.) (2013). *Eesti õigekeelsussõnaraamat. ÕS 2013*. Tallinn: Eesti Keele Sihtasutus [Dictionary of Correct Usage of Estonian].