Electronic Lexicography in the 21st Century

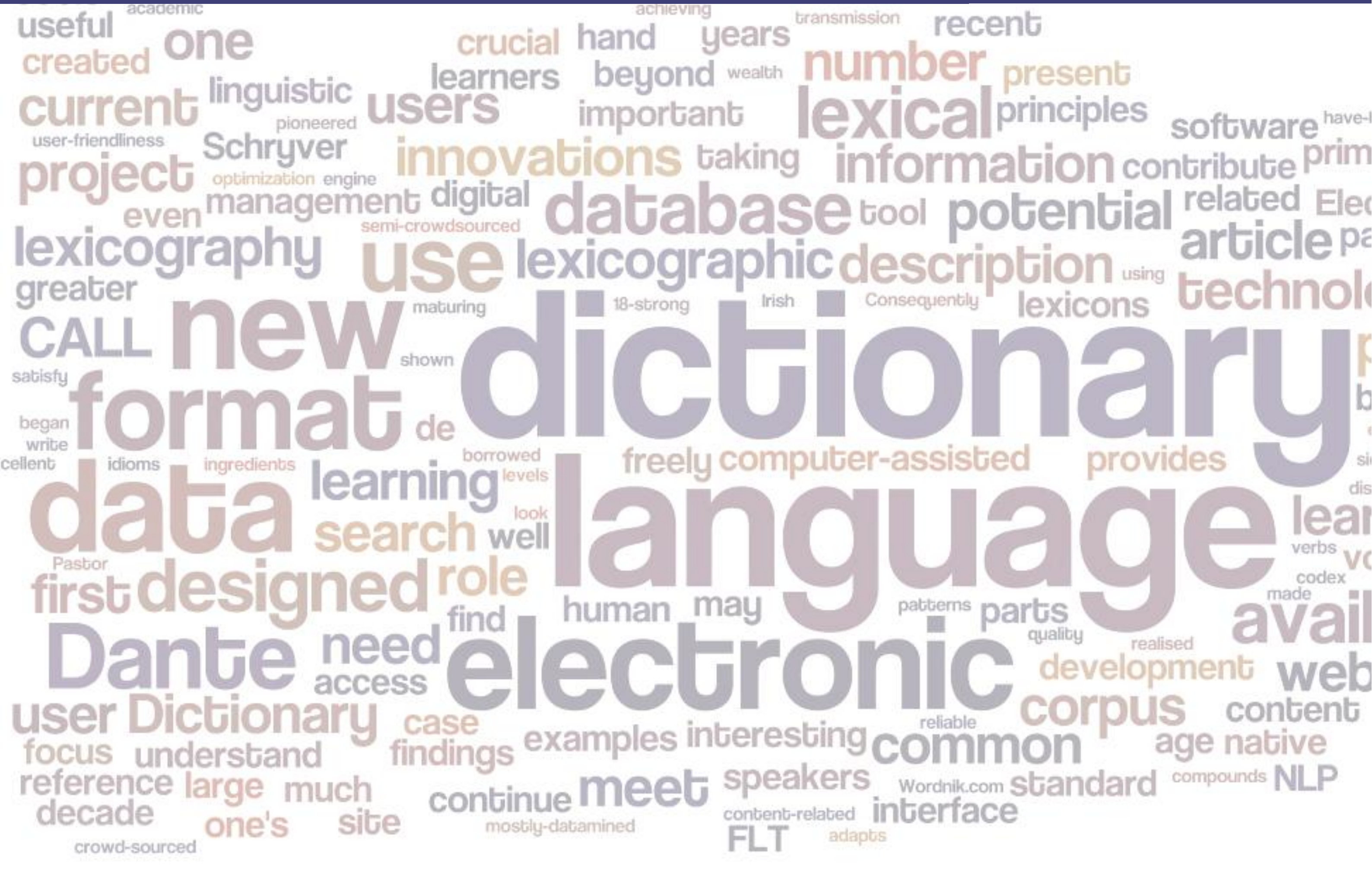
New Applications for New Users

Proceedings of eLex 2011, Bled, 10-12 November 2011

Edited by

Iztok Kosem and Karmen Kosem

trojína
Institute
for Applied Slovene Studies



Electronic lexicography in the 21st century:

New applications for new users

Proceedings of eLex 2011

Edited by

Iztok Kosem and Karmen Kosem

http://www.trojina.si/elex2011

10-12 November 2011

Bled, Slovenia

Electronic lexicography in the 21st century:
New applications for new users

Proceedings of eLex 2011, Bled, 10-12 November 2011

Edited by Iztok Kosem, Karmen Kosem

Front cover designed by Karmen Kosem

Published by Trojina, Institute for Applied Slovene Studies

© Trojina, Institute for Applied Slovene Studies

Ljubljana, November 2011

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

81'374:004.9(082)(0.034.2)

Electronic lexicography in the 21st century [Elektronski vir]: new applications for new users : proceedings of eLex 2011, 10-12 November 2011, Bled, Slovenia / edited by Iztok Kosem and Karmen Kosem. - El. knjiga. - Ljubljana : Trojina, Institute for Applied Slovene Studies, 2011

Način dostopa (URL):
http://www.trojina.si/elex2011/elex2011_proceedings.pdf

ISBN 978-961-92983-3-6 (pdf)
1. Kosem, Iztok

258763008

Acknowledgements

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

Academic partners

EURALEX

META-NET

Main sponsors

iLEX

Integrated Lexicography, Editing and Xml

erlandsen media publishing

Design
Edit
Publish

Lexical 文 Computing

ABBYY®

Supporting sponsors

KDICTIONARIES

OXFORD
UNIVERSITY PRESS

htc
quietly brilliant

HRUSKA.SI
SOČNE REŠITVE

FOREWORD

Continuous technological progress has had a considerable impact on dictionary-making, both on how dictionaries are made, and the format(s) in which they are presented to the user. There has been an ongoing discussion on whether electronic dictionaries will slowly replace paper dictionaries, and in many parts of the world, this has already taken place. It is now the time of online dictionaries and dictionary apps – the users want to click, tap, slide, etc. But dictionary users that use these new technologies are putting new demands on dictionary-makers. It is now often expected that dictionaries are free, offer quick access to all the types of information in them, contain every single word in existence, and include other types of features, e.g. grammar, games, blogs and forums. The question is thus no longer about electronic format competing with the paper format, but more about how to utilize the many advantages of electronic medium to make dictionaries as user friendly as possible.

There is another group of users that have been affected by technological progress, namely lexicographers themselves. As corpora get larger and larger, there is more and more data to analyse. Also, the existence of different dictionary formats means that the needs of different types of users have to be met. It is thus essential that the lexicographers are provided with tools that speed up their work, and automate the procedures that do require little human intervention.

The papers found in these proceedings from the eLex 2011 conference on electronic lexicography, which took place between November 10th and 12th in Bled, Slovenia, contain the reports on electronic dictionaries or ongoing lexicographic projects that seek to address some of these issues. The interest in the conference by both members of the academia and representatives of the industry is clear evidence that electronic lexicography needs an event where current projects are presented and topical issues are discussed.

We would like to thank everyone who contributed to the success of the conference: the keynote speakers, the presenters, the sponsors, the programme committee, and the organising committee.

Iztok Kosem, editor

CONFERENCE COMMITTEES

Organising Committee

Iztok Kosem, chair
Simon Krek, co-chair
Špela Arhar Holdt
Darja Fišer
Polona Gantar
Polonca Kocjančič
Karmen Kosem
Nataša Logar Berginc

Programme Committee

Andrea Abel
Nicoletta Calzolari
Frantisek Čermak
Gilles-Maurice de Schryver
Patrick Drouin
Christiane Fellbaum
Darja Fišer
Thierry Fontenelle
Dušan Gabrovšek
Polona Gantar
Sylviane Granger
Gregory Grefenstette
Patrick Hanks
Ulrich Heid
Adam Kilgarriff
Polonca Kocjančič
Iztok Kosem
Simon Krek
Lothar Lemnitzer
Robert Lew
Nataša Logar Berginc
Rosamund Moon
Hilary Nesi
Vincent Ooi
Magali Paquot
Paul Rayson
Tadeja Rozman
Michael Rundell
Marko Stabej
Mojca Šorli
Sven Tarp
Carole Tiberius
Yukio Tono
Agnes Tutin
Darinka Verdonik
Serge Verlinde
Piek Vossen
Ana Zwitter-Vitez

TABLE OF CONTENTS

Towards a Dynamic Combinatorial Dictionary: A Proposal for Introducing Interactions between Collocations in an Electronic Dictionary of English Word Combinations

Moisés Almela, Pascual Cantos, Aquilino Sánchez
Universidad de Murcia (Spain)
Depto. de Fil. Inglesa, Fac. de Letras, Campus de la Merced, 30071 Murcia (Spain)
E-mail: moisesal@um.es, pcantos@um.es, asanchez@um.es

Abstract

This paper presents an academic (non-commercial) lexicographic project called *Dynamic Combinatorial Dictionary*, which is currently being developed by members of the LACELL Research Group at the University of Murcia. The aim of this project is to bring e-Lexicography in closer alignment with lexical models that cannot be implemented in printed dictionaries. Theoretically, the project is informed by the Lexical Constellation Model. The main difference between this model and the mainstream approaches to collocation lies in its suitability for recognising more than one domain of lexical attraction within the same collocational window. We will distinguish two different manifestations of this multiplicity of domains. The first one is the phenomenon of *indirect collocation*, which has been investigated in previous Lexical Constellation research, and the second one is *inter-collocability*. This concept refers to positive or negative dependency relations between collocational pairs (not between words). It will be argued that incorporating inter-collocability features into lexical entries can lead to significant advances in the field of combinatorial lexicography.

Keywords: collocation; lexical constellations; corpus linguistics; e-Lexicography; combinatorial dictionaries.

1. Introduction

The potential of electronic formats for increasing the variety of contextual data offered to the user, as well as for facilitating an interactive management of the information contained in lexical entries, is underexploited in current combinatorial dictionaries. This is in part due to the fact that the design of electronic combinatorial dictionaries is to a large extent informed by the design of earlier printed dictionaries. At present, the difference between electronic and printed developments in combinatorial lexicography lies more in the material format (i.e. in the medium) than in the kind and amount of information provided.

In this study, we present a proposal for exploiting more effectively and thoroughly the opportunities created by the electronic format in combinatorial lexicography. Our research is motivated by the idea that in an electronic dictionary it is possible to incorporate collocational information of a qualitatively different kind from the one that is offered to the user of a conventional collocation dictionary. More specifically, we submit that collocational information in an electronic dictionary need not be restricted to dependencies between words, and that it can be extended to include dependencies between different collocations.

The paper is structured as follows. First, in the next section we shall explain the theoretical framework of the proposal, which is based on Cantos & Sánchez's (2001) Lexical Constellation Model. It will be argued that the analysis of collocation as a relationship between lexical items is incomplete and should be complemented with a description of interactions between different collocations of a lemma. In section 3 the workings of the model are illustrated with reference to the collocational profile of the noun *goods*. The lexicographic treatment of this information is illustrated in Section 4, where we present parts of a sample entry from the *Dynamic Combinatorial Dictionary* (DCD). The advantages of the DCD over conventional approaches to combinatorial lexicography are also explained in this section.

2. The Lexical Constellation Model

The Lexical Constellation Model (henceforth: LCM) originated from the observation that the node, i.e. the word under investigation in corpus collocational research, does not exert an unlimited influence on its environment (Cantos & Sánchez, 2001). This means that the node is not the only lexical item to restrict the range of lexical choices in its textual environment. In the syntagmatic context of the node there are other lexical items which can be endowed with a context-predictive potential. To express it in more formal terms, we can say that what differentiates the LCM approach from mainstream approaches to collocation is its determination to resolve difficulties caused by the phenomenon of *lexical gravity overlaps* (or *lexical gravity interference*).

The term *lexical gravity*, as is well known, was used by Mason (2000) to denote the context-predictive potential associated with the selection of a word in the discourse. To quote the author, lexical gravity can be defined as "the restriction a word imposes on the variability of its context" (Mason, 2000: 270). The lexical gravity of a word is the influence it exerts on restricting the choice of possible words in specific positions of its textual environment.

The problem brought to the fore by LCM research is that lexical gravity can be exerted by more than one item in the same textual window. The imposition of restrictions on lexical choices in the context of the node is not an exclusive function of the node. The lexical gravity exerted by collocates of the node can interfere with the gravity attributed to the node. This begs the need to distinguish which features of lexical gravity are a contribution of the node and which ones are contributions of other elements. In this respect, the LCM outperforms the conventional approaches to collocation.

The received models of collocation are not suitable for dealing with the problem of lexical gravity interference. The reason for this is that they are linear, in the sense that they fail to divide the collocational patterns of the node into different domains of lexical attraction. The LCM seeks to resolve this problem by comparing the influence of the node and the influence exerted by other items or structures that co-exist within the same textual window.

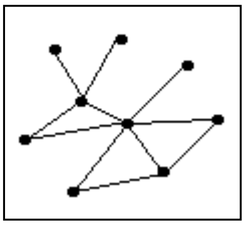


Figure 1: Structure of a plain collocational network

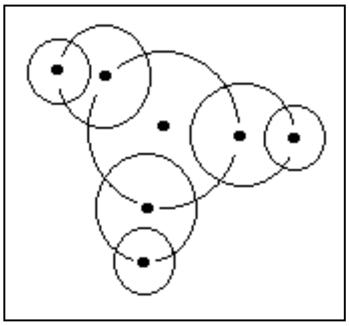


Figure 2: Structure of a lexical constellation (type 1)

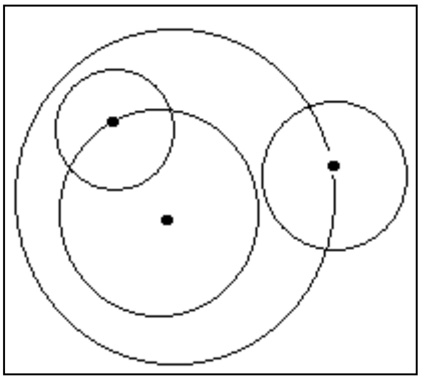


Figure 3: Structure of a lexical constellation (type 2)

The differences between plain (or linear) collocational analysis and constellational analysis are graphically represented in Figure 1, 2 and 3. In the three figures, a dot represents a lexical item, and a line represents a relationship of statistically significant co-occurrence.[1] Thus, a pair of dots connected by a line represents a collocational bi-gram. Additionally, in Figures 2 and 3 each circle symbolises a domain of lexical attraction.

Figures 2 and 3 represent different types of *lexical constellations*, the central category of description in the LCM. A lexical constellation is a collocational network hierarchically organised in two or more centres of lexical attraction. The first type of lexical constellation shown above (Figure 2) corresponds to the phenomenon of *indirect collocation*; the second type corresponds to patterns of *inter-collocability*[2] (Figure 3). These two classes of lexical constellations are described separately in the next subsections.

2.1 Indirect collocation

The phenomenon of indirect collocation was the first problem of lexical gravity interference to be investigated within the framework of the LCM (Cantos & Sánchez, 2001). This problem originates when a word so to say "intrudes" one of its collocates into the context of a another word. The phenomenon of indirect collocation is thus responsible for a large part of the "unwanted" items that are found in collocate lists.

The strategy adopted by the proponents of the LCM in order to detect cases of indirect lexical attraction among statistical collocates is based on comparisons of conditional probabilities. Once the statistically significant co-occurrences of a node have been extracted from the corpus using the conventional parameters of collocational analysis (defining a span, establishing a frequency threshold, selecting an association measure, etc.), the method proceeds to calculate the values of conditional probabilities between three different words: the node, the collocate and a candidate sub-collocate (i.e. a collocate which is suspected of being indirectly attracted because it shares more semantic features with other collocates than with the node).

[1] Following the Sinclairian line of thinking, *collocation* is defined in this study in statistical terms. Thus, it denotes a pair or group of words which co-occur with a probability greater than chance. However, we must be aware that this definition of collocation is controversial and has been criticised by notable experts in the field, especially by Bosque (2001). We will not tackle the debate here because the issue lies beyond the specific aims set for the present investigation.

[2] To avoid possible misunderstandings, a brief terminological note is in place here. The term *inter(-)collocation* is sometimes used in the literature to denote a reciprocal relationship of collocation. Thus, if a word *a* is a statistically significant co-occurrence of *b*, and *b* is a statistically significant co-occurrence of *a*, the two terms are said to form an inter-collocation. This notion of inter-collocation is not equivalent to the phenomenon that we call inter-collocability. The latter refers to a relationship between different collocational pairs.

Conditional probabilities are indicative of the strength of the dependency of one event on another event. For example, if we want to know how probable is the event *a* (say, the occurrence of a word *a*) given the occurrence of *b* as a fact, we can divide the total number of occurrences of *a* by the number of joint occurrences of *a* and *b* in a corpus. The value indicates the proportion of occurrences of *a* that take place in the company of *b*. This can be interpreted as an estimation of the dependency of the event *a* on the event *b*. The notation is P(b|a), which is read as follows: "the probability of *b* given the occurrence of *a*".

Thus, in previous research it was shown that *dental* collocates with *incidence* not because it is attracted towards *incidence* but because it has a strong dependency on another collocate of *incidence*, i.e. *caries* (Almela, 2011; Almela, Cantos & Sánchez, 2011). The probability of finding *dental* given *caries* in the Bank of English (55.2%) is more than a hundred times higher than that of finding *dental* given the occurrence of *incidence* in the same corpus (0.5%). This data is consistent with the observation that *dental* shares more semantic features with *caries* than with *incidence.* Thus, in Figure 2, the biggest circle can stand for *incidence*, the intermediate one for *caries*, and the smallest one for *dental*.

More generally, it was also found that collocates of *incidence* referring to body parts (*dental, heart, lung*, etc.) are more strongly attracted to other collocates of *incidence*, especially to those denoting a 'disease' or 'health problem' (e.g. *caries, attack, cancer*, etc.), than they are to the node. Hence, in expressions such as *incidence of dental caries*, *incidence of heart attack* or *incidence of lung cancer*, the modifier can be categorised as an indirect collocate of *incidence* (Almela, 2011; Almela, Cantos & Sánchez, 2011).

It is important to point out that grammar alone does not provide an explanation for the phenomenon of indirect collocation. Admittedly, in the above example there is a close correlation between phrase structure and the structure of the lexical constellation: the noun that modifies *incidence* is the direct collocate, and in turn, the modifier of the second noun is the indirect collocate of *incidence*. However, it should be added that in other cases the node has a closer syntactic connection with the indirect collocate than with the direct collocate. For instance, in expressions such as *caused by faulty design* or *caused by a defective gene*, the adjective is a direct collocate of the verb, and the noun is an indirect collocate (the data supporting this conclusion will be published in forthcoming research).

In sum, the analysis of indirect collocation in the LCM serves to uncover some discrepancies between statistical significance and lexical relevancy. From the fact that two or more words co-occur significantly in a corpus it does not necessarily follow that they are attracted to one another. One of the reasons for this is that there can be more than one centre of attraction within the same textual window.

The detection of cases of indirect collocation is useful in combinatorial lexicography, because it helps us to optimise the criteria for selecting the collocates of a headword. However, the implications are similar in printed and electronic dictionaries — the exclusion of irrelevant collocates is advisable in both types of dictionaries. Therefore, in what follows we will concentrate our analysis on the phenomenon of inter-collocability. As will be argued below, this second aspect of lexical constellations has important implications for the micro-structural design of collocation dictionaries, and consequently, it bears greater relevance for the discussion of issues that are specific to the field of electronic lexicography.

2.2 Inter-collocability

The concept of "inter-collocability" denotes the existence of dependency relations between different collocations of a word. The manifestations of inter-collocability are varied. In a positive sense, inter-collocability can be defined as the contribution which a collocation makes to the activation of another collocation of the same node. In a negative sense, inter-collocability can be defined in terms of restrictions on the combinational possibilities among different collocates of the same node.

As a method for identifying cases of inter-collocability we can use a variant of the technique employed for identifying cases of indirect collocation. Instead of calculating and comparing conditional probabilities between individual members of overlapping collocations, e.g. P(*a*|*b*), P(*a*|*c*), P(*b*|*c*), P(*b*|*a*), etc., we can calculate conditional probabilities between events of a larger size, for instance, the probability that a collocate of the node is selected given as a fact the co-occurrence of the node and another collocate: P(*c1*|*n,c2*), where *n* stands for the node, and *c1* and *c2* represent two different collocates. This value can then be compared with the corresponding conditional probability at the intra-collocational level, namely: P(*c1*|*n*).

Thus, given two collocates *c1* and *c2* of a node *n*, we can say that there is a relationship of positive inter-collocability between the pairs (*n,c1*) and (*n,c2*) if the probability of (*n,c1*) co-occurring with *c2* is higher than the probability of the node occurring with *c2* alone, or if the probability that (*n,c2*) co-occurs with *c1* is higher than the probability of the node co-occurring with *c1*. In the first case, we can say that *c2* is a "positive co-collocate" of *c1*, because the collocation (*n,c2*) is made more probable by the selection of *c1*; conversely, in the second case we say that *c1* is a *positive co-collocate* of *c2*, because the collocation (*n,c1*) is made more probable by the selection of *c2*. The relationship of positive co-collocation can be mutual — that is, it can be observed in the two directions, from *c1* to *c2* and vice versa.

As for negative inter-collocability, we can say that *c2* is a *negative co-collocate* of *c1* if the capacity of the node for predicting the choice of *c1* is higher than the capacity of the collocation (*n,c2*) for predicting the choice of *c1*. This indicates that the collocation of the node with c2 diminishes the probability of finding the collocation (*n,c1*); conversely, we can say that *c1* is a negative co-collocate of *c2* if the selection of the collocation (*n,c2*) diminishes the probability of (*n,c1*). Like positive inter-collocability, negative inter-collocability can be mutual: *c1* and *c2* can be negative co-collocates of one another.

Inter-collocability is extremely frequent in patterns consisting of a verb and a noun phrase, especially when the noun phrase features a modifier-noun collocation. This reflects a characteristic of argument structure that we can describe as *valency stratification*. The capacity of a predicative lexeme, typically a verb, for restricting the lexical class of its arguments can extend over more than one layer of phrase structure.

At one level, the valency carrier restricts the class of the head of the valency filler (i.e. the noun heading the argument phrase). For instance, *return* selects nouns denoting 'data' (e.g. *value, string, integer, list*, *row, zero*, *tuple*, etc.) or 'goods' (*goods*, *vehicle, equipment, medicines*, etc.), among many others. This aspect of argument structure has been extensively described under different names. In valency theory it is described as a feature of *semantic valency*, along with semantic roles. In generative grammar, the terms employed are *selectional restrictions* and *s-selection*. Bosque opts for the term *lexical restrictions* (Bosque, 2001, 2004). This aspect of valency patterning has also been extensively described in valency dictionaries and similar reference works. In Herbst et al.'s (2004) *Valency Dictionary of English*, the arguments of verbs are assigned general semantic categories. For instance, the direct object of *translate* (in its primary meaning) is categorised as 'text'. Similarly, in P. Hanks' *Pattern Dictionary* the same argument of *translate* is categorised as 'document' — for more detailed information on the semantic categorisation of arguments in this dictionary, see Hanks & Pustejovsky (2005).

Less explored, however, is the second stratum of semantic valency. On top of restricting the lexical class of the head noun, the verb can also impose constraints on the collocability of different words within the argument phrase. Generally, these constraints exhibit a high level of semantic regularity — this justifies the treatment of valency stratification as a special feature of semantic valency patterning rather than as an idiosyncratic restriction.

One way of discovering patterns of valency stratification is to analyse adjectival co-collocates of verbs. The probability that the noun co-occurs with one or other adjectival collocate is often readjusted to the selection of a specific verbal collocate. Another way of approaching the phenomenon of valency stratification is by analysing inter-collocability relations in the reverse direction — that is, by analysing adjectival co-collocates of verbs. Because different modifier-noun collocations are associated with different verbs, the probability of finding a specific verb-noun collocation will be affected by the selection of different modifier-noun collocations. In principle, we can assume that these procedures are complementary. Both of them will be applied in the next section.

3. Lexical constellations at work

In this section the analytical framework sketched out above is applied to the description of lexical constellations formed with the noun *goods*. The analysis will be focused on capturing features of inter-collocability and valency stratification in verb-noun and modifier-noun collocations.

3.1 Method and results

The data and the examples have been extracted from the *ukWaC* corpus (1,565,274,190 tokens), accessible at the SketchEngine query system. All queries are syntactically restricted. We have taken into account only occurrences of the noun phrase (i.e. the adjective-noun collocation) as a direct object of the verb in an active construction, or as the subject in a passive construction (the connection between the two constructions is that in both cases the collocation ADJ+*goods* performs the semantic role THEME). The WordSketch function proved very useful in limiting our queries to the foregoing grammatical scheme. Nevertheless, manual supervision was required in order to detect possible parsing errors.

Following the remarks made at the end of section 4, we have approached the phenomenon of inter-collocability from two complementary perspectives. Tables 1-3 reflect the perspective provided by the analysis of adjectival co-collocates, and Tables 4-6 reflect the perspective provided by verbal collocates.

The criteria applied in the selection of the potential co-collocates were aimed at testing the initial hypothesis that the lexical constellations of *goods* follow highly systematic semantic patterns (at this point it should be remembered that in section 3 valency stratification was described as a special feature of semantic valency). The verbs *return, replace* and *reject* have been selected because they share important aspects of meaning. In collocation with *goods* they denote an action whereby the consumer does not accept the goods initially bought or received. As for the adjectives *faulty*, *defective* and *damaged*, they all describe a 'flaw' or 'imperfection'.

The results are shown in Tables 1-6. In each table the left-most column is a list of collocates of *goods*. In Tables 1-3 these collocates are modifiers (more specifically adjectives)[3], and in Tables 4-6 they are verbs. A frequency threshold and a statistical filter were applied to all the collocates. We made sure that all of them co-occur at least three times with *goods* (in the specified grammatical framework), and that they all are statistically significant co-occurrences of this noun. Statistical significance was defined in terms of *logDice* — for an explanation of the advantages of this measure the reader is referred to Rychlý (2008). Again, these data were obtained from the WordSketch function at SketchEngine.

The next two columns indicate raw frequency data. The first of them indicates the frequency of the whole 3-gram (verb, modifier, noun) in the corpus (for instance, the frequency of *return defective goods*). A minimum frequency threshold of 3 was also applied in this column. This was motivated by purely practical reasons that are independent of the research methodology: the list of 3-grams with a frequency lower than two would generate excessively long tables difficult to fit into the size of this paper.

The second frequency data column corresponds to the collocational pair formed by the noun (*goods*) and each of the collocates listed in the left-most column. Thus, in Tables 1-3 this column specifies the frequency of modifier-noun collocations (e.g. *faulty* + *goods, defective* + *goods*, etc.), while in Tables 4-6 the same column indicates the frequency of verb-noun collocations (e.g. *return* + *goods, replace* + *goods*, and so on). These data were obtained by checking the results in different SketchEngine tools (Concordance, WordSketch, Collocation, etc.).

As for the last two columns, they indicate values of conditional probabilities between collocations and between words, respectively. The first of these columns returns the value of P(*m*|*v,n*) in Tables 1-3, and of P(*v*|*m,n*) in Tables 4-6. The first formula can be read as "the probability that the modifier occurs given the occurrence of the verb+noun collocation" (where the noun is always *goods*). Correspondingly, the second formula can be read as "the probability that the verb occurs given the occurrence of the modifier+noun collocation". Finally, the right-most column returns the value of P(*m*|*n*) in Tables 1-3, and of P(*v*|*n*) in Tables 4-6. The first value reflects the probability that the modifier occurs given the occurrence of the noun; the second one specifies the probability that the verb occurs given the selection of the noun. In all the tables the order of the rows is determined by the difference between the values of these two columns. Thus, the word at the top of the list is the best candidate for positive co-collocate.[4]

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
faulty	35	354	2.35%	0.36%
unwanted	21	149	1.41%	0.15%
defective	20	137	1.34%	0.14%
unused	7	20	0.47%	0.02%
undamaged	6	10	0.40%	0.01%
damaged	8	209	0.54%	0.21%
non-faulty	4	11	0.27%	0.01%
stolen	8	434	0.54%	0.44%

Table 1: Adjectival co-collocates of *return*.[5]

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
faulty	30	354	19.11%	0.36%
defective	12	137	7.64%	0.14%
damaged	12	209	7.64%	0.21%
electrical	6	850	3.82%	0.86%

Table 2: Adjectival co-collocates of *replace*.[6]

	f(v,m,n)	f(m,n)	P(m|v,n)	P(m|n)
faulty	6	354	5.41%	0.36%
defective	3	137	2.70%	0.14%

Table 3: Adjectival co-collocates of *reject*.[7]

	f(v,m,n)	f(v,n)	P(v|m,n)	P(v|n)
return	35	1491	9.89%	1.50%
replace	30	157	8.47%	0.16%
receive	19	913	5.37%	0.92%
buy	17	1592	4.80%	1.60%
reject	6	111	1.69%	0.11%
supply	6	961	1.69%	0.97%
collect	3	270	0.85%	0.27%
sell	7	2237	1.98%	2.25%

Table 4: Verbal co-collocates of *faulty*.[8]

[3] A priori we did not decide to exclude noun modifiers from this list (e.g. *consumer goods*, *household goods*, etc.). However, for some reason, none of the modifiers that met the conditions set in the first three columns were nouns; all of them were adjectives.

[4] In Tables 5 and 6, the position of *deliver* at the bottom of the list might be misleading. The value of P(v|n) in this row is inflated by occurrences of the idiom *deliver the goods*. If we were able to exclude this idiom from the count of collocations of *deliver* + *goods*, the difference with P(v|m,n) would be greater in Table 5, and in Table 6 the value of P(v|m,n) would be higher than P(v|n). However, the occurrences of *deliver the goods* as an idiom cannot be separated automatically from those of *deliver the goods* as a collocation, and doing it manually is far too time-consuming a task to be considered a convenient method in lexicography.

[5] F(*return,goods*) = 1491

[6] F(*replace,goods*) = 157

[7] F(*reject,goods*) = 111

	f(v,m,n)	f(v,n)	P(v|m,n)	P(v|n)
return	20	1491	14.60%	1.50%
replace	12	157	8.76%	0.16%
reject	3	111	2.19%	0.11%
inspect	3	121	2.19%	0.12%
deliver	4	1930	2.92%	1.94%

Table 5: Verbal co-collocates of *defective*.[9]

	f(v,m,n)	f(v,n)	P(v|m,n)	P(v|n)
receive	15	813	7.18%	0.82%
replace	12	157	5.74%	0.16%
return	8	1491	3.83%	1.50%
inspect	3	121	1.44%	0.12%
deliver	4	1930	1.91%	1.94%

Table 6: Verbal co-collocates of *damaged*.[10]

The frequency of the noun remains constant in all the tables. The frequency of the noun *goods* in the corpus is 99393 (substantivisations of the adjective *good* were excluded from this count). Besides, the frequency of verb-noun collocations remains constant within each of the first three tables. Likewise, the frequency of modifier-noun collocations remains constant within each of the last three tables (4-6). Therefore, the figures are indicated in a footnote added to the caption.

3.2 Analysis and discussion

The results displayed in Tables 1-6 lend strength to the initial hunch that the lexical constellations of *goods* exhibit a high degree of semantic systematicity. The strongest positive co-collocates tend to be grouped together around a common core of meaning.

In Tables 1-3 the dominant group of adjectives is formed by words depicting a 'flaw': *faulty, defective, damaged*. Observe that *faulty* and *defective* occur in the three tables, and that in all of them *faulty* lies at the top. The fact that *unwanted* is a stronger co-collocate than *defective* in Table 1 does not run counter to the general pattern, because the meaning of *unwanted* is conceptually related to *faulty*, *defective* and *damaged* (as a rule, goods that are in a bad condition are not desired by the consumer).

Particularly significant are the values of conditional probabilities in Table 2. Observe that the capacity of the collocation *replace goods* for predicting the choice of *faulty* reaches 19.11 percent, a figure more than 50 times higher than the capacity of the noun *goods* for predicting the selection of *faulty*. This constellation is thus a very good example of the kind of dependency relation depicted in Figure 3 (see section 2). If we insert these lexical items in Figure 3 we obtain the picture below:

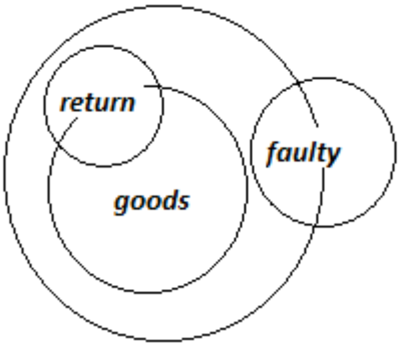


Figure 4: Positive inter-collocability

The results displayed in Tables 4-6 are equally coherent from the point of view of meaning. The dominant group is formed by verbs implying a decision of 'non-acceptance of the goods received': *return, replace, reject*. The verbs *return* and *replace* appear in the three tables, and in two of them, *return* is the strongest co-collocate.

Overall, the semantic regularities observed in these lexical constellations suggest that verb-noun collocations expressing 'non-acceptance of goods' are likely to converge with adjective-noun collocations describing goods as 'having a flaw'. This speaks strongly for the conception of lexical constellations as surface lexical realisations of underlying conceptual (cognitive) structures. In the same line of reasoning, it would be interesting to determine the extent to which lexical constellations are language-independent. Obviously, this objective cannot be pursued in the present article, because it requires more empirical research in English and in other languages.

Another interesting remark concerns the consistency of the findings obtained in the two groups of tables (1-3 and 4-6). The output of Tables 1-3 overlaps with the input of Tables 4-6, and vice versa. The dominant adjectives in Tables 1-3 coincide roughly with the elements analysed in Tables 4-6, and conversely, the dominant verbs in Tables 4-6 contain the elements analysed in Tables 1-3. This confirms the claim made in section 3.2 that co-collocation can be mutual. *Defective* is a co-collocate of *return*, and conversely, *return* is a co-collocate of *defective* (see Tables 1 and 4). The same holds true for other pairs: (*defective, replace*), (*defective, reject*), (*faulty, return*), (*faulty, replace*), (*faulty, reject*), (*damaged, return*), (*damaged, replace*). This reinforces the idea that the two perspectives on valency stratification (the one provided by verbal co-collocates and the one provided by modifiers) are complementary and lead to relatively similar results.

Finally, it should be noted that the prevalence of positive co-collocates over negative ones in Tables 1-6 results

[8] F(*faulty,goods*) = 354
[9] F(*defective,goods*) = 137
[10] F(*damaged,goods*) = 209

mainly from the decision to set a minimum frequency threshold for the 3-gram. If the analysis had been focused on verbs or adjectives occurring in low-frequency 3-grams, we would have obtained several prominent patterns of negative inter-collocability. Interestingly, these patterns can also be characterised by a high degree of semantic regularity.

A case in point is the relationship between verbs such as *ship* and *transport* and the adjectives analysed in tables 4-6. There is evidence that *ship* and *transport*, which are quasi-synonyms, are negative verbal co-collocates of *faulty*, *defective* and *damage*. The probability of these verbs given the occurrence of *goods* is 0.31 percent in the case of *ship* (309/99393), and 0.43 percent in the case of *transport* (426/99393). These figures, however low, are considerably greater than the probability of these verbs occurring in the context of modifier-noun collocations such as *faulty goods*, *defective goods*, or *damaged goods*. In almost all these cases the probability is zero. In the whole ukWaC corpus, which, it should be emphasised, contains more than one billion words, there is no single instance of 3-grams such as *ship defective goods*, *transport faulty goods*, *transport damaged goods*, etc. The sequence *ship faulty goods* yields one hit, but obviously the value of P(*ship*|*faulty goods*) is lower than P(*ship*|*goods*). Clearly, the collocations *ship/transport goods* tend to avoid the selection of modifiers describing a 'flaw' or 'imperfection'. This can be interpreted as an indication that semantic systematicity is a characteristic both of positive and of negative inter-collocability.

4. Lexical constellations in lexicography

From the previous sections we can draw the overall conclusion that the choice of a collocation influences the range of choice of other collocations in the same context. The choice of a collocation can contribute to activating or blocking other collocations of the same node. Once this fact has been established, the question that needs to be addressed is: should lexical constellations be recorded in combinatorial dictionaries, and if so, what are the appropriate lexicographic techniques for dealing with them? The first part of the question is answered in 4.1. The answer to the second part of the question is given in 4.2. In Section 4.3 we explain the guidelines for our lexicographic project and present some examples.

4.1 The relevance of constellational information

Lexical constellations provide a potentially useful type of information in a collocation dictionary. One of the main functions of this kind of dictionary is to assist the user –typically a foreign or second language speaker– in achieving native-like, fluent composition. Precisely, lexical constellations are one of the principal resources of fluency and cohesion in a text, because they make the word fit within a context broader than the simple collocational bi-gram. Compared to the simple collocation, a lexical constellation provides, so to say, an extended pattern of lexical cohesion.

Apart from this general consideration, there are two more specific arguments for introducing lexical constellations into collocation dictionaries. The first of these arguments concerns the strength of constellational patterns. In some respects, these patterns are stronger than most of the simple collocational bi-grams recorded in a conventional combinatorial dictionary. Observe, for example, that the dependency of the collocation *defective goods* on *return*, measured in terms of conditional probability, is ten times higher than the dependency of *goods* on *return* (see Table 5). In this light it is difficult to justify why the weaker pattern (the bi-gram) should be included in a dictionary while the stronger pattern (the constellation) is omitted.

A further argument for the incorporation of collocational data refers to the connection of form and meaning. The syntagmatic behaviour of words is closely associated with their semantic properties. Therefore, collocation is more than a surface co-occurrence pattern; it also provides a representation of word meaning (Renouf, 1996). Knowing the collocations of words is a contributing factor to the development of lexical semantic competence. This idea, which was formulated by Firth in his well-known definition of "meaning by collocation", has inspired much of the work conducted in corpus-driven lexicology, both theoretical and applied. Lexical constellations can help to provide a much more detailed and refined account of the connections between context and meaning. Notice, for example, that some semantic aspects of adjectives such as *faulty*, *defective*, or *damage* are better represented by their verbal co-collocates (*reject, return, replace*) than by the noun (*goods*). The discovery of a 'flaw' is causally connected with the decision of 'non-acceptance', and this decision is implied by the meaning of verbs such as *reject, return*, or *replace*, but not by the meaning of *goods*.

Considering these arguments, we can conclude that lexical constellations can improve the utility of collocation dictionaries. Having answered this question, the next problem to be resolved concerns the know-how. Clearly, the incorporation of lexical constellations requires the development of innovative practices, because current collocational dictionaries do not provide this kind of information. This gives rise to the question: what exactly are changes that have to be introduced in order to accommodate lexical constellations into collocation dictionaries? The issue is addressed below.

4.2 The treatment of constellational information

By definition, a lexical constellation always involves some form of interaction between different collocations of a node. However, in a standard collocation dictionary the different words in an entry are directly related to the headword and not to one another. Therefore, the main obstacle that has to be overcome in order to integrate constellations within collocation dictionaries is the lack of explicit connection between different collocates of a headword.

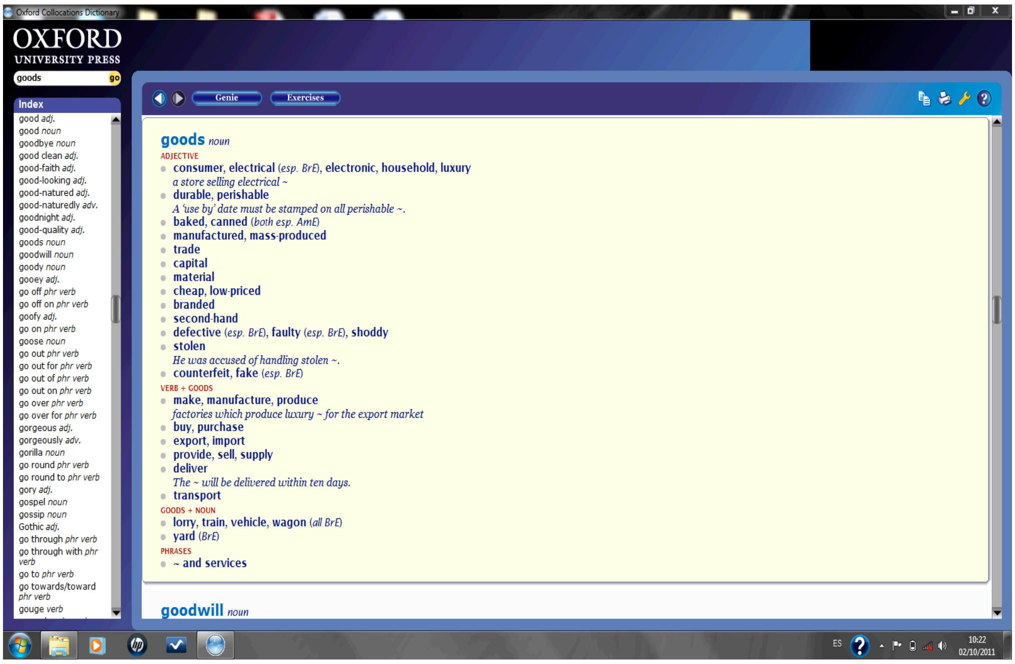
The incorporation of lexical constellations requires us to take a step from an "intra-collocational" to an "inter-collocational" perspective. Thus far, the analysis of syntagmatic dependencies in collocation dictionaries, both printed and electronic, has been focused on relationships between the parts of a collocation. In this sense, we can say that combinatorial lexicography has done justice to Sinclair's (1991) remark that the choice of a word affects the choice of other words in its vicinity. What combinatorial lexicography has so far failed to reflect is the fact that the choice of a collocation can also affect the choice of other collocations in its vicinity.

In a conventional collocation dictionary the user is not provided with information concerning how different elements and sections in an entry can or tend to be combined in the discourse. There is, of course, information about the relationship between the headword and each of the collocates. However, this is not complemented by any specification of whether particular collocations or groups of collocations of the lemma tend to attract or repel each other.

For example, in *Macmillan Collocations Dictionary* (MCD), to quote the most recent major dictionary of English collocations, *faulty* and *return* are presented as different categories of collocates of the noun *goods*. *Faulty* is one of the three adjectives in this entry, along with *defective* and *damaged*, which are labelled as expressing the meaning 'not working properly'. *Return* is one of the four verbal collocates in the same entry which are ascribed the meaning 'send goods' (the others are *deliver*, *transport* and *ship*). From this information we can gather that *return faulty goods* and *transport faulty goods* are possible lexical combinations expressing the meaning 'send goods which do not work properly'. What we are not told, however, is that the selection of the modifier is adjusted to choice of a verbal collocate, and that the selection of *return goods* makes the selection of the adjective *faulty* highly probable, while the collocations *transport goods* and *faulty goods* tend to avoid each other. That is to say, the MCD does not inform us that some pairs of verbal and adjectival collocates of *goods* are more likely to converge in the same complex expression than others.

These facts are not reflected in the MCD or in any other major collocation dictionary, because the design does not contemplate any form of interaction between different collocations in an entry. The same remark applies to other important combinatorial dictionaries of English, notably the BBI and the OCD, or of other languages such as Spanish (e.g. REDES).

This criticism can also be made of electronic collocation dictionaries, such as the *Diccionario de Colocaciones del Español* (DiCE, an online dictionary of Spanish collocations), as well as of electronic versions of printed dictionaries (e.g. the OCD on CD-ROM). In none of these resources is the user provided with specifications of how the selection of a collocate influences the range of choice of further collocates of the same headword. Observe, for example, Figure 5, where we reproduce an entry from the OCD on CD-ROM. Here, we find some of the adjectival and verbal collocates of *goods* mentioned above (*faulty, defective, deliver, transport*, etc.), but again, no specification is given of their inter-collocability.



goods *noun*

ADJECTIVE

- **consumer, electrical** (*esp. BrE*), **electronic, household, luxury**
  a store selling electrical ~
- **durable, perishable**
  A 'use by' date must be stamped on all perishable ~.
- **baked, canned** (*both esp. AmE*)
- **manufactured, mass-produced**
- **trade**
- **capital**
- **material**
- **cheap, low-priced**
- **branded**
- **second-hand**
- **defective** (*esp. BrE*), **faulty** (*esp. BrE*), **shoddy**
- **stolen**
  He was accused of handling stolen ~.
- **counterfeit, fake** (*esp. BrE*)

VERB + GOODS

- **make, manufacture, produce**
  factories which produce luxury ~ for the export market
- **buy, purchase**
- **export, import**
- **provide, sell, supply**
- **deliver**
  The ~ will be delivered within ten days.
- **transport**

GOODS + NOUN

- **lorry, train, vehicle, wagon** (*all BrE*)
- **yard** (*BrE*)

PHRASES

- **~ and services**

goodwill *noun*

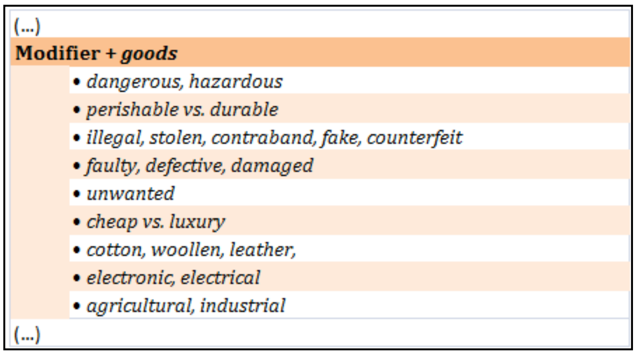
Figure 5: An entry from the OCD on CD-ROM

Logically, the possibilities of accommodating lexical constellations are not equal for printed dictionaries and electronic dictionaries. The printed format imposes a number of material conditions which render the incorporation of lexical constellations virtually impracticable. Supplying this kind of information in a printed dictionary would imply an excessive increase in size, probably beyond what is commercially viable. However, these practical difficulties can be resolved in an electronic dictionary. The user interface allows an interactive management of the information contained in lexical entries. With a simple click, the user can choose to expand the information on the collocations associated with a particular item, and precisely, one of the choices that can be made available in this menu is the generation of a list of collocates that are attracted to specific collocations of the lemma. For these reasons, we think that, in the present state of the art, the project of developing a collocation dictionary that includes lexical constellations is conceivable only in electronic format.

4.3 The *Dynamic Combinatorial Dictionary*

The treatment of lexical constellations in our lexicographic project, the DCD, follows four main guidelines: *dynamicity*, *progressiveness*, *compactness* and *systematicity*. Firstly, the micro-structural design is dynamic, because the information presented in a lexical entry is readjusted to the selections made by the dictionary user. This is why the project has been called a *Dynamic Combinatorial Dictionary*. This means, for example, that by clicking on the collocate *faulty* under the entry for *goods* the positive verbal co-collocates (e.g. *return, replace, reject*) are foregrounded, and the negative ones are omitted.

Secondly, the step from simple collocational bi-grams to lexical constellations is made in a progressive manner. As a default option, the entry offers only plain collocational information. The user is not provided with information on lexical constellations before s/he clicks on a specific collocate in search for more detailed information, and when this happens, the entry zooms in to show only the most relevant contextual data. That is, in the transition from purely collocational information to constellational information the dictionary leaves out all those elements which are not positive co-collocates of the items selected by the user. The principle behind this criterion is one of user-friendliness. It is not advisable to increase at the same time the level of detail and the amount of information. An increase in the depth of information should be compensated by a decrease in the width of information.



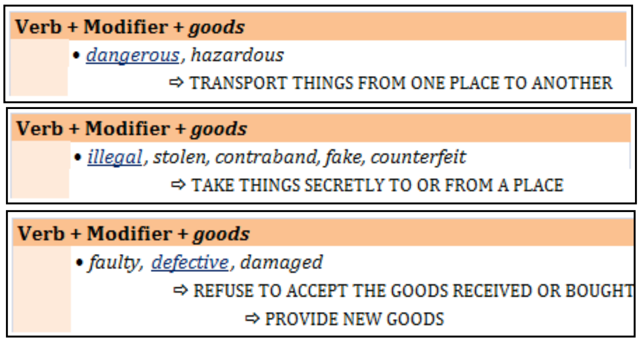
(...)

Modifier + *goods*

- *dangerous, hazardous*
- *perishable vs. durable*
- *illegal, stolen, contraband, fake, counterfeit*
- *faulty, defective, damaged*
- *unwanted*
- *cheap vs. luxury*
- *cotton, woollen, leather,*
- *electronic, electrical*
- *agricultural, industrial*

(...)

Figure 6: Extract from a DCD entry (first stage)



Verb + Modifier + *goods*

- *dangerous, hazardous*

⇨ TRANSPORT THINGS FROM ONE PLACE TO ANOTHER

Verb + Modifier + *goods*

- *illegal, stolen, contraband, fake, counterfeit*

⇨ TAKE THINGS SECRETLY TO OR FROM A PLACE

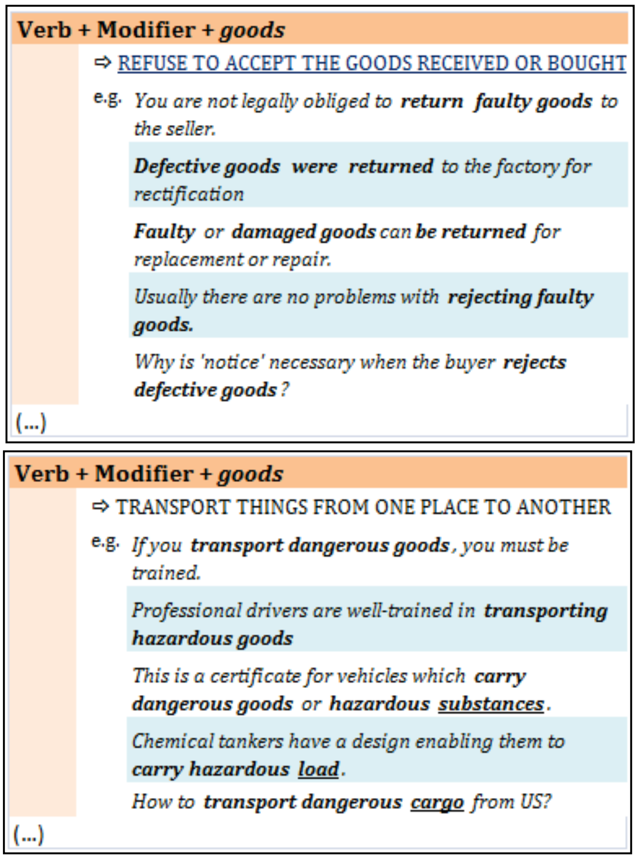
Verb + Modifier + *goods*

- *faulty, defective, damaged*

⇨ REFUSE TO ACCEPT THE GOODS RECEIVED OR BOUGHT

⇨ PROVIDE NEW GOODS

Figure 7: Extracts from a DCD entry (second stage)



Verb + Modifier + *goods*

⇨ REFUSE TO ACCEPT THE GOODS RECEIVED OR BOUGHT

e.g. *You are not legally obliged to **return faulty goods** to the seller.*

***Defective goods were returned** to the factory for rectification*

***Faulty** or **damaged goods** can **be returned** for replacement or repair.*

*Usually there are no problems with **rejecting faulty goods.***

*Why is 'notice' necessary when the buyer **rejects defective goods**?*

(...)

Verb + Modifier + *goods*

⇨ TRANSPORT THINGS FROM ONE PLACE TO ANOTHER

e.g. *If you **transport dangerous goods**, you must be trained.*

*Professional drivers are well-trained in **transporting hazardous goods***

*This is a certificate for vehicles which **carry dangerous goods** or **hazardous substances**.*

*Chemical tankers have a design enabling them to **carry hazardous load**.*

*How to **transport dangerous cargo** from US?*

(...)

Figure 8: Extracts from a DCD entry (third stage)

The level of detail or granularity in DCD entries unfolds gradually through three different steps. In a first stage, the screen displays collocational pairs, similarly to conventional collocation dictionaries (Figure 6). In a second stage, the screen displays a semantic description of lexical constellations related to the collocate on which the user has clicked (see Figure 7). Finally, in a third stage, the user is provided with a series of examples representing different lexical realisations of the constellation (see Figure 8). This list is accessed by clicking on the semantic description of the constellation. Where relevant, the list includes references to other headwords sharing in the same lexical constellation pattern (e.g. *cargo, load, substance*, etc., in the lower part of Figure 8). In these cases, the words are underlined so that the user can follow the link to the corresponding noun entry.

Concerning the third guideline, i.e. compactness, information about lexical constellations is presented in a format as succinct as possible. One implication is that labels such as "lexical constellation", "inter-collocability" or "positive co-collocate" are not explicitly mentioned by any means in the entry. This marks a difference with some collocation dictionaries, especially in the Meaning-Text Theory (MTT) framework (notably the DiCE), which make extensive use of specialised terms that are not known to the wider audience and the lay speaker. These terms include MTT jargon such as *gloss* and lexical function labels such as 'Magn', 'Anti Bon', etc. In the DCD project we try to make the dictionary accessible by keeping metalinguistic data to a minimum. Metalinguistic information is reduced to basic grammatical categories (Verb, Noun, Adjective, etc,) and to semantic labels. For similar reasons, probability and statistical data are not shown to the user. The structure of constellations is signalled only by means of symbols such as arrows, and by highlighting words in authentic examples (see Figures 7 and 8).

Finally, the fourth guiding principle is the maximisation of systematicity. This apparently trivial statement contains important implications for the design of dictionary entries. It entails, among other things, the attempt at subsuming as much lexical information as possible under general combination rules. This implies first and foremost that semantic labels will be used to show the interconnectedness of several collocational patterns.

This practice, i.e. the grouping of different collocations under meaning categories, has been adopted to a greater or lesser extent by previous collocation dictionaries such as MCD, REDES and the DiCE, but no by others such as the OCD or the BBI. The specific challenge faced now by the DCD is to extend this strategy to apply to the description of semantic regularities underlying lexical constellations. This problem is resolved by inserting semantic paraphrases of constellations at an intermediate stage between collocational information and real examples of constellations (see Figures 7 and 8).

The rationale behind this emphasis on the connection of combinatorial and semantic properties of words is our strive for abridging the distance between the collocation dictionary and the general-purpose dictionary. In the line of neo-Firthian thinking, it is our conviction that a well-organised, detailed description of the syntagmatic behaviour of a word has a definitional value. Collocation provides a representation of word meaning, as Firth suggested.

5. Conclusion

In this article we have argued that the mainstream approaches to collocation have missed an important aspect of collocational patterning, namely, the operation of dependency relations between different collocations. Crucially, this level of analysis should not be confused with observation of dependency relations between the parts of a collocation. Collocability must be analysed at a different level than inter-collocability.

It has also been argued that the LCM provides an adequate analytical framework for inter-collocability. After applying the methodology of constellational analysis to collocational patterns of the noun *goods*, we have confirmed that different collocations influence in different ways the selection of other collocations of the same noun.

Finally, we have explained that dealing with lexical constellations in a dictionary is only possible in an electronic format and requires us to introduce a number of substantial changes with respect to the conventional micro-structural design of collocation dictionaries (including electronic ones). Some of these changes have been illustrated with reference to sample parts from the DCD.

6. Acknowledgements

The project presented in this paper is generously funded by a grant from *Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia* (Ref. 08594/ PHCS/08). We are most grateful for this financial support.

7. References

Almela, M. (2011). Improving corpus-driven methods of semantic analysis: a case study of the collocational profile of 'incidence'. *English Studies*, 92(1), pp. 84-99.

Almela, M., Cantos, P. & Sánchez, A. (2011). From collocation to meaning: revising corpus-based techniques of lexical semantic analysis. In I. Balteiro (ed.) *New Approaches to Specialized English Lexicology and Lexicography*. Newcastle u. T.: Cambridge Scholars Press, pp. 47-62.

The BBI Dictionary of English Word Combinations

(1997). Compiled by M. Benson, E. Benson & R. Ilson. Amsterdam: John Benjamins.

Bosque, I. (2001). Sobre el concepto de 'colocación' y sus límites. *Lingüística Española Actual*, 23(1), pp. 9-40.

Bosque, I. (2004). La direccionalidad en los diccionarios combinatorios y el problema de la selección léxica. In T. Cabré (ed.) *Lingüística teórica: anàlisi i perspectives*. Bellaterra: Universitat Autonoma de Barcelona, pp. 13-58.

Cantos, P., Sánchez, A. (2001). Lexical constellations: what collocates fail to tell. *International Journal of Corpus Linguistics*, 6(2), pp. 199-228.

DiCE: Diccionario de colocaciones del español. Accessed at: http://www.dicesp.com.

Hanks, P., Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Révue Française de Linguistique Appliquée*, 10, pp. 63-82.

Herbst, T., Heath, D., Roe, I.F. & Götz, D. (2004). *A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin: Mouton de Gruyter.

Mason, O. (2000). Parameters of collocation: the word in the centre of gravity. In J.M. Kirk (ed.) *Corpora Galore. Analyses and techniques in describing English*. Amsterdam: Rodopi, pp. 267-280.

Macmillan Collocations Dictionary for Learners of English (2010). Compiled by M. Rundell. Oxford: Macmillan.

Oxford Collocations Dictionary for Students of English (2009). Compiled by C. McIntosh. Oxford: Oxford University Press.

A Pattern Dictionary of English Verbs. Accessed at: http://deb.fi.muni.cz/pdev/.

REDES: Diccionario combinatorio del español contemporáneo (2004). Compiled by I. Bosque. Madrid: SM.

Renouf, A. (1996). Les nyms: en quête du thésaurus des textes. *Lingvisticae Investigationes*, 20(1), pp. 145-165.

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka, A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*. Brno: Masaryk University, pp. 6-9.

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

Collocational networks and their application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci)

Araceli Alonso, Chrystel Millon, Geoffrey Williams
Equipe LiCoRN - Université de Bretagne-Sud
Faculté de Sciences Humaines et Sociales
4 rue Jean Zay 56321 Lorient CEDEX (France)
E-mail: araceli.alonso@univ-ubs.fr, chrystel.millon@univ-ubs.fr, geoffrey.williams@univ-ubs.fr

Abstract

The present article deals with a situation that lies between the needs of an advanced learner's dictionary and those of a specialised dictionary in attempting to build a pattern dictionary for verbs which are being used in scientific research papers. Current dictionaries do not necessary assist in the particular production environment of the scientific article. This can be tackled by building a bottom-up phraseological dictionary which will help both with decoding and encoding. The building method uses collocational networks in order to compile a dictionary which will demonstrate usage of individual verbs, grouping them into a natural classification system that will grow from the corpus data. This organic dictionary ultimately makes wide use of mind mapping technology to allow the user to navigate within the dictionary. It contains both individual entries containing phraseological information and super entries linking quasi-synonyms and writing assistance. The dictionary provides the environment which can link phraseological patterns to the corpus data so as to limit the information retrieval process whilst providing real examples of language in use in specialised contexts.

Keywords: learner's dictionary; specialised dictionary; organic dictionary; phraseology; collocational networks; verbal patterns

1. Introduction

In recent years, developments in technology have brought about some major changes in dictionary-writing. The ground-breaking work of Sinclair and the COBUILD team in the 1980s introduced a move in lexicographical practice towards the creation of corpus-based dictionaries on the basis that users need to know not only the meaning of the word, but the way the word is used in context. Many monolingual and, especially, learner's dictionaries have applied corpus-based techniques for representation of word uses by giving examples taken from a corpus. Although the corpus is now integrated as a source, most of these dictionaries, whether print or cd-rom in format, have not implemented the full potential of adopting a corpus-driven approach to what may be extracted from a corpus, such as the networks of relations between words.

The rise of electronic dictionaries due to the widespread use of computers and especially of Internet has also contributed to pushing lexicographical practice further, even if technology changes much more quickly than the dictionary-writing process. As a result, many on-line dictionaries do not take full advantage of the potential offered by web technology. In fact, many of them are just a copy of the paper dictionary. Some attempts at creating genuine on-line dictionaries, such as, the visual dictionaries, or *Wordnik* (http://www.wordnik.com), based on the web 2.0 or social web, and the like, have been made, but there is still a long way to go. New approaches to dictionary-making practice are needed. In a society of knowledge and technology, dictionaries must be updated and adapted to the users' needs.

In the case of science dictionaries, there is a real need for innovation. Most dictionaries of science are very traditional in outlook or simply take the form of terminological databases applying an onomasiological approach which supply the user with a definition and a context, and in some cases, relations between the units, but fail to give detailed information on the syntagmatic and paradigmatic relations between technical words, and technical words and 'general' words. In reality, specialised communication is not just about technical words. In most cases, scientists already know the definition of the technical word, but look up the 'specialised' meaning of a general word in the dictionary, for getting information of the behaviour of the word in a domain-specific context.

The wealth of language lies in semitechnical words and general words in specific contexts. As has been stated by many authors (Cabré, 1999; Meyer, 2000; Ciapuscio, 2003; Hunston & Sinclair, 2003; ten Hacken, 2008), the dichotomy between general and specialised languages must be seen in terms of a continuum; they are not clearly separable entities. It can be stated that there is a transfer of lexical units from one side to the other; processes of *determinologization* or *banalization*, *terminologization* and *pluriterminologization* take place (Meyer, Mackintosh & Varantola, 1997; Cabré, 1999). This passage of meaning potentials from general language to specialised language, and back, is particularly a problem for non-native speakers who need to communicate in scientific contexts.

Furthermore, most specialised dictionaries consider only nouns as entries of the dictionary, as according to a classical perspective of terminology, the noun was considered the only category to have a terminological value, and do not take into account the role that other categories, such as verbs, can play in specialised discourse. In order to produce a text, non-native speakers need to understand the characteristics of the specialised discourse and it is not only the noun which plays a relevant role. Verbs, for instance, can help to organize the discourse, to articulate and structure the text, to establish links between different referential lexical units, to express the point of view of the author, to interactuate with the reader, to understand the meaning of a word, etc. As Hanks states '*meanings are constructed around the verb, the pivot of the clause*' (Hanks, 2010a:3). Therefore, for a language learner, it is extremely important to get to know the behaviour and use of verbs in order to be able to produce and understand a specialised discourse. A dictionary for verbs used in the sciences can assist by helping users to overcome their basic communication problems.

The main objective of this communication is to present the potential of collocational networks for a new approach to an experimental dictionary conceived from the beginning as a virtual dictionary, the *E-Advanced Learner's Dictionary of Verbs in Science* (DicSci). Collocational networks for the building-up of dictionary entries will be discussed and exemplified with reference to the most frequent verbs extracted from a corpus related to BioSciences.

This paper shows how specialised learner's dictionaries have evolved. The article presents the initial premises of the lexicographical project *DicSci*, paying special attention to the 'organic' nature of the *E-Advanced Learner's Dictionary of Verbs in Science*, and describes the work methodology and building-up of the dictionary by showing the verb *to take* as an example. Finally, some conclusions and perspectives are outlined.

2. Learner's dictionaries of science

Dictionaries of science or specialised dictionaries are usually terminologically based which have been elaborated taking into account terminological theoretical and methodological framework rather than those of lexicography, particularly those of advanced learner's dictionaries. In many cases, they are terminological databases. Most of these terminological dictionaries or databases are based on an onomasiological perspective, that is, the different entries are organized by means of the concepts. The terminological units are just means of the linguistic expression of the conceptual organisation of a particular domain. The focus is on explaining the concept, with the terminological unit is only observed as a way of designing the concept. Therefore, no attention is being paid to the different senses of a term, as the term is not considered as a lexical unit. More recent approaches to terminology advocate a semasiological approach to terminology — see L'Homme (2005) for more detailed information —, considering the term as a lexical unit which can have the same characteristics of other lexical units of general language. Despite this fact, little progress has been made in specialised dictionary practice. As Williams (2003:94) states, most of these dictionaries, whether multilingual or bilingual terminologies, mostly address the translator, whereas the monolingual encyclopaedic dictionaries principally address the subject specialist. The latter are prescriptive dictionaries whose main aim is to fix and explain terms for native-speakers of the language. They have not been compiled with the foreign learner's needs in mind. They do not explain use of terms in context and, therefore, are of little help for encoding purposes. In the 80s Moulin (1983:151) already considered that existing specialised dictionaries were of little use to foreign learners. Not much progress has been made since then.

Even though, some attempts have been made during these years to answer foreign learner's needs, learner's dictionaries of science are not really satisfactory. Some authors (Bergenholtz & Tarp, 1995; Fuertes-Olivera, 2009, 2010; Tarp, 2008) have defended a functional approach to lexicography, usually referred to as the *Function Theory of Lexicography*, considering lexicography as an area of social practice where the dictionary must take into account users' specific types of problems and situations and satisfy user's needs. From this perspective some specialised dictionaries for foreign learners have been compiled. And even though more attention has been paid to the linguistic characteristics of the terminological units, many problems of grammar and usage have received only minimal attention.

On the other hand, learner's dictionaries of English as a foreign language have a strong tradition, but aim at general usage with little coverage of the sciences. Over the years, learner's dictionaries of English as a foreign language have increased in number and variety. Since the first learner's dictionaries much work has been done for giving more information — see Cowie (2002, 2009) for a detailed history of English dictionaries for foreign learners —, paying special attention to the linguistic features of language. However, most advanced learner's dictionaries have paid little attention to the representation of specialised lexical units, being primarily aimed at learners of the language for general purposes. A similar situation can be found with standard bilingual dictionaries which essentially provide decontextualized equivalents with a minimum of encoding assistance. Consequently, many scientists have to rely on 'native English speakers', hopefully with an awareness of genre specificities, to correct their texts.

Learner's dictionaries of English as a foreign language deal with grammatical and usage aspects of lexical units, as the learner of the language need information not only for understanding texts, but also for producing texts in the foreign language. For instance, many dictionaries have made an attempt to introduce information on collocations, "lexical co-occurences of words" (Sinclair, 1991:170), in order to give more information about the use of words in context, taking into consideration information extracted from corpora. This has also been recognized as a useful addition to specialised dictionaries, especially in relation to the user's needs for encoding. However, as explained by L'Homme & Leroyer (2009:259) "there does not seem to be a general agreement as to what types of word combinations should be listed, nor as to how they should be presented in specialised reference works."

A learner's dictionary of science must be a tool for an ongoing learning process where specific collocations and lexical patterns can help non-native speakers who need to produce scientific texts in English. To do this, we propose to use a bottom up model to create an experimental dictionary dealing with verbs used in scientific texts. We pay our attention to the verbal category, as verbs are the centre of the clause which link nodes of specific terminology and are of phraseological interest.

From a classical perspective of terminology, verbs were not considered of interest as they were not proper terminological units. Recent approaches to terminology have shown that not only does the nominal category can have a terminological value, but that other categories, such as adjectives or verbs, can also be domain-specific lexical units — see Lorente (2007, 2009) for more information. According to Lorente (2009:59) verbs are not *per se* terminological units, but can acquire a 'specialised value' in context when their immediate environment also provides specialised knowledge. Lorente (2007) establishes a classification of verbs used in scientific texts: a) *verbos casi-términos* ('near-term verbs'), such as *to ionize*; b) *verbos fraseológicos* ('phraseological verbs'), such as *to codify* (i.e *codify a protein*); c) *verbos de relación lógica* ('verbs of logic relation'), such as *to present*; d) *verbos performativos del discurso* ('verbs performative of discourse'), such as *to conclude*.

As it can be observed by Lorente's classification, in most cases, the 'specialised value' of a verb is determined by the company it keeps. As Hanks (2010b) establishes, taking into consideration Sinclair's distinction (Sinclair, 1991) between the *open-choice principle* and the *idiom principle*, many units have both a terminological tendency (*open-choice principle*) and a phraseological tendency (*idiom principle*). Verbs have mainly a phraseological tendency. It is impossible to know the meaning of some of these verbs without knowing the phraseological context in which the verb is used. This phraseological context is the information to which a learner of the language needs to pay particular attention. The difficulty of the learner of science is in the phraseology being used and not in the designation of a concept. An advanced specialised learner's dictionary must pay special attention to those units with a phraseological tendency. In order to write scientific texts in a foreign language, the learner of the language needs to know the meaning of the specific words used in specific contexts and, as it has been mentioned before, there are many words whose meaning can only be understood by knowing the environment where the word is used.

The DicSci is an advanced learner's dictionary of verbs whose main aim is to give account of the functioning of a verb in an scientific context, showing its phraseological behavior, taking into account its collocates and its textual environment.

3. DicSci – An E-Advanced Learner's Dictionary of Verbs in Science

The lexicographical project *DicSci* starts off from ongoing work that is both theoretical and practical in nature related to two research projects on science corpora coupled with and analyse of the place of scientific usage in advanced learner's dictionaries and the application of the methodology of *collocational networks* and *collocational resonance* (Williams, 1998, 2002, 2003, 2006, 2008a, 2008b, 2008c; Williams & Millon, 2009, 2010) and that of the technique developed by Patrick Hanks — see Hanks (2004, 2006), Hanks & Ježek (2008) for more detailed information —, named *Corpus Pattern Analysis* or CPA and supported by the *Theory of Norms and Exploitations* or *TNE* (Hanks, forthcoming).

On the theoretical side the objective of the lexicographical project is to show how collocational networks, collocational resonance and lexical patterns can assist with understanding not just meaning change, but the carry-over of aspects of meaning from changing contextual environments, and also the relations between the technical and the general lexical units. The practical final outcome is an *E-Advanced Learner's Dictionary of Verbs in Science* (DicSci) built bottom-up using corpus-driven methodologies both for selection of headwords, semantic organisation of the data, representation of norms and exploitations, word syntagmatic and paradigmatic relations and movement of meanings between contexts.

The working methodology is based on the use of collocational networks and collocational resonance. This can be further enhanced by applying *Corpus Pattern Analysis* or CPA. In previous studies (Alonso, 2009; Williams, 1998, 2002, 2006, 2008a, 2008b, 2008c; Williams & Millon, 2009, 2010), these statistically based chains of collocations have been used to demonstrate thematic patterns in texts, as well as means for selecting the lexis for a specialised language dictionary, for observing the movement of meanings between contexts, establishing syntagmatic and paradigmatic relations between units and determining the difference between the 'specialised' and 'general' language use.

The methodology proposed is influenced by John Sinclair's insights into collocation and the idiom principle (Sinclair, 1991), Wittgenstein's approach to prototypes (1953), the work on scientific texts developed by Roe (1977) and the later studies of the phraseology of scientific texts developed by Gledhill (2000), the work on pattern grammar by Hunston & Francis (1999), the study on semantic prosody by Louw (1993, 2000|2008), the theory of Lexical Priming proposed by Hoey (2005). Finally, as has been shown in previous studies (Alonso, 2009; Renau & Alonso, in press), the application of *Corpus Pattern Analysis* proposed by Hanks (2004) for building-up a *Pattern Dictionary of English Verbs* (PDEV)[1] seems to be useful for analysing the normal use of the lexical units in scientific contexts and establishing differences between the general and specialised use of a lexical unit, as well as it can help to improve the dictionary entry as it provides a systematic and very fine analysis of language in use. CPA is a technique which can complement the information given by the collocational networks.

Collocational networks, proposed by Williams (1998, 2002), are statistically based chains of collocations, a web of interlocking conceptual clusters realised in the form of words linked through the process of collocation. The idea that collocations "cluster" forming interwoven meaning networks comes from Phillips (1985). Phillips's aim was the study of metastructure within texts and the notion of 'aboutness'. Williams (1998) considered Phillip's work and hypothesised that "the patterns of co-occurrence forming the collocational networks will be unique to any one sublanguage and serve to define the frames of reference within that sublanguage" (Williams 1998:157). From a high frequency lexical unit, considered as node of the network, the collocates are calculated using a statistical measure, mainly MI or Z-Score, even though other statistical measures can be considered. The collocates are then treated as nodes and the collocates of each collocate is then calculated. The network will be allowed to extend through collocational chains until a point is reached where either no more significant collocates are found or where a word-form that has occurred earlier in the network is encountered. A detailed description of the procedure for the creation of collocational networks is shown in Williams (1998).

It must be taken into account the importance of the statistical measure selected for calculating the more significant collocates of a lexical unit, as different measures will give different results. For instance, Mutual Information displays more rarer items whereas Z-score gives more general collocates — see Church & Hanks (1990) for more information on measuring word association norms. It is also important to bear in mind that the collocational network can vary depending on the form of the lexical unit. For instance, in texts related to Molecular Biology, the environment developed by the use of 'gene' in singular is quite different to the environment of the form in plural:

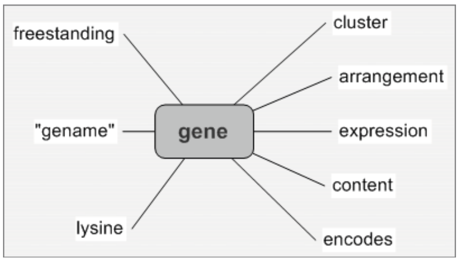


Figure 1: First level of the collocational network of 'gene' extracted from Williams (2008c:140)

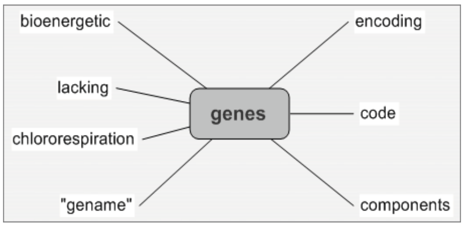


Figure 2: First level of the collocational network of 'genes' extracted from Williams (2008c:140)

Despite the different collocates associated to each form of the lexical unit, the lemmatised network must be also considered in order to have a complete panorama of the total environment of a word.

On the other hand, collocational resonance is also a tool being used at DicSci to show how elements of meaning are carried over from on textual environment to another. The mechanism of collocational resonance has been described in Williams (2008b) and Williams & Millon (2009). The notion of collocational resonance is based on the assumption that language users carry aspects of meaning from previously encountered usage, consciously and subconsciously, subcategorised for topic and genre, coloning the meanings and prosodies in use. This can be mapped by using lexicographical prototypes. For instance, if we consider the word 'culture', one of its meanings is that of *farming*. When 'we culture children', there are pieces of meaning that still carry a resonance of the meaning of 'culture' as *farming*. A detailed explanation of resonance with reference to the word 'probe' can be found in Williams & Millon (2009). Collocational resonance is used to explain particular patterns of usages. It can assist in understanding the movement from general to specialised usage of language, or from specialised to general. It can also help to build up the definition of dictionary entries. In the present

[1] The *Pattern Dictionary of English Verbs* (PDEV) is an ongoing project whose first results are free available on the Internet (http://deb.fi.muni.cz/pdev).

study we concentrate on collocational networks rather than on collocational resonance, as collocational network is the primary tool for building-up the dictionary DicSci.

The third element of the work methodology in compiling DicSci is the use of *Corpus Pattern Analysis* to give a more accurate account of the normal uses of each of the significant collocates which form the collocational network. CPA is a work-in-progress corpus-driven methodology developed by Hanks for 'mapping meanings onto use' (Hanks, 2002). According to Hanks (2010c:590), "a corpus does not show directly what a word means, but it provides evidence on the basis of which meanings can be inferred." It provides evidence on the word use. Most of these uses are highly patterned. Each unique pattern is usually associated with a specific meaning. CPA is a methodology for identifying prototypical syntagmatic patterns with which words in use are associated. As Hanks (2006:1165) explains "a pattern consists of a verb with its valencies, plus semantic values for each valency and other relevant clues, and is associated with an implicature that associates the meaning with the context rather than with the word in isolation." A pattern is based on the structure of English clause roles described in systemic grammar (Halliday, 1961) — subject, predicator, object, complement, adverbial. Each clause role or argument is 'populated' by a set of collocations. The more significant collocates of a verb are usually nouns which share a semantic aspect of meaning. The meaning of a group of collocates is expressed by a semantic type. Using Hanks' words, semantic types represent 'folk concepts.' All semantic types are stored in a hierarchically structure shallow ontology which is continuously under review. The CPA ontology is corpus-driven. There are cases in which the argument slot is populated by one or more lexical items which cannot be grouped together into semantic types; these are considered as lexical sets. In other cases, the semantic type is complemented by a semantic role. The semantic type is an intrinsic property of the collocate, while a semantic role is an extrinsic property assigned by context. For instance, if we consider the verb *to filter*, one normal pattern would be [[Human]] *filter* [[Liquid]]. However, the corpus can show cases in which not all kinds of liquids are filtered but only some specific ones, such as *water*. The pattern in this case would be [[Human]] filter [[Liquid=Water]]. The organisation of semantic types and semantic roles is not easy and it is only by corpus evidence that this task can be achieved. For more detailed information on the general principles of CPA, see Hanks (2004, 2006, 2010a); for an explanation of the CPA ontology, see Ježek & Hanks (2010).

As mentioned before, the DicSci is a corpus-driven dictionary which takes into account the use of words in scientific texts. Therefore, it is obvious that a corpus of scientific texts is needed. To begin with the building-up of the dictionary a corpus was compiled, the BioMed Central corpus (BMC). The BMC is a 33-million-word English language built as part of the Scientext initiative. The Scientext initiative was a project for the creation of comparable corpora carried out by a consortium of three French universities led by the Université de Grenoble 3. The BMC corpus, which is now freely online at the Scientext website[2], stands at 33 million words drawn from 8945 scientific texts from 137 different journals, made freely accessible online by the independent publishing house BioMed Central[3]. The texts have been selected from a number of journals dating from 1997 to 2005. All texts have been formatted according to the TEI guidelines and have been part-of -speech tagged and lemmatized using *Treetagger*[4]. The texts in the BMC corpus encompass a large number of topics and genres, all related to two main areas: biology and medical research. Each text has been informed with XML-TEI annotation to which topic(s) and to which genre is belonged.

The corpus cannot be considered as fully representative of published scientific research, as it is focused on articles related to Biosciences. The distribution of topics and genres is not well-balanced, as stated in Williams & Millon (2009). In the present work, however, the subcategorisation of the corpus has not been exploited. Despite the limitations of the corpus, due to its size the BMC corpus provides adequate data for work on an experimental dictionary such as DicSci. More details about the corpus can be found on the Scientext website.

Finally, the experimental dictionary presented is considered an 'organic' dictionary. It is 'organic' in the sense that it refers to a living dictionary that will organised itself in a natural way thanks to the links between words shown by means of collocational networks. Collocational networks are used for headwords selection, for structuring and classifying verbs together into classes and as means of navigation. This dictionary will ultimately make wide use of mind mapping technology to allow user navigate within the different entries. The dictionary will provide the environment which can link phraseological patterns to the corpus data whilst providing real examples of language in use in specialised contexts. In the following chapter the use of collocational networks for building-up our dictionary is illustrated through the exploration of the verb *to treat.*

4. Collocational networks and dictionary making: the verb to treat

To treat is the 49th most frequent verb in the BMC corpus with 13018 occurrences. The collocational network was created by measuring the most significant collocates of the verb. Due to space restrictions, Figure 3 below shows

[2] http://scientext.msh-alpes.fr/scientext-site/?article30
[3] http://www.biomedcentral.com
[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

only the first level of the collocational network of the verb *to treat*, as the main aim in this paper is to demonstrate the principles, not to expose full networks. This network contains the eight most statistical significant noun collocates of *to treat*, namely *animal*, *rat*, *mouse*, *patient*, *intention*, *control*, *vehicle* and *cell* — showed in red in Figure 3 —, and the first ten most statistical significant verb collocates of each of the nouns. The collocates are calculated by means of Z-score in a span of 5:5, and the collocations that have less than 3 occurrences are kept out. Yet, five verbal collocates were removed from the network, that are *deciduoma-bearing*, *coimmunized*, *frequency-matched*, *transfected*, and *exhaust*. The first four are word-forms not recognized by the *Treetagger tool*, and the last one *exhaust* was removed because in the noun-verb collocation 'vehicle exhaust', *exhaust* correspond to a noun which belongs to the syntagmatic lexical unit *motor vehicle exhaust*.

In total, 54 verb collocates have been considered for the network. Among them, seven are amongst the 100 more frequent verbs in the BMC corpus: *compare*, *express*, *grow*, *include*, *receive*, *stain* and *use*. Moreover, there are eight verbs (without counting *treat*) that are shared by some of the seven noun collocates, namely *anesthetize*, *compare*, *feed*, *immunize*, *inject*, *receive*, *sacrifice*, and *stain* — marked in green in Figure 3.

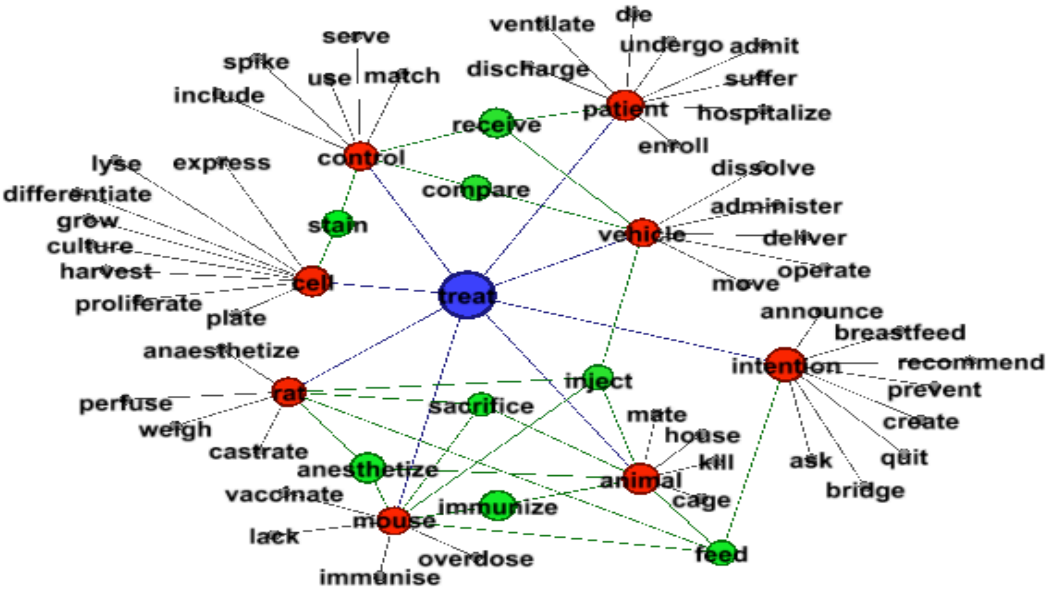


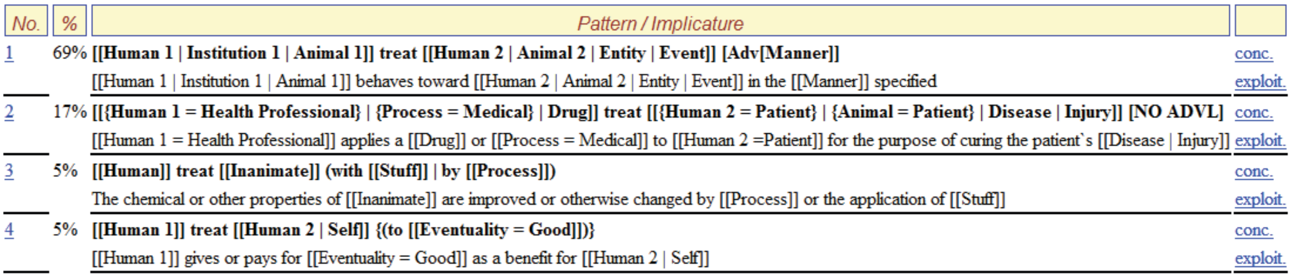
Figure 3: Collocational network from the verb *to treat*

Through the collocational network, verbs that are not in the top 100 verbs list are then introduced. In our illustration, this concerns 47 verbs of the network. Naturally, amongst this set of 'new' verbs, some could have been already enter in the dictionary, as they may have been introduced in a previous analysis. However, not all verbs present in the network will be selected as headwords and considered as entries of the dictionary. Indeed, this depends as well on the frequency.

This brief exemplification illustrates the organic nature of the constitution of the dictionary, which will grow in a natural way, by selecting what is statistically significant in the textual environment of the words. It is through the study of the 100 more frequent verbs that other verbs attested in the BMC corpus will in turn be enter in the dictionary. The constitution of the dictionary follows thus an iterative process: the analysis of one verb of the top-100 verb list leads to the consideration of verbs that are not in this list, and the analysis of one of them leads to the consideration of new verbs, and so on. As mentioned in the previous chapters, collocational networks are a mechanism for headwords selection. It also give a first picture of the environment of scientific texts, showing the most significant lexical units which are 'pivots', — using Hanks' terminology — of the clauses or are the main cognitive nodes that form the texts' framework.

The collocational network brings about a global picture of the node of the network, in this case the verb *to treat*. A lexicographical analysis of the network also show that collocates can be grouped in different conceptual classes. In previous research (Williams & Millon, 2009), Levin's classification of verbs was considered (Levin, 1993). However, this classification does not suit all cases as it has not been built taking into account corpus data. Another option would have been that of using a vast hiercharchical ontology such as WordNet, but as Hanks (2006) points out not all lexical items fit into a hierarchical ontology. The relations between lexical units are not always of the same kind. Indeed, Hanks' point of view has been an inspiration for getting a way to group the different collocates into classes. Moreover, an analysis of the different collocations observed in the network brings about different semantic patterns of usage. These different lexical patterns are determined by using CPA.

In relation to our example, in general texts, four are the CPA patterns established by Hanks, as shown in Figure 4:



<table>
<tr><th>No.</th><th>%</th><th>Pattern / Implicature</th><th></th></tr>
<tr><td>1</td><td>69%</td><td><b>[[Human 1 | Institution 1 | Animal 1]] treat [[Human 2 | Animal 2 | Entity | Event]] [Adv[Manner]]</b></td><td>conc.</td></tr>
<tr><td></td><td></td><td>[[Human 1 | Institution 1 | Animal 1]] behaves toward [[Human 2 | Animal 2 | Entity | Event]] in the [[Manner]] specified</td><td>exploit.</td></tr>
<tr><td>2</td><td>17%</td><td><b>[[{Human 1 = Health Professional} | {Process = Medical} | Drug]] treat [[{Human 2 = Patient} | {Animal = Patient} | Disease | Injury]] [NO ADVL]</b></td><td>conc.</td></tr>
<tr><td></td><td></td><td>[[Human 1 = Health Professional]] applies a [[Drug]] or [[Process = Medical]] to [[Human 2 =Patient]] for the purpose of curing the patient`s [[Disease | Injury]]</td><td>exploit.</td></tr>
<tr><td>3</td><td>5%</td><td><b>[[Human]] treat [[Inanimate]] (with [[Stuff]] | by [[Process]])</b></td><td>conc.</td></tr>
<tr><td></td><td></td><td>The chemical or other properties of [[Inanimate]] are improved or otherwise changed by [[Process]] or the application of [[Stuff]]</td><td>exploit.</td></tr>
<tr><td>4</td><td>5%</td><td><b>[[Human 1]] treat [[Human 2 | Self]] {(to [[Eventuality = Good]])}</b></td><td>conc.</td></tr>
<tr><td></td><td></td><td>[[Human 1]] gives or pays for [[Eventuality = Good]] as a benefit for [[Human 2 | Self]]</td><td>exploit.</td></tr>
</table>

Figure 4: CPA patterns of the verb *to treat* extracted from Hanks' *PDEV*[5]

As can be inferred, the four patterns stand for different meanings of the verb *to treat*. The percentages assigned to patterns show the distribution of the four patterns within the corpus. At first sight, pattern 2 and 3 seem to be more thematically marked, the first related to Medical and the latter related to Chemistry domain. It could be thought that these patterns would also be commonly used in scientific texts. However, by analysing the BMC corpus applying CPA, differences of usage are brought about. The collocational network already shows that not all patterns are always coincident to those patterns distinguished in general texts.

In illustrating our work methodology with the verb *to treat*, using to the BMC corpus, pattern 1 of the verb *to treat* is close to the second CPA pattern (see Figure 4), in that it refers to a medical context. Indeed, in the BMC corpus the following normal pattern is found:

- X treat Y *with* Z
  - [[Human 1 | Human Group]] treat [[Human 2 = Patient | Laboratory Animal = *Rat, Mouse* | Organism= *Cell*]] (with [[Drug= *Vehicle*]])

In this pattern, the different collocates are gathered in different semantic types, as in CPA. By trying to apply CPA to the BMC corpus is clearly not always possible to use the same ontology. The ontology being used in CPA is a corpus-driven shallow ontology created from a general corpus. Many semantic types are not necessary in our case; on the contrary, semantic types that are not considered in CPA ontology are needed for explaining specific uses of a word in Biomedical texts. It is in fact the selection of specific semantic types, semantic roles and lexical sets which makes the difference between the general and specialised use of a lexical unit. For instance, in the pattern shown above, not all animals are treated. The semantic type specifically refers to 'Laboratory Animals'. There is a restriction on what is being treated. In reality, the lexical sets that define a given semantic type change according to each verb. For example, we treat *rats* and *mice*, but we do not treat neither *lion* or *elephant*. Hanks & Ježek (2008) has referred to this change as 'shimmering lexical sets.'

By looking at the concordances of *treat*, a slightly difference between CPA pattern 2 and our pattern 1 can also be detected. Most occurrences of *treat* refer to medical research and not to medical practices. An animal is treated *not for the purpose of being cured*, but *for getting a cure to a disease*. The implicature is not exactly the same.

The collocational network shown in Figure 3, also shows that the collocate *vehicle* is polysemic. Indeed, in the collocational network, the verbal collocates of the nominal collocates of the central verb *treat*, do not necessarily collocates with *treat*, since the nouns have been taken in turn as word-nodes. Thus, collocational networks do not stand for one particular meaning of the verb from which they are built. If we consider the noun *vehicle* — see Figure 3 —, within the occurrences of the collocations (on the lemma level) *vehicle – operate* and *vehicle – move*, the noun *vehicle* denotes a means of transport, whereas within the syntagmatic lexical relations with the verb *treat*, or its other verbal collocates in the network, it is a medical term used to refer to an excipient. Hence the presence in the network of its verbal

[5] http://deb.fi.muni.cz/pdev/?action=patterns&id=treat

collocates *dissolve*, *deliver*, *administer*, *receive* and *inject*. These two meanings are therefore linked, because an excipient serves to 'transport' the active ingredients of a medication. This will lead us to draw two nominal semantic types to which the noun *vehicle* will be attached: 'Transport' and 'Drug'. The verbs *dissolve*, *deliver*, *administer*, *receive* and *inject* are in lexical relation with the semantic type 'Drug', gathering themselves in a verbal conceptual class that we could name 'Giving drugs'. Concerning the conceptual classes in which the verbs of DicSci will be gathered, Framenet is consulted, but, ultimately, the verbal clustering in the dictionary DicSci is based on the specialised contexts of the BMC corpus.

Using CPA has brought about the necessity of using a shallow ontology in order to explain the phraseological tendency of verbs used in science. Indeed, phraseology occupies a main place in language use, notably through the use of collocations. In the lexicon of a given language, there are strong syntagmatic links between words. The phraseology of a given language implies that speaker (or writer), especially a non-native one, could product unnatural speech if he/she uses a 'wrong' word even if it matches the idea to be expressed. Language use is mainly filled with conventional lexical combinations that a native speaker has unconsciously memorised because he/she has already met them during their life. Non-native speakers, who do not have this linguistic experience, would construct their speech according to the semantic compatibility between words, and not to the lexical compatibility between words. Thus, the speaker, especially the non-native one, has to know the phraseology in use within the language in order to produce natural speech. Naturally, inside the same language, lexical preferences may differ notably between the general language and specialised ones, as notably state L'Homme (1998) or Heid & Freibott (1991).

The mechanism used allows conceptual classes that semantically link verbs in the dictionary to grow naturally as new verbs are analysed, and thus eventually split in several sub-classes. This has been illustrated in Williams & Millon (2009). In addition to conceptual classes of verbs, nominal ones are also created, according to the collocational network of the verbs, and notably, through the shared collocates reported in them.

It is important to underline that although networks can be automatically built, the eye of the lexicographer is essential. What we are extracting are potential collocates, only through analysis of the concordance can potential definitions be made. The semantic groupings of verbs or nouns follows the same procedure as, although they do fall together naturally, their interpretation and naming is the work of the lexicographer. Nevertheless, we project to apply the word sense discrimination algorithm written by Millon (2011), as we believe that this processing would help us with this task.

The next step in the creation of DicSci is that of adding the information extracted from the collocational networks and verbal patterns to the entries of the dictionary. For that, the dictionary production software *TshwaneLex*[6] is being used. The *E-Advanced Dictionary of Verbs in Science* is conceived as a virtual dictionary. By using visualisation techniques, the idea is to enter the dictionary by means of the collocational networks and from there go into the verbal patterns, concordances and dictionary entries. The grouping of verbs into classes will also give more options for the user to visualise not only syntagmatic relations but also paradigmatic relations between different lexical units.

5. Conclusions

The first aim of the DicSci project is to build an organic online dictionary of verbs use in sciences which will reflect usage and assist non-native speakers of English with production. In doing so a work methodology based on *collocational networks*, *collocational resonance* and Hanks' *Corpus Pattern Analysis-CPA* is being developed.

In this article, special attention has been paid to the use of collocational networks and application of CPA for building-up the dictionary. Collocational networks provide a natural selection of the main cognitive nodes of scientific texts, show links between lexical units, demonstrate thematic patterns in texts, and facilitate observation of what it is the 'normal' use/s of a specific lexical unit in a scientific context. By taking each collocate at a time, a number of lexico-semantic patterns can be detected. For that, the procedure *Corpus Pattern Analysis* described by Patrick Hanks is used. CPA method allows us to show the central and prototypical uses of a verb in science. By looking at the own output from Patrick Hanks' CPA, the *PDEV*, differences between 'general' and 'specialised' uses can be highlighted. From the patterns, the meaning potentials of the verbs can be inferred in a second stage.

Furthermore, collocational networks and semantic patterns show similarities and differences between the different uses of a lexical unit. Both mechanisms facilitate sense disambiguation of polysemic words. The methodology proposed shows also differences and similarities between different lexical units. Words that that are semantically related can be clustered together naturally in a conceptual class. In this way, both paradigmatic and syntagmatic relations can be illustrated.

The work methodology permits different ways to structure, organise and access the DicSci entries. In this sense, the dictionary is structured and organized according to the collocational networks. Apart from the traditional alphabetically ordering of entries, in the DicSci each central node of a network, which

[6] http://tshwanedje.com/tshwanelex/

corresponds to a verb, is an access to the entries of the dictionary. Each verbal collocate can also be a central node of another network and, therefore, another way to enter the dictionary. At the same time, other collocates, such as nouns or adjectives, can also be a means of access. The groupings of verbs will also permit access to the main verbal lexical units. The dictionary is both semasiologically and onomasiologically conceived.

The DicSci is an ongoing bottom-up, corpus-driven dictionary which describes how verbs are used in science. It is an organic dictionary in the sense that it is being developed in a natural and continuous process. It is dynamic, a moving system. Each collocational network can bring about new uses and new relations between other verbs and lexical units which have been already included in the dictionary. The relations between the units are continuously in motion.

In this paper, we have explored the first stage of the building-up of the dictionary which affects the global organisation and structure of the dictionary, the selection of headwords, the establishment of classes and the demonstration of semantic patterns. Further development is needed in relation to the definition and naming of conceptual classes and the microstructure of each entry. In a second stage, it is also expected to apply the mechanism of collocational resonance to assist in a better understanding of the movement from general to specialised usage of language, or from specialised to general.

The final aim of the DicSci project is to compile a dictionary which provides a way to explain not only the terminological tendency of words used in science, but also the phraseological tendency. The information included will help non-native speakers of English who need to produce scientific texts in English to improve their communication skills at different levels.

6. Acknowledgements

This paper was carried out during a postdoctoral research stay by one of the authors at the Equipe LiCoRN-Laboratoire HCTI of the Université Bretagne-Sud, directed by Prof. Geoffrey Williams, in the framework of the National Mobility Programme of Human Resources of the R+D National Programme 2008-2011, financed by the Spanish Ministry of Education. This research has also been funded by the European Project Metricc, and the Spanish National Project HUM2009-07588/FILO, supported by the Spanish Ministry of Education and Science.

7. References

Alonso, A. (unpublished 2009). Características del léxico del medio ambiente y pautas de representación en el diccionario general. PhD Thesis. Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra, Barcelona.

Bergenholt, H., Tarp, S. (1995). *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries.* Benjamins Translation Library 12. Amsterdam/Philadelphia: John Benjamins.

Cabré, M.ª T. (1999). La terminología. Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra.

Ciapuscio, G. (2003). *Textos especializados y terminología*. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra.

Cowie, A.P. (2002). *English Dictionaries for Foreign Learners. A History*. Oxford: Oxford University Press.

Cowie, A.P. (2009). The Oxford History of English Lexicography. Volumen II. Oxford: Clarendon.

Church, K., Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational linguistics*, 16(1), pp. 22-29.

Fuertes-Olivera, P.A. (2009). Specialised Lexicography for Learners: Specific Proposals for the Construction of Pedagogically oriented Printed Business Dictionaries. *Hermes – Journal of Language and Communication Studies*, 42, pp. 167-188.

Fuertes-Olivera, P.A. (ed.) (2010). *Specialised Dictionaries for Learners.* Lexicographica Series Maior, 136. Berlin/New York: De Gruyter.

Gledhill, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.

Halliday, M.A.K. (1961). Categories of the Theory of Grammar. *Word*, 17, pp. 241-292.

Hanks, P. (2002). Mapping Meaning onto Use. In M.-H-Corréard (ed.) *Lexicography and Natural Language Processing: a Festschrift in honour of B. T. S. Atkins*. United-Kingdom: Euralex, Göteborg University, pp. 156-198.

Hanks, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.

Hanks, P. (2006). The Organization of the lexicon: Semantic Types and Lexical Sets. In C. Marello *et al.* (eds.) *Proceedings of the XII EURALEX International Congress*. Torino: Università di Torino, pp. 1165-1168.

Hanks, P. (2010a). How People Use Words to Make Meanings. In B. Sharp, M. Zock (eds.) *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2010*. In conjunction with ICEIS 2010, Funchal, Madeira, Portugal, pp. 3-13.

Hanks, P. (2010b). Terminology, Phraseology, and Lexicography. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden, Pays Bas: Fryske Akademy.

Hanks, P. (2010c). Compiling a Monolingual Dictionary for Native Speakers. *Lexikos* 20, 580-598.

Hanks, P. (forthcoming). *Lexical Analysis: Norms and Exploitations*. Massachusetts: The MIT Press.

Hanks, P., Ježek, E. (2008). Shimmering Lexical Sets. In E. Bernal, J. DeCesaris, J. (eds.) *Proceedings of the*

XIII Euralex International Congress. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra, pp. 391-402.

Heid, U., Freibott, G. (1991). Collocations dans une base de données terminologique et lexicale. *Meta,* 36(1), pp. 77-91.

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

Hunston, S., Francis, G. (1999). *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Hunston, S., Sinclair, J. (2003). A local grammar of evaluation. In S. Hunston, G. Thompson (eds.) *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press, pp. 74-101.

Ježek, E., Hanks, P. (2010). What lexical sets tell us about conceptual categories. In *Corpus Linguistics and the Lexicon, Special issue of Lexis, E-Journal in English Lexicology*, 4, pp. 7-22.

L'Homme, M.-C. (1998). Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale. In *Proceedings of the VIII EURALEX International Congress*. Liège, Belgium, pp. 513-522.

L'Homme, M.-C. (2005). Sur la notion de «terme». *Meta: Translators' Journal* 50(4), pp. 1112-1132. On line: http://id.erudit.org/iderudit/012064ar

L'Homme, M.-C. & Leroyer, P. (2009). Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology* 15(2), pp. 258-283.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago: University of Chicago Press.

Lorente, M. (2007). Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació. In M. Lorente et al. (eds.) *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví. Vol. 2: De deixebles 2*. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra; Documenta Universitaria, pp. 365-380.

Lorente, M. (2009). Verbos y fraseología en los discursos de especialidad. In M. Casas, R. Márquez (ed.) *XI Jornadas de Lingüística: homenaje al profesor José Luis Guijarro Morales (Cádiz, 22 y 23 de abril de 2008)*. Cádiz: Universidad de Cádiz. Servicio de Publicaciones, pp. 55-84.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker (ed.) *Text and Technology.* Amsterdam: John Benjamins, pp. 157-76.

Louw, B. (2000|2008). Contextual Prosody Theory: bringing Semantic Prosodies to Life. In C. Heffer, H. Sauntson (eds.) *Words in Context: A Tribute to John Sinclair on his Retirement*. CD-ROM: English Language Research Discourse Analysis Monograph No. 18. Reprinted in online journal *Texto* (2008): http://www.revue-texto.net/index.php?id=124.

Meyer, I. (2000). Computer Words in Our Everyday Lives: How are they interesting for terminography and lexicography? In U. Heid *et al.* (eds.) *Proceedings of IX EURALEX International Conference 2000*. Stuttgart: Universität Stuttgart, pp. 39-57.

Meyer, I., Mackintosh, K., Varantola, K. (1997). Exploring the reality of *virtual*: on the lexical implications of becoming a knowledge society. *Lexicology,* 3(1), pp. 129-163.

Million, C. (unpublished 2011). Acquisition automatique de relations lexicales désambiguïsées à partir du Web. PhD Thesis. Université de Bretagne-Sud, Lorient.

Moulin, A. (1983). LSP Dictionaries for EFL Learners. In R. R. K. Hartmann (ed.) *Lexicography: Principles and Practice*. London: Academic Press, pp. 144-152.

Phillips, M. (1985). Aspects of Text Structure: An investigation of the lexical Organisation of Text, Amsterdam, North Holland.

Renau, I., Alonso, A. (in press). Using Corpus Pattern Analysis for the Spanish Learner's Dictionary DAELE (Diccionario de aprendizaje del español como lengua extranjera). In *Proceedings Corpus Linguistics Conference 2011*. Birmingham: University of Birmingham.

Roe, P. (unpublished 1977). The notion of difficulty in Scientific Text. PhD thesis. University of Birmingham, Birmingham.

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

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-knowledge General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Lexicographica Series Maior 134. Tübingen: Max Niemeyer Verlag.

ten Hacken, P. (2008). Prototypes and discreteness in terminology. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress. Barcelona, 15-19 july 2008.* Papers de l'IULA. Sèrie Activitats. 20. Barcelona: Documenta Universitaria. Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra.

Williams, G. (1998). Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics*, 3(1), pp. 151-171.

Williams, G. (2002). In search of representativity in specialised corpora: categorisation through collocation. *International Journal of Corpus Linguistics*, 7(1), pp. 43-64.

Williams, G. (2003). From meaning to words and back: Corpus linguistics and specialised lexicography. *Asp, la revue du GERAS* 39-40, pp. 91-106. On line: http://asp.revues.org/1320.

Williams, G. (2006). Advanced ESP and the Learner's Dictionary. In C. Marello *et al.* (eds.) *Proceedings of the XII EURALEX International Congress*. Torino: Università di Torino, pp. 795-801.

Williams, G. (2008a). Verbs of Science and the Learner's Dictionary. In J. DeCesaris, E. Bernal (eds.) *Proceedings of the XIII Euralex International Congress. Barcelona, 15-19 july 2008. Papers de l'IULA. Sèrie Activitats*. 20. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra; Documenta Universitaria.

Williams, G. (2008b). The Good Lord and his works: A corpus-based study of collocational resonance. In S. Granger, F. Meunier (eds.) *Phraseology: an interdisciplinary perspective*. Amsterdam: John Benjamins, pp. 159-174.

Williams, G. (2008c). Les corpus et le dictionnaire dans les langues scientifiques. In F. Maniez et al. (eds.) *Corpus et dictionnaires de langues de spécialité*. Grenoble: Presses Universitaires de Grenoble.

Williams, G., Millon, C. (2009). The General and the Specific: Collocational resonance of scientific language. In *Proceedings of the Corpus Linguistics Conference CL2009, 20-23 July 2009*. Liverpool: University of Liverpool.

Williams, G., Millon, C. (2010). Going organic: Building an experimental bottom-up dictionary of verbs in science. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden, Pays Bas: Fryske Akademy, pp. 1251-1257.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

Lexical Profiling for Arabic

Mohammed Attia, Pavel Pecina, Lamia Tounsi, Antonio Toral and Josef van Genabith
School of Computing
Dublin City University, Dublin, Ireland
E-mail: {mattia, ppecina, atoral, ltounsi, josef}@computing.dcu.ie

Abstract

We provide lexical profiling for Arabic by covering two important linguistic aspects of Arabic lexical information, namely morphological inflectional paradigms and syntactic subcategorization frames, making our database a rich repository of Arabic lexicographic details. First, we provide a complete description of the inflectional behaviour of Arabic lemmas based on statistical distribution. We use a corpus of 1,089,111,204 words, a pre-annotation tool, knowledge-based rules, and machine learning techniques to automatically acquire lexical knowledge about words' morpho-syntactic attributes and inflection possibilities. Second, we automatically extract the Arabic subcategorization frames (or predicate-argument structures) from the Penn Arabic Treebank (ATB) for a large number of Arabic lemmas, including verbs, nouns and adjectives. We compare the results against a manually constructed collection of subcategorization frames designed for an Arabic LFG parser. The comparison results show that we achieve high precision scores for the three word classes. Both morphological and syntactic specifications are combined and connected in a scalable and interoperable lexical database suitable for constructing a morphological analyser, aiding a syntactic parser, or even building an Arabic dictionary. We build a web application, AraComLex (Arabic Computer Lexicon), available at: http://www.cngl.ie/aracomlex, for managing and maintaining the standardized and scalable lexical database.

Keywords: Arabic; subcategorization frames; morphological analysis; morphological paradigms

1. Introduction

In a typical dictionary entry of a word, it is expected to find basic information pertaining to the word's morphology (possible inflections) and syntax (part of speech, whether it is transitive or intransitive, in the case of verbs, and what prepositions it can co-occur with). Yet, existing Arabic dictionaries have several limitations. Most of them do not rely on a corpus for attesting the validity of their entries (as in a COBUILD approach (Sinclair, 1987)), but they typically include either refinements, expansions, corrections, or organisational improvements over the previous dictionaries. Therefore, they tend to include obsolete words not in contemporary use. Furthermore, they often do not explicitly state all the possible inflection paradigms, and they do not provide sufficient syntactic information on word's obligatory combinations (or argument list).

The aim here is to attempt to resolve these shortcomings by automatically providing a complete description of the inflectional and syntactic behaviour of Arabic lexical entries based on statistical distribution in treebanks and un-annotated corpora. The work described in this paper is divided into two major parts. The first is focused on examining the statistical distribution of inflection paradigms for lexical entries in a large corpus pre-annotated with MADA (Roth et al., 2008), a tool which performs morphological analysis and disambiguation using the Buckwalter morphological analyser (Buckwalter, 2004) and machine learning. The second is related to the automatic extraction of syntactic information, or subcategorization frames, from the Arabic Treebank (ATB) (Maamouri and Bies, 2004).

To the best of our knowledge, this is the first attempt at extracting subcategorization frames from the ATB. The subcategorization requirements of lexical entries are important type lexical information, as they indicate the argument(s) a predicate needs in order to form a well-formed syntactic structure. Yet producing such resources by hand is costly and time consuming. Moreover, as Manning (1993) indicates, dictionaries produced by hand will tend to lag behind real language use because of their static nature. Therefore a complete, or at least complementary, automatic process is highly desirable.

This paper is structured as follows. In the introduction we describe the motivation behind our work. We differentiate between Modern Standard Arabic (MSA), the focus of this research, and Classical Arabic (CA) which is a historical version of the language. We briefly explain the current state of Arabic lexicography and describe how outdated words are still abundant in current dictionaries. Then we outline the Arabic morphological system to show what layers and tiers are involved in word derivation and inflection. In Section 2, we present the results obtained to date in building and extending the lexical database using a data-driven filtering method and machine learning techniques. We also explain how we use knowledge-based pattern matching in detecting and extracting broken plural forms. In Section 3, we explain the method we followed in extracting and evaluating the subcategorization frames for Arabic verbs, nouns and adjectives. In Section 4, we describe AraComLex, a web application we built for curating and combining our lexical resources. Finally, Section 5 gives the conclusion.

1.1 Modern Standard Arabic vs. Classical Arabic

Modern Standard Arabic (MSA), the subject of our research, is the language of modern writing, prepared speeches, and the language of the news. It is the language universally understood by Arabic speakers around the world. MSA stands in contrast to both Classical Arabic (CA) and vernacular Arabic dialects. CA is the language which originated in the Arabian Peninsula centuries before the emergence of Islam and continued to be the standard language until the medieval times. CA continues to the present day as the language of religious teaching, poetry, and scholarly literature. MSA is a direct descendent of CA and is used today throughout the Arab World in writing and in formal speaking (Bin-Muqbil, 2006).

MSA is different from CA at the lexical, morphological, and syntactic levels (Watson, 2002; Elgibali and Badawi, 1996; Fischer, 1997). At the lexical level, there is a significant expansion of the lexicon to cater for the needs of modernity. New words are constantly coined or borrowed from foreign languages while many words from CA have become obsolete. Although MSA conforms to the general rules of CA, MSA shows a tendency for simplification, and modern writers use only a subset of the full range of structures, inflections, and derivations available in CA. For example, Arabic speakers no longer strictly abide by case ending rules, which led some structures to become obsolete, while some syntactic structures which were marginal in CA started to have more salience in MSA. For example, the word order of object-verb-subject, one of the classical structures, is rarely found in MSA, while the relatively marginal subject-verb-object word order in CA is gaining more weight in MSA. This is confirmed by Van Mol (2003) who pointed out that MSA word order has shifted balance, as the subject now precedes the verb more frequently, breaking from the classical default word order of verb-subject-object.

1.2 The Current State of Arabic Lexicography

Until now, there is no large-scale lexicon (computational or otherwise) for MSA that is truly representative of the language. Al-Sulaiti (2006) emphasises that existing dictionaries are not corpus-based. Ghazali and Braham (2001) stress the need for new dictionaries based on an empirical approach that makes use of contextual analysis of modern language corpora. They point out the fact that traditional Arabic dictionaries are based on historical perspectives and that they tend to include obsolete words that are no longer in current use. The inclusion of these rarities inevitably affects the representativeness of dictionaries and marks a significant bias towards historical or literary forms. In recent years, some advances have been made (Van Mol, 2000; Boudelaa and Marslen-Wilson, 2010), but they are not enough in terms of size or the breadth of linguistic description.

The Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) is widely used by the Arabic NLP research community. It is a *de facto* standard tool, and has been described as the "most respected lexical resource of its kind" (Hajič et al., 2005). It is designed as a main database of 40,648 lemmas supplemented by three morphological compatibility tables used for controlling affix-stem combinations. Other advantages of BAMA are that it provides information on the root, reconstructs vowel marks and provides an English glossary. The latest version of BAMA is renamed SAMA (Standard Arabic Morphological Analyzer) version 3.1 (Maamouri et al., 2010).

Unfortunately, there are some drawbacks in the SAMA lexical database that raise questions for it to be a truthful representation of MSA. We estimate that about 25% of the lexical items included in SAMA are outdated based on our data-driven filtering method explained in Section 2.2.1. SAMA suffers from a legacy of heavy reliance on older Arabic dictionaries, particularly Wehr's Dictionary (Wehr Cowan, 1976), in the compilation of its lexical database.

Therefore, there is a strong need to compile a lexicon for MSA that follows modern lexicographic conventions (Atkins and Rundell, 2008) in order to make the lexicon a reliable representation of the language and to make it a useful resource for NLP applications dealing with MSA. Our work represents a further step to address this critical gap in Arabic lexicography. We use a large corpus of more than one billion words to automatically create a lexical database for MSA. We enrich the lexicon with syntactic information by extracting subcategorization frames and significant preposition collocates from the ATB.

1.3 Arabic Morphological System

Arabic morphology is well-known for being rich and complex. Arabic morphology has a multi-tiered structure where words are originally derived from roots and pass through a series of affixations and clitic attachments until they finally appear as surface forms. Morphotactics refers to the way morphemes combine together to form words (Beesley and Karttunen, 2003). Generally speaking, morphotactics can be concatenative, with morphemes either prefixed or suffixed to stems, or non-concatenative, with stems undergoing internal alterations to convey morpho-syntactic information (Kiraz, 2001). Arabic is considered as a typical example of a language that employs both concatenative and non-concatenative morphotactics. For example, the verb استعملوها {istaEomaluwha[1] 'they-used-it' and the noun والاستعمالات wAl{istiEomAlAt 'and-the-uses' both originate from the root عمل Eml.

Figure 1 shows the layers and tiers embedded in the representation of the Arabic morphological system. The derivation layer is non-concatenative and opaque in the sense that it is a sort of abstraction that affects the choice of a part of speech (POS), and it does not have a direct explicit surface manifestation. By contrast, the inflection

[1] All examples are written in Buckwalter Transliteration.

layer is more transparent. It applies concatenative morphotactics by using affixes to express morpho-syntactic features. We note that verbs at this level show what is called 'separated dependencies' which means that some prefixes determine the selection of suffixes.

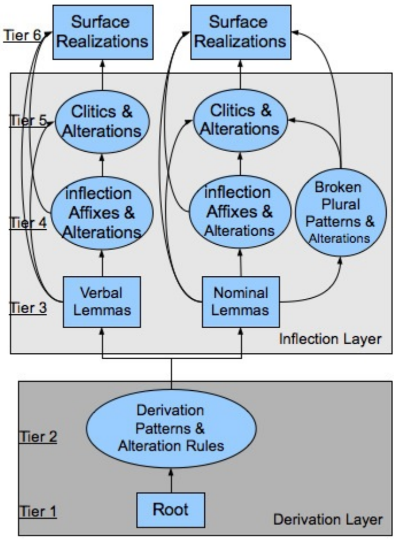


Figure 1: Arabic Morphology's Multi-tier Structure

In the derivational layer Arabic words are formed through the amalgamation of two tiers, namely root and pattern. A root is a sequence of three consonants and the pattern is a template of vowels with slots into which the consonants of the root are inserted. This process of insertion is called interdigitation (Beesley and Karttunen, 2003). An example is shown in Table 1.

Root	درس drs			
POS	V	V	N	N
Pattern	$R_1aR_2aR_3a$	$R_1aR_2R_2aR_3a$	$R_1AR_2iR_3$	$muR_1aR_2\sim iR_3$
Stem	darasa 'study'	darrasa 'teach'	dAris 'student'	mudar~is 'teacher'

Table 1. Root and Pattern Interdigitation

2. Extending the Existing Lexicon

In this section, we describe the small-scale, manually-constructed lexical resources that we had, and how we managed to significantly extend these resources. We explain how we filter out obsolete words, how we use machine learning to acquire knowledge on morphological paradigms (or continuation classes) for new entries, and how we extract broken plural forms from our corpus. The corpus we use contains 1,089,111,204 words, consisting of 925,461,707 words from the Arabic Gigaword (Parker et al., 2009), in addition to 163,649,497 words from news articles we collected from the Al-Jazeera web site.[2]

2.1 Existing Lexical Resources

There are three key components in the Arabic morphological system: root, pattern and lemma. For accommodating these components, we acquire three lexical databases: one for lemmas, one for word patterns, and one for lemma-root lookup. The lemma database is collected from Attia (2006) which was developed manually. It includes 5,925 nominal lemmas (nouns and adjectives) and 1,529 verb lemmas. The advantage of the lemma entries in this resource is that they are fully specified with necessary morpho-syntactic information. In addition to the usual specification of gender, number and person, it provides information on continuation classes for nominals (as shown in Table 2), whether the noun indicates a human or non-human entity. For verbs it gives details on the transitivity, whether the passive voice is allowed or not, and whether the imperative mood is allowed or not.

We automatically create the lemma-root lookup database relying on the SAMA database. We manually developed a database for Arabic patterns that includes 490 patterns (456 for nominals and 34 for verbs). These patterns can be used as indicators of the morphological inflectional and derivational behaviour of Arabic words. Patterns are also powerful in the abstraction and coarse-grained categorisation of word forms.

2.2 Extending the Lexical Database

In extending our lexicon, we rely on Attia's manually-constructed lexicon (Attia, 2006) and the lexical database in SAMA 3.1 (Maamouri et al., 2010). Creating a lexicon is usually a labour-intensive task. For instance, Attia took three years in the development of his morphology, while SAMA and its predecessor, BAMA, were developed over more than a decade, and at least seven people were involved in updating and maintaining the morphology.

Our objective here is to automatically extend Attia's lexicon (Attia, 2006) using SAMA's database. In order to do this, we need to solve two problems. First, SAMA suffers from a legacy of obsolete entries and we need to filter out these outdated words, as we want to enrich the lexicon only with lexical items that are still in current use. Second, Attia's lexicon requires features (such as humanness for nouns and transitivity for verbs) that are not provided by SAMA, and we want to automatically induce these features.

2.2.1 Lexical Enrichment

To address the first problem, we use a data-driven filtering method that combines open web search engines and our pre-annotated corpus. Using frequency statistics[3] on lemmas from three web search engines (Al-Jazeera,[4] Arabic Wikipedia,[5] and the Arabic BBC website[6]), we find that 7,095 lemmas in SAMA have zero hits.

[2] http://aljazeera.net/portal. Collected in January 2010.

[3] Statistics were collected in January 2011.

[4] http://aljazeera.net/portal

[5] http://ar.wikipedia.org

[6] http://www.bbc.co.uk/arabic/

	Masculine Singular	Feminine Singular	Masculine Dual	Feminine Dual	Masculine Plural	Feminine Plural	Continuation Class
1	معلم muEal~im, 'teacher'	معلمة muEal~imap	معلمان muEal~imAn	معلمتان muEal~imatAn	معلمون muEal~imuwn	معلمات muEal~imAt	F-Mdu-Fdu-Mpl-Fpl
2	طالب TAlib, 'student'	طالبة TAlibap	طالبان TAlibAn	طالبتان TAlibatAn	-	طالبات TAlibAt	F-Mdu-Fdu-Fpl
3	تحضيري, taHoDiyriy~, 'preparatory'	تحضيرية taHoDiyriy~ap	تحضيريان taHoDiyriy~An	تحضيريتان taHoDiyriy~atAn	-	-	F-Mdu-Fdu
4	-	بقرة baqarap 'cow'	-	بقرتان baqaratAn	-	بقرات baqarAt	Fdu-Fpl
5	تنازل, tanAzul 'concession'	-	-	-	-	تنازلات tanAzulAt	Fpl
6	-	ضحية DaHiy~ap 'victim'	-	ضحيتان DaHiy~atAn	-	-	Fdu
7	محض maHoD 'mere'	محضة maHoDap	-	-	-	-	F
8	امتحان {imotiHAn, 'exam'	-	امتحانان {imotiHAnAn	-	-	امتحانات {imotiHAnAt	Mdu-Fdu
9	طيار Tay~Ar, 'pilot'	-	طياران Tay~ArAn	-	طيارون Tay~Aruwn	-	Mdu-Mpl
10	كتاب kitAb, 'book'	-	كتابان kitAbAn	-	-	-	Mdu
11	ديمقراطي diymuqoratiy~, 'democrat'	-	-	-	ديمقراطيون diy-muqoratiy~uwn	-	Mpl
12	خروج xuruwj, 'exiting'	-	-	-	-	-	NoNum
13	مباحث mabAHiv, 'investigators'	-	-	-	-	-	Irreg_pl

Table 2: Arabic Continuation Classes based on the inflection grid

Frequency statistics from our corpus show that 3,604 lemmas are not used in the corpus at all, and 4,471 lemmas occur less than 10 times. Combining frequency statistics from the web and the corpus, we find that there are 29,627 lemmas that returned at least one hit in the web queries and occurred at least 10 times in the corpus. Using a threshold of 10 occurrences here is discretionary, but the aim is to separate the stable core of the language from instances where the use of a word is perhaps accidental or somewhat idiosyncratic. We consider the refined list as representative of the lexicon of MSA as attested by our statistics.

No.	Classes	Features	P	R	F
	Nominals				
1	Continuation Classes: 13 classes	number, gender, case, clitics	0.62	0.65	0.63
2	Human: yes, no, unspecified		0.86	0.87	0.86
3	POS: noun, adjective		0.85	0.86	0.85
	Verbs				
4	Transitivity: transitive, intransitive	number, gender, person, aspect, mood, voice, clitics	0.85	0.85	0.84
5	Allow passive: yes, no		0.72	0.72	0.72
6	Allow imperative: yes, no		0.63	0.65	0.64

Table 3: Results of the Classification Experiments.

2.2.2 Feature Enrichment

To address the second problem, we use a machine learning classification algorithm, the Multilayer Perceptron (Haykin, 1998). The main idea of machine learning is to automatically learn complex patterns from existing (training) data and make intelligent decisions on new (test) data.

In our case, we have a seed lexicon (Attia, 2006) with lemmas manually annotated with classes, and we want to build a model for predicting the same classes for each new lemma added to the lexicon. The classes (second column in Table 3) for nominals are continuation classes (or inflection paths), the semantico-grammatical feature of humanness, and POS (noun or adjective). The classes for verbs are transitivity, allowing the passive voice, and allowing the imperative mood. From our seed lexicon we extract two datasets of 4,816 nominals and 1,448 verbs. We feed these datasets with frequency statistics from our pre-annotated corpus and build the statistics into a vector grid. The features (third column in Table 3) for nominals are number, gender, case and clitics; for verbs, number, gender, person, aspect, mood, voice, and clitics. For the implementation of the machine learning algorithm, we use the open-source application Weka version 3.6.4.7[7]. We split each dataset into 66% for training and 34% for testing. We conduct six classification experiments to provide the classes that we need to include in our lexical database. Table 3 gives the results of the experiments in terms of precision, recall, and f-measure.

The results show that the highest f-measure scores (above 80%) are achieved for 'Human', 'POS', and 'Transitivity'. Typically one would assume that these features are hard to predict with any reasonable

[7] http://www.cs.waikato.ac.nz/ml/weka/

accuracy without taking the context into account. It was surprising to obtain such good prediction results based on statistics on morphological features alone. We also note that the f-measure for ‘Continuation Classes’ is comparatively low, but considering that here we are classifying for 13 classes, the results are in fact quite acceptable. Using the machine learning model, we annotate 12,974 new nominals and 5,034 verbs.

2.3 Handling Broken Plurals

Broken plurals are an interesting phenomenon in Arabic where the plural is formed not through regular suffixation, but by changing the word pattern. In our seed lexicon (Attia, 2006), we have 950 broken plurals which were collected manually and clearly tagged. In SAMA, however, broken plurals are rather poorly handled. SAMA does not mark broken plurals as “plurals” either in the source file or in the morphology output. There is no straightforward way to automatically collect the list of all broken plural forms from SAMA. For example, the singular form جانب jAanib “side” and the broken plural jawAnib “sides” are analysed as in (1) and (2) respectively.

```
(1) <lemmaID>jAnib_1</lemmaID>
<voc>jAnib</voc> <pos>jAnib/NOUN</pos>
<gloss>side/aspect</gloss>
(2) <lemmaID>jAnib_1</lemmaID>
<voc>jawAnib</voc> <pos>jawAnib/NOUN</pos>
<gloss>sides/aspects</gloss>
```

The only tags that distinguish the singular from the broken plural form are the gloss (or translation) and voc (or vocalisation). We also note that MADA passes this problem on unsolved, and broken plurals are all marked num=s, meaning that the number is singular. We believe that this shortcoming can have a detrimental effect on the performance of any syntactic parser based on such data.

To extract broken plurals from our large MSA corpus (which is annotated with SAMA tags), we rely on the gloss of entries with the same LemmaID. We use Levenshtein Distance which measures the similarity between two strings. For example, using Levenshtein Distance to measure the difference between “sides/aspects” and “side/aspect” will give a distance of 2. When this number is divided by the length of the first string, we obtain 0.15, which is within a threshold (here set to 0.4). Thus the two entries pass the test as possible broken plural candidates. Using this method, we collect 2,266 candidates. We believe, however, that many broken plural forms went undetected because the translation did not follow the assumed format. For example, the word حرب harb has the translation “war/warfare” while the plural form huruwb has the translation “wars”.

To validate the list of candidates, we use Arabic word pattern matching. For instance, in the jAnib example above, the singular form (vocalisation) follows the pattern fAEil (or the regular expression .A.il) and the plural form follows the pattern fawAEil (or .awA.i.). In our manually developed pattern database we have fawAEil as a possible plural pattern for fAEil. Therefore, the matching succeeds, and the candidate is considered as a valid broken plural entry. We compiled a list of 135 singular patterns that choose from a set of 82 broken plural patterns. The choice, however, is not free, but each singular form has a limited predefined set of broken plural patterns to select from. From the list of 2,266 candidates produced by Levenshtein Distance, 1,965 were validated using the pattern matching, that is 87% of the instances. When we remove the entries that are intersected with our 950 manually collected broken plurals, 1,780 forms are left. This means that in our lexicon now we have a total of 2,730 broken plural forms.

There are some insights that can be gained from the statistics on Arabic plurals in our corpus. The corpus contains 5,570 lemmas which have a feminine plural suffix, 1,942 lemmas with a masculine plural suffix (out of these 1,273 forms intersect with the feminine plural suffix), and about 1,965 lemmas with a broken plural form. This means that the broken plural formation in Arabic is as productive as the regular plural suffixation. Currently, we cannot explain why the feminine plural suffix enjoys this high preference, but we can point to the fact that masculine plural suffixes are used almost exclusively with the natural gender, while the feminine plural suffix, as well as broken plurals, are used liberally with the grammatical gender in addition to the natural gender.

3. Automatic Extraction of Subcategorization Frames

The encoding of syntactic subcategorization frames is an essential requirement in the construction of computational and paper lexicons alike. In English, the construction and extraction of subcategorization frames received a lot of attention, one example is the specialized lexicon COMLEX (Grishman et al., 1994) which is an extensive computational lexicon containing syntactic information for approximately 38,000 English headwords, with detailed information on subcategorization, containing 138 distinct verb frames for 5,662 active verbs lemmas.

For Arabic, the attention has been directed, almost exclusively, to the construction and automatic extraction of semantic roles (Palmer et al., 2008; Attia et al. 2008). Semantic roles are related to syntactic functions and surface phrase structures, but the three are at totally different and distinct layers of analysis. Grammatical functions are in the intermediary position between phrase structures and semantic roles. It is a major concept in semantic role labelling to make greater level of generalization. There is an emphasis on that the semantic labels do not vary in different syntactic constructions (Palmer et al., 2008). For example, the Arabic verb لاحظ lAHaZ "noticed" has two subcategorization frames: <subj,obj> for لاحظ الفرق lAHaZa Al-faroq "He noticed the difference" and <subj,comp> for لاحظ أن المحصول ينقص lAHaZa >an~a Al-maHoSuwla yanoquS "He noticed that the crop is decreasing" Yet, in the Arabic Propbank annotation[8] both frames have the same roleset:

Arg0: observer
Arg1: thing noticed or observed

To our knowledge, the only resource that currently exists for Arabic subcategorization frames is the lexicon manually developed for the Arabic LFG Parser (Attia, 2008). It is published as an open-source resource under the GPLv3 license[9]. It contains 64 frame types, 2,709 lemmas types, and 2,901 lemma-frame types, averaging 1.07 frames per lemma. The resource incorporates control information and details of specific prepositions with obliques. We use this resource in the evaluation of our automatically induced lexicon of Arabic subcategorization frames.

3.1 LFG Subcategorization Frames

The LFG syntactic theory (Dalrymple, 2001) distinguishes between governable (subcategorizable) and non- governable (non-subcategorizable) grammatical functions. The governable grammatical functions are the arguments required by some predicates in order to produce a well-formed syntactic structure, and they are SUBJ(ect), OBJ(ect), OBJ_{Θ}, OBL(ique) $_{\Theta}$, COMP(lement) and XCOMP. The non-governable grammatical functions are not required in the sentence to form a well-formed structure, and they are ADJ(junct) and XADJ. The subcategorization requirements in LFG are expressed in the following format (O'Donovan et al., 2005):

$$\pi<gf_1,gf_2,\ldots gf_n>$$

where π is the lemma (predicate or semantic form) and *gf* is a governable grammatical function. The value of the argument list of the semantic form ensures the well-formedness of the sentence. For example, in the sentence اعتمد الطفل على والدته {iEotamada Al-Tifolu EalaY wAlidati-hi "The child relied on his mother", the verb {iEotamada "to rely" has the following argument structure {iEotamada<(↑SUBJ)(↑OBL$_{>alaY}$)>. By including a subject and an oblique with the preposition >alaY, we ensure that the verb's subcategorization requirements are met and that the sentence is well-formed, or syntactically valid.

3.2 Extracting Subcategorization frames from the Arabic Treebank

We follow here the successful model by the previous language resource extraction efforts for other languages including English (O'Donovan et al., 2005) and German (Rehbein and van Genabith, 2009) taking into consideration the specifics of the Arabic language and the resources available for evaluation. We automatically extract the Arabic syntactic-function based subcategorization frames by utilizing an automatic Lexical-Functional Grammar (LFG) f-structure annotation algorithm for Arabic developed in (Tounsi et al., 2009). The syntactic annotations in the ATB provides explicit information on deep representation in the phrase structure such as dealing with traces in the case of pro-dropped arguments which helped the automatic extraction of subcategorization frames to be complete. After we extract the surface forms we lemmatize all forms by re-analysing all the words using the Buckwalter morphology and then choosing the analysis where the word diacrization and the tag set in the ATB match those in the Buckwalter analysis.

We provide information on the prepositions for obliques, distinguish between active and passive frames, and provide information on the probability score for each frame and the frequency count for each lemma. We extract 240 frame types for 3,295 lemmas types, with 7,746 lemma-frame types (for verbs, nouns and adjectives), averaging 2.35 frames per lemma. We make this resource available under the open-source license GPLv3 [10]. Table 5 shows the list of grammatical functions included in our frames with examples. We compare and evaluate the complete set of subcategorization frames extracted against the manually developed subcategorization frames in the Arabic LFG Parser.

Our extraction algorithm deals with the passive voice and its effect on subcategorization behaviour. We find that in Arabic the passive forms stand at 12% of the active forms compared to 31% in English (O'Donovan et al., 2005), as shown in Table 4. Our explanation of the low frequency of the use of passive in Arabic is that there is a tendency to avoid passive verb forms when the active readings are also possible in order to avoid ambiguity and improve readability. For example, the verb form نظم nZm "organize" can have two readings, one for active and one for passive depending on diacritization, or how the word is pronounced. Therefore, instead of the ambiguous passive form, the alternative syntactic construction تم tam~a "performed/done" + verbal noun is used, giving تم تنظيمه tam~a tanZiymuhu "lit. organizing it has been done / it was organized". One evidence for the validity of our explanation is that the verb tam~a is the seventh most frequent verb in the ATB following كان kAn "be", قال qAla "say", أعلن >aEolana "declare", أكد >ak~ada "confirm", أضاف >aDAfa "add" and اعتبر {iEotabar "consider".

[8] http://verbs.colorado.edu/propbank/framesets-arabic

[9] http://arasubcats-lfg.sourceforge.net

[10] http://arabicsubcats.sourceforge.net

	Active	Passive	Passive %
Arabic verb frames	5,915	681	12
English verb frames	16,000	5,005	31

Table 4: Comparing active and passive subcategorization frames in Arabic and English

		Treebank Tag	Source	Meaning	Example
1	subj	-SBJ	L-T	subject	جاء الوقت jaA'a Al-waqt lit. came the time, “The time came”
2	obj	-OBJ	L-T	object	عرفت الطريق Earaftu Al-Tariyq “I knew the way”
3	obj2	-DTV/-BNF	L-T	secondary object	أعطاه طعاما >aEoTA-hu TaEAmA “gave him food”
4	obl	-CLR	L-T	oblique	اعتمد على والدته {iEotamad EalaY wAlidi-hi “relied on his father”
5	obl2		L	secondary oblique	تنافس معه في السباق tanAfasa maEa-hu fi Al-sibAq “competed with him in the race”
6	obl-betweenAnd		L	oblique for between … and	تنقل بين العراق والكويت tanaq~ala bayona Al-EirAq wa-Al-kwiyt “moved between Iraq and Kuwait”
7	obl-fromTo		L	oblique for from … to	سافر من العراق إلى الكويت sAfara min Al-EirAq <ilA Al-kwiyt “travelled from Iraq to Kuwait”
8	obl-dir	-DIR	T	oblique for direction	شحنها إلى جدة $aHana-hA <ila jad~ap “shipped it to Jeddah”
9	compL		L	light complementizer >an	أمكنه أن يراها >amkana-hu >an yarAhaA “became possible for him to see it”
10	compH		L	heavy complementizer >an~a	أذاع أنهم هربوا >a*aEa >an~a-hum harabuwA “announced that they escaped”
11	vcomp		L	verb complement	بدأ يسقط bada>a yasoquT lit. started fall, “started to fall”
12	xcomp		L	obligatory control	أراد أن يسافر >arAda >an yusAfir “wanted to travel”
13	xcomp-pred	-PRD	T	copular complement	كان مريضا kAna mariDA “was sick”
14	xcomp-verb	(VP)	T	verb complement	same as 11
15	comp-sbar	(SBAR)	T	complement with complementizer	same as 9 and 10
16	comp-nom	(S-NOM)	T	gerund (masdar) complement	نفى علمه بالواقعة nafaY Eilma-hu bi-Al-wAqiEap “denied knowing the incident”
17	comp-s	(S)	T	sentential complement	قال لابد من التشاور qAla lAbud~a min Al-ta$Awur “he said there must be negotiations”

L: LFG Parser, T: Treebank

Table 5: List of Arabic subcategorization frames suffixed with phrase structure information

3.3 Estimating the Subcategorization Probability

In order to estimate the likelihood of the occurrence of a certain argument list with a predicate (or lemma), we compute the conditional probability of subcategorization frames based on the number of token occurrences in the ATB, according to the following formula (O’Donovan et al., 2005);

$$P(ArgList \mid \Pi) = \frac{count(\Pi\langle ArgList\rangle)}{\sum_{i=1}^{n} count(\Pi\langle ArgList_i\rangle)}$$

where ArgList$_1$ … ArgList$_n$ are all the possible argument lists that co-occur with Π. Because of the variations in verbal subcategorization, probabilities are useful for discriminating prominent frames from accidental ones. An example is shown in table 6 for the verb شاهد $Ahada “watch” which has a frequency of 40 occurrences in the ATB.

Lemma with argument list	Conditional Probability
$Ahad_1([subj,obj,comp-s])	0.0250
$Ahad_1([subj,obj,comp-sbar])	0.0500
$Ahad_1([subj,passive])	0.1000
$Ahad_1([subj,obj])	0.8000
$Ahad_1([subj])	0.0250

Table 6: Subcategorization frames with probabilities.

3.4 Evaluating the Subcategorization Frames

We compare our resource on subcategorization frames against a manually created subcategorization frames lexicon used in a rule-based LFG Parser. The Arabic LFG Parser has detailed subcategorisation information for lexical entries that includes the preposition of obliques, control relationships (or XCOMPs), and the type of complementizer in verbs that have complements. The number of subcategorization frames collected in the ATB induced resource is comparable to the manually constructed frames in the Arabic LFG parser for nouns and adjectives, but it is almost four times larger for verbs, as shown in Table 7. Figure 2 compares the size of the two resources in proportional intersecting circles. The circles on the left represent the treebank-induced resource, and the circles on the right represent the manually constructed resource.

	Verbs	Nouns	Adjectives
lemma-subcat pairs in ATB	6596	855	295
lemma-subcat pairs in the LFG Parser	1621	991	289
Common lemmas	1447	268	70

Table 7: Number of subcat frames in the ATB and the Arabic LFG Parser

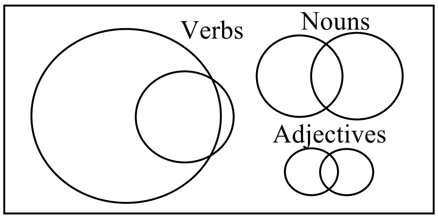


Figure 2: Intersecting circles of ATB subcategorization frames (left) and the LFG Parser (right)

We compare the subcategorization frames in terms of precision, defined here as the number of exact matches of the argument list divided by the number of common lemmas. Table 8 shows results of matching on all grammatical functions and on selected grammatical relations. We conduct the evaluation experiment at four levels: (1) we match the full argument list between the two data sets, (2) we remove the value of the preposition in obliques, (3) we also remove COMPs and XCOMPs, and (4) we only leave SUBJs, OBJs and OBJ2s. Number (4) denotes transitivity, or the most important type of argument. The smaller the number, the less important the argument type is considered in our perspective.

		Precision		
		Verbs	Nouns	Adjectives
1	Full argument list	0.78	0.50	0.53
2	Without preps	0.82	0.52	0.66
3	Without preps, comps and xcomps	0.84	0.54	0.67
4	Without obls, comps and xcomps	0.97	0.73	0.86

Table 8: Evaluating the Tree-induced subcategorization frames against the resource in the Arabic LFG Parser.

Table 8 shows that, at level 4, there is a high level of agreement between the two resources. At level 1, although the precision is comparatively low for nouns and adjectives, we notice that the precision is high for verbs which constitute the largest portion of the data and the most important type of predicates when dealing with subcategorization frames.

4. AraComLex Lexical Management Application

In order to manage our lexical database, we have developed the AraComLex (Arabic Computer Lexicon) authoring system which provides a Graphical User Interface for human lexicographers to review, modify and update the automatically derived lexical and morphological information. We use AraComLex for storing the lexical resources mentioned in this paper as well as generating data for other purposes, such as frequency counts and data for extending our morphological transducer.

The data used in the AraComLex is stored in a relational database, with all various tables connected together as shown in Figure 3 which presents a diagram of the entity relationship (Chen, 1976) of the database. In this diagram, entities are drawn as rectangles and relationships as diamonds. Relationships connect pairs of entities with given cardinality constraints (represented as numbers surrounding the relationship). Three types of cardinality constraints are used in the diagram: 0 (entries in the entity are not required to take part in the relationship), 1 (each entry takes part in exactly one relationship) and n (entries can take part in an arbitrary number of relationships). Entities correspond to tables in the database, while relationships model the relations between the tables.

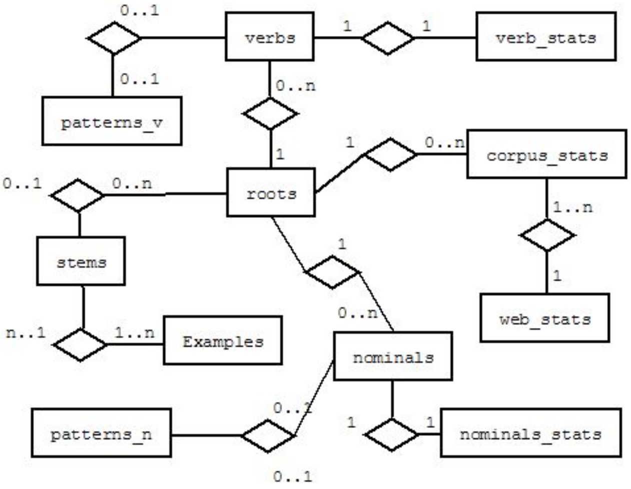
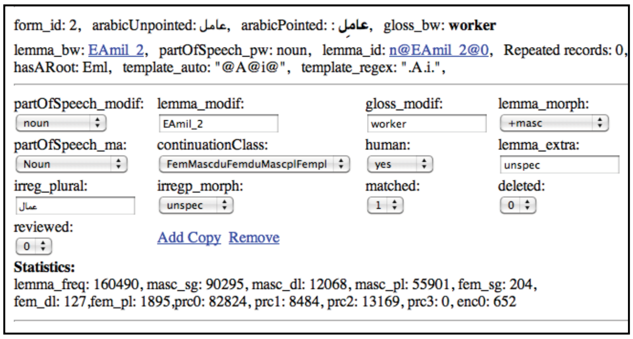


Figure 3: Entity Relationship diagram of AraComLex

AraComLex lists all the relevant morphological and morpho-syntactic features for each lemma. We use finite sets of values implemented as drop-down menus to allow lexicographers to edit entries while ensuring consistency, as shown in Figure 4. Two of the innovative features added are the “human” feature and the 13 continuation classes which stand for the inflection grid, or all possible inflection paths, for nominals as shown in Table 2 above. Statistics show the total frequency of the lemma in the corpus and the weights of each morpho-syntactic feature.


form_id: 2, arabicUnpointed: عامل, arabicPointed: : عامِل, gloss_bw: **worker**

lemma_bw: EAmil_2, partOfSpeech_pw: noun, lemma_id: n@EAmil_2@0, Repeated records: 0, hasARoot: Eml, template_auto: "@A@i@", template_regex: ".A.i.",

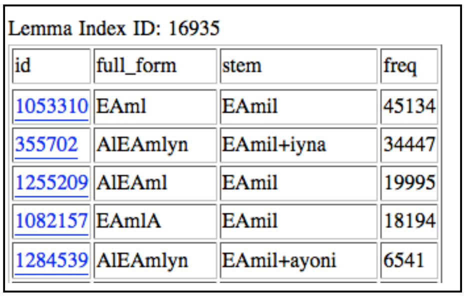
partOfSpeech_modif: noun
lemma_modif: EAmil_2
gloss_modif: worker
lemma_morph: +masc
partOfSpeech_ma: Noun
continuationClass: FemMascduFemduMascplFempl
human: yes
lemma_extra: unspec
irreg_plural: عمال
irregp_morph: unspec
matched: 1
deleted: 0
reviewed: 0

Statistics:
lemma_freq: 160490, masc_sg: 90295, masc_dl: 12068, masc_pl: 55901, fem_sg: 204, fem_dl: 127,fem_pl: 1895,prc0: 82824, prc1: 8484, prc2: 13169, prc3: 0, enc0: 652

Figure 4: A nominal entry in AraComLex

Figure 4 shows the features specified for nominal lemmas in AraComLex. The feature “lemma_morph” feature can be either ‘masc’ or ‘fem’ for nouns and can also be ‘unspec’ (unspecified) for adjectives. Following SAMA, “partOfSpeech” can be ‘noun’, ‘noun_prop’, ‘noun_quant’, ‘noun_num’, ‘adj’, ‘adj_comp’, or ‘adj_num’.

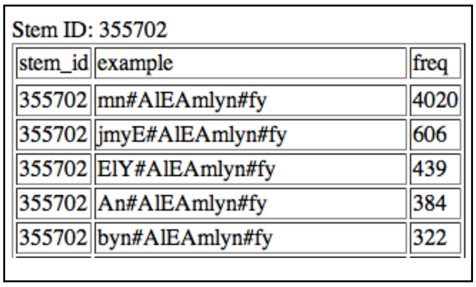
For lexicographic purposes, a lexicographer can review the lemma in detail by looking into the stems and full forms, as shown in Figure 5.


Lemma Index ID: 16935

id	full_form	stem	freq
1053310	EAml	EAmil	45134
355702	AlEAmlyn	EAmil+iyna	34447
1255209	AlEAml	EAmil	19995
1082157	EAmlA	EAmil	18194
1284539	AlEAmlyn	EAmil+ayoni	6541

Figure 5: Lemma Stems

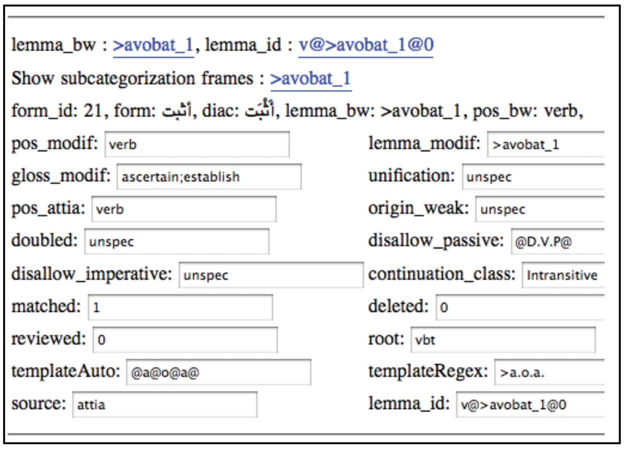
The lexicographer can go even further by reviewing the examples in which the words occurred, sorted according to frequency, as shown in Figure 6. For practical reasons and to keep the size of the database within reasonable bound, we only keep records of the word’s bigrams, which in most cases are enough to provide a glimpse of the context and possible collocates.


Stem ID: 355702

stem_id	example	freq
355702	mn#AlEAmlyn#fy	4020
355702	jmyE#AlEAmlyn#fy	606
355702	ElY#AlEAmlyn#fy	439
355702	An#AlEAmlyn#fy	384
355702	byn#AlEAmlyn#fy	322

Figure 6: Word Examples

For verb lemmas, as shown in Figure 7, we provide information on whether the verb is transitive or intransitive and whether it allows passive and imperative inflection, as well as the usual information on the template and the root. One of the features that can be highly valuable for a lexicographer is the link to subcategorization frames.


lemma_bw : >avobat_1, lemma_id : v@>avobat_1@0

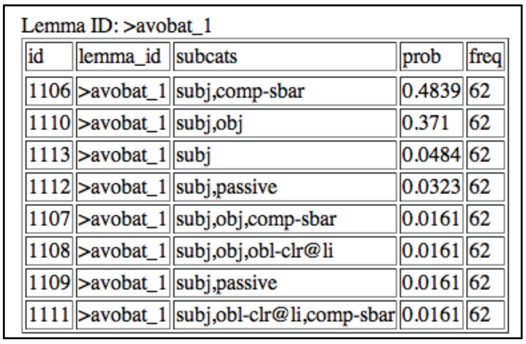
Show subcategorization frames : >avobat_1

form_id: 21, form: أثبت, diac: أَثْبَتَ, lemma_bw: >avobat_1, pos_bw: verb,

pos_modif: verb
lemma_modif: >avobat_1
gloss_modif: ascertain;establish
unification: unspec
pos_attia: verb
origin_weak: unspec
doubled: unspec
disallow_passive: @D.V.P@
disallow_imperative: unspec
continuation_class: Intransitive
matched: 1
deleted: 0
reviewed: 0
root: vbt
templateAuto: @a@o@a@
templateRegex: >a.o.a.
source: attia
lemma_id: v@>avobat_1@0

Figure 7: A verb entry in AraComLex

The subcategorization frames, as shown in Figure 8, are sorted by probability, ensuring that more frequent subcategorization frames appear on the top. As the figure shows, information on passive occurrences and prepositions for obliques are also included.



Lemma ID: >avobat_1

id	lemma_id	subcats	prob	freq
1106	>avobat_1	subj,comp-sbar	0.4839	62
1110	>avobat_1	subj,obj	0.371	62
1113	>avobat_1	subj	0.0484	62
1112	>avobat_1	subj,passive	0.0323	62
1107	>avobat_1	subj,obj,comp-sbar	0.0161	62
1108	>avobat_1	subj,obj,obl-clr@li	0.0161	62
1109	>avobat_1	subj,passive	0.0161	62
1111	>avobat_1	subj,obl-clr@li,comp-sbar	0.0161	62

Figure 8: Verb Subcategorization Frames

5. Conclusion

We build a lexicon for MSA focusing on the problem that existing lexical resources tend to include a large subset of obsolete lexical entries, no longer attested in contemporary data, and they do not contain sufficient syntactic information. We start with a manually constructed lexicon of 10,799 MSA lemmas and automatically extend it using lexical entries from SAMA's lexical database, carefully excluding obsolete entries and analyses. We use machine learning on statistics derived from a large pre-annotated corpus for automatically predicting inflectional paradigms, successfully extending the lexicon to 30,587 lemmas. We also provide essential lexicographic information by automatically building a lexicon of subcategorization frames from the ATB. We develop a lexicon authoring system, AraComLex,[11] to aid the manual revision of the lexical database by lexicographers. As an output of this project, we create and distribute an open-source finite-state morphological transducer. [12] We also distribute a number of open-source resources that are of essential importance for lexicographic work, including a list of Arabic morphological patterns, [13] subcategorization frames,[14] and Arabic lemma frequency counts.[15]

6. Acknowledgements

This research is funded by Enterprise Ireland (PC/09/037), the Irish Research Council for Science Engineering and Technology (IRCSET), and the EU projects PANACEA (7FP-ITC-248064) and META-NET (FP7-ICT- 249119).

7. References

Al-Sulaiti, L., Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11(1), pp. 135-171.

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

Attia, M., Rashwan, M., Ragheb, A., Al-Badrashiny, M., Al-Basoumy, H. & Abdou, S. (2008). A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields. In B. Nordström, A. Ranta (eds.) *GoTAL '08 Proceedings of the 6th international conference on Advances in Natural Language Processing.* Götenburg, Sweden: Springer-Verlag Berlin, Heidelberg, pp. 65-76.

Attia, M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. *In Challenges of Arabic for NLP/MT Conference*, The British Computer Society, London, UK.

Attia, M. (2008). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. Ph.D. Thesis. The University of Manchester, Manchester, UK.

Beesley, K.R., Karttunen, L. (2003). *Finite State Morphology*: CSLI studies in computational linguistics. Stanford, California.: CSLI.

Bin-Muqbil, M. (2006). *Phonetic and Phonological Aspects of Arabic Emphatics and Gutturals*. Ph.D. thesis in the University of Wisconsin, Madison.

Boudelaa, S., Marslen-Wilson, W.D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, 42(2), pp. 481-487.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer (BAMA) Version 2.0. Linguistic Data Consortium (LDC) catalogue number: LDC2004L02, ISBN1-58563- 324-0.

Chen, P.P. (1976). The Entity-Relationship Model: Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1, pp. 9-36.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Volume 34 of Syntax and Semantics. New York: Academic Press.

Elgibali, A., Badawi, E.M. (1996). *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said M. Badawi*. Egypt: American University in Cairo Press.

Fischer, W. (1997). Classical Arabic. In R. Hetzron (ed.) *The Semitic Languages*. London: Routledge, pp. 397-405.

Ghazali, S., Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. Arabic NLP Workshop at ACL/EACL. Toulouse, France.

Grishman, R., MacLeod, C. & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, pp. 268–272.

Hajič, J., Smrž, O., Buckwalter, T. & Jin, H. (2005). Feature-Based Tagger of Approximations of

[11] http://www.cngl.ie/aracomlex

[12] http://aracomlex.sourceforge.net

[13] http://arabicpatterns.sourceforge.net

[14] http://arabicsubcats.sourceforge.net

[15] http://arabicwordcount.sourceforge.net

Functional Arabic Morphology. In *The 4th Workshop on Treebanks and Linguistic Theories* (TLT 2005), Barcelona, Spain.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2 ed.). Prentice Hall.

Kiraz, G.A. (2001). Computational Nonlinear Morphology: With Emphasis on Semitic Languages. Cambridge: Cambridge University Press.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S. & Kulick, S. (2010). LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.0. LDC Catalog No. LDC2010L01. ISBN: 1-58563-555-3.

Maamouri, M., Bies, A. (2004). Developing an Arabic Treebank: Methods, guidelines, procedures, and tools. In *Workshop on Computational Approaches to Arabic Script-based Languages, COLING*.

Manning, C. (1993). Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 235–242.

O'Donovan, R., Burke, M., Cahill, A., van Genabith, J. & Way, A. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics*, 31(3), pp. 329-366.

Owens, J. (1997). The Arabic Grammatical Tradition. In R. Hetzron, ed.) *The Semitic Languages*. London: Routledge, pp. 46-48.

Palmer, M., Bies, A., Babko-Malaya, O., Diab, M., Maamouri M., Mansouri, A. & Zaghouni, W. (2008). A pilot Arabic Propbank. In *Proceedings of LREC*, Marrakech, Morocco.

Parker, R., Graff, D., Chen, K., Kong, J. & Maeda, K. (2009). Arabic Gigaword Fourth Edition. LDC Catalog No. LDC2009T30. ISBN: 1-58563-532-4.

Rehbein, I., van Genabith, J. (2009). Automatic Acquisition of LFG Resources For German - As Good As It Gets. In *Proceedings of the LFG09 Conference*. Cambridge, UK.

Roth, R., Rambow, O., Habash, N., Diab, M. & Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of Association for Computational Linguistics (ACL)*, Columbus, Ohio.

Sinclair, J. M. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

Tounsi, L., Attia, M. & van Genabith, J. (2009). 'Automatic Treebank-Based Acquisition of Arabic LFG Dependency Structures.' EACL Workshop on Computational Approaches to Semitic Languages, Athens, Greece.

Van Mol, M. (2000). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the ninth EURALEX International Congress*, Stuttgart, pp. 831-836.

Van Mol, M. (2003). *Variation in Modern Standard Arabic in Radio News Broadcasts, A Synchronic Descriptive Investigation in the use of complementary Particles*. Leuven, OLA 117.

Watson, J. (2002). *The Phonology and Morphology of Arabic*. New York: Oxford University Press.

Wehr, H., Cowan, J.M. (1976). *Dictionary of Modern Written Arabic*, pp. VII-XV. Ithaca, New York: Spoken Language Services.

A model for integrated dictionaries of fixed expressions

Henning Bergenholtz[1,2,3], Theo Bothma[2] and Rufus Gouws[3]
[1] Aarhus University; [2] University of Pretoria; [3] Stellenbosch University

[1] Center for Lexicography
Fuglesangs Allé 4
8210, Aarhus V
Denmark

[2] Department of Information Science
Private Bag X20
Hatfield, 0028
South Africa

[3] Department of Afrikaans and Dutch
Private Bag X1
Matieland, 7602
South Africa

E-mail: hb@asb.dk, tbothma@up.ac.za, rhg@sun.ac.za

Abstract

This paper discusses a project for the creation of a theoretical model for integrated e-dictionaries, illustrated by means of an e-information tool for the presentation and treatment of fixed expressions using Afrikaans as example language. To achieve this a database of fixed expressions is compiled wherein data are treated in such a way that access can be provided through a variety of dictionaries for specific situations, based on specific lexicographic functions, e.g. the cognitive function as well as the communicative functions of text reception and text production. From one database, the user will have access to six monofunctional dictionaries of fixed expressions. Each one of these dictionaries provides a view on selected fields of the database, i.e. a search is carried out on selected fields in the database and only the data in specific fields that are relevant for the specific dictionary are displayed. There are unique user needs that may not necessarily be satisfied by means of these six dictionaries. Individualised search facilities will therefore be provided to enable a user to retrieve data from a single data field or a user-specified selection of data fields. Phase two will provide the option of setting up a user profile, an extension of data fields and linking to external data sources. The result of the project will therefore be a comprehensive database of Afrikaans fixed expressions that may be accessed through six monofunctional dictionaries, as well as individualised search options, user profiling and the possibility to display additional data on demand.

Keywords: Fixed expressions; databases; integrated dictionaries; monofunctional and polyfunctional dictionaries

1. A purportedly user-friendly e-idiom dictionary

For many centuries lexicographers have proudly claimed that specifically their own dictionary was user-friendly and satisfied the needs of all users as well. This was, and still is, an immunising and self-serving assertion in most cases. It is based on real, factual research only to a limited extent, and at the same time it is an advertising measure to persuade potential dictionary buyers. However, one thing has changed. Up to some 30 years ago, dictionary user research was *de facto* nonexistent. This was formulated quite succinctly by Wiegand, who referred to the "known unknown" that needed to be researched (Wiegand, 1977: 59). Apparently, much research has indeed been done. Numerous surveys of all kinds have been conducted, but much of this assumed the form of memory-based questions such as: "How often do you use a dictionary? Daily? Weekly? Monthly? Rarely?"; "What kind of information do you look for? Grammatical? Orthographic? Semantic?"; "What kind of entries do you look for? Collocations? Examples? Items about style?" These days it is hardly possible any more to read and understand all contributions made to and by user studies. In our view this is not worthwhile either, as such surveys mostly ask questions which are constructed instead of being rooted in reality. The research should be conducted on real users with their real and specific needs and on their use of dictionaries, but in most cases it is not. A user with a cognition-related information need may be looking explicitly for certain types of data (examples, for instance). That is not what users with a need for communication-related information do. They have a problem concerning reception, text production or translation and are hoping to get the necessary help in this regard. Ordinary users do not know exactly what lexicographers (and linguists in particular) call such items. We therefore hold the view that lexicographers should not act as dictionary philologists or interpreters of what users remember about their use of dictionaries, but should especially develop new concepts on the basis of theoretical considerations concerning the needs of certain types of users in foreseen situations. It is not the actual user that matters, but the potential user and his/her potential needs in situations anticipated by the lexicographer. For these needs the lexicographer develops a (new) tool of which he/she assumes that it can satisfy the needs he/she foresees. Within the function theory such dictionaries are typically monofunctional, i.e., they address a specific need of a specific user group in a specific situation (see, for example, Bergenholtz, 2010; Bergenholtz, 2011; Bergenholtz & Bergenholtz, 2011, Tarp, 2007; Tarp, 2008; Tarp, 2009a; Tarp, 2009b; Tarp, 2011). Yet a majority of practical and theoretical lexicographers assume that all dictionaries should always provide as much as possible data for the identified user groups, and therefore always were, and should continue to be, polyfunctional (see Bergenholtz 2010).

The above introduction reflects the arguments which have been frequently put forward in lexicographic discussions in recent years. Tarp (2002) proposed the following basic division into two types of lexicography: In **contemplative lexicography**, existing dictionaries are analysed and users are questioned about their use of existing dictionaries to date. In **transformative**

lexicography, theoretical analyses of the potential user situations, the respective user conditions and the user needs are used to develop new concepts for compiling new dictionaries, typically monofunctional dictionaries. On the basis of theoretical analyses the lexicographer therefore decides what the characteristics are of the monofunctional dictionaries that will satisfy specific user needs. In the case of the Centlex dictionaries developed at the University of Aarhus, no general surveys on the use of these dictionaries are undertaken, but feedback in the form of e-mails is analysed and taken into consideration. Moreover, log file analyses are done which, in selected cases, are linked to enquiries among a handful of users (see Bergenholtz & Johnsen 2005; Bergenholtz & Johnsen 2007).

Such log file analyses and feedback can lead to small changes, but also to a complete redesign of the dictionary, as was the case with the e-idiom dictionary by Vrang, Bergenholtz & Lund (2003-2005) (*Den danske Idiomordbog*). This was a dictionary of idioms containing the relatively large number of 8500 idioms. It had been designed especially as a reception dictionary, as it contained only meaning items. In the user guide and in the outer text on the structure of the dictionary, the meaning of 'idiom' was explained clearly. The authors received a fair number of e-mails from users with feedback on this dictionary. None of these mails asked what actually constitutes an idiom. They were, however, quite frequently asked why this or that combination of words could not be found in *Den danske Idiomordbog*. The typical answer to this question was that the expression in question is a proverb, not an idiom, and is therefore not in this dictionary. This happened regardless of the fact that the terms were clearly defined in the user guide.

During the period from mid-2003 until mid-2004 the number of unsuccessful dictionary searches was relatively high. (Misspelled search terms are included here, but amounted to fewer than 3% of searches; searches for unlemmatised idioms are also included, but these searches amounted to less than 1% of the searches.)

Number of searches in *Den danske Idiomordbog*	
With result	70.4%
Without result	29.6%

Table 1: Percentage of successful and unsuccessful searches in *Den danske Idiomordbog*, 2003-2004

There are two sides to the bare figures for successful and unsuccessful uses of *Den danske Idiomordbog*. The positive side is that the users find the idiom they were looking for in more than 70% of all enquiries. The negative side is that in about 26% of all enquiries (cases with incorrect spelling and deficient lemmas have been deducted from the 29.6%) users were looking for 'idioms' which are not idioms but proverbs, sayings ('winged words'), standard formulations, multiword expressions from other languages and many more. We do not believe that another definition of idioms would have improved this rate. On the other hand, for an internet dictionary of idioms there is an obvious solution: Don't make one at all; make one that contains all forms of fixed expressions. Moreover, the user with a reception problem does not even need to know what kind of fixed expression he/she is dealing with; he/she needs only the meaning. This insight led to a new concept for a new Danish database with fixed expressions from which several monofunctional dictionaries are offered to users (see Bergenholtz, 2011).

The preceding insight led to the decision to compile a database for fixed expressions in Afrikaans, rather than a database of idioms.

The concept for this database with several monofunctional dictionaries is the point of departure for the concept of the Afrikaans database presented here. The new database differs from the previous concept in some respects, however, especially as regards the number of fields for item types. The Danish database has 14 fields, the new database has 36. Also, this is a database for two languages, viz. Afrikaans and English, not for one. Nevertheless, the basic concept remains intact. A database and a dictionary are not the same thing. A single dictionary can be extracted from a database, and the result will normally be a polyfunctional dictionary. From a database, on the other hand, as many dictionaries could be extracted as are deemed relevant on the basis of theoretical considerations and experience with earlier databases, and these should be function-oriented monofunctional dictionaries (see Bergenholtz & Tarp 2002; Bergenholtz & Tarp 2003; Bergenholtz & Tarp 2005).

2. Afrikaans dictionaries with fixed expressions

Afrikaans dictionaries represent a wide-ranging typological variety, compiled to assist different users in finding assistance with regard to both language for general purposes and language for special purposes. Within the category of general dictionaries various monolingual and bilingual dictionaries offer an extensive presentation of fixed expressions. The category of restricted dictionaries also include a few dictionaries that focus on fixed expressions, cf. Malherbe (1924), De Villiers & Gouws (1988), Botha, Kroes & Winckler (1994), Prinsloo (1997) and Prinsloo (2009). The extent and nature of both the macrostructural coverage and the treatment in these dictionaries of fixed expressions differ considerably. They share one feature and that is that they have been produced in printed format. The only non-printed version of Afrikaans fixed expressions can be found in the presentation and treatment of this category of lexical items in those general monolingual and bilingual dictionaries that are available in CD ROM format or online. Afrikaans has a real need for an e-dictionary of fixed expressions. The advantage of the fact that all dictionaries of fixed expressions are only available in printed format is that no bad e-dictionary exists, and a transformative approach to the planning and compilation of an e-dictionary of fixed expressions does not have to pay any attention to any electronic predecessor. In the following sections the plan for an innovative e-dictionary that deals with Afrikaans fixed expressions will be discussed.

3. Concept for a database system for Afrikaans fixed expressions

The database system consists of the database itself and the database management system. The database is developed in MySQL. It is integrated in a database management system that has been developed using open source software (HTML, XML, XMLT, Perl, CGI and related technologies). The database management system has a comprehensive administrative back-end which manages access, data security and integrity, including aspects such as version control and back-up. The system has two further interfaces, viz. an interface for the researchers contributing the data for the different fields in the database and an interface for end-users through which they obtain access to the dictionaries and other customization functionalities.

In principle, hundreds or even thousands of fields related to one or more phenomena could be provided in a database with one or more languages. For instance, there is a total of 84 fields in a Danish, English and Spanish accounting database, from which 23 different dictionaries are offered to users at present (see Bergenholtz 2011). In this case, only fields with different types of data – apart from the lexicographers' notes on the work in progress – which also finds its way into one or more dictionaries are provided for in the database. If one or more collaborating lexicographers have data at their disposal which are not intended to be presented in at least one dictionary, specific fields could be created for such data so that it could perhaps be accommodated in one or more additional dictionaries. The limit must be drawn where the number of fields becomes so large that the lexicographer loses sight of the big picture and the first presentation of the database takes too long.

The order of the fields in the "Field" column in Table 2 is a working order; the order in the individual dictionaries is determined for each respective dictionary.

Field
1. Core field
2. Meaning in Afrikaans
3. Internet link to meaning
4. Further meaning item in Afrikaans
5. Meaning in English
6. Grammar
7. Comment on grammar
8. Internet link to grammar
9. Background remark(s)
10. Comment on background remark(s)
11. Internet link to background remark(s)
12. Fixed expression(s) in Afrikaans
13. Remarks on the fixed expression(s)
14. References to fixed expression(s)
15. Internet link to variants, e.g. statistical
16. Fixed expression(s) in English translated from Afrikaans
17. Style
18. Comment on style
19. Internet link to style
20. Classification of the fixed expression
21. Comment on classification
22. Collocation(s)
23. Comment on collocation(s)
24. Internet link to collocation(s)
25. Example(s)
26. Comment on example(s)
27. Internet link to example(s)
28. Synonym(s)
29. Comment on synonym(s)
30. Internet link to synonym(s)
31. Antonym(s)
32. Comment on antonym(s)
33. Internet link to antonym(s)
34. Associated concept(s)
35. Key word(s)
36. Memo field

Table 2: Data fields for the database of fixed expressions in Afrikaans

There is not enough space here to justify all fields. However, some fields which are not self-explanatory do require some explanation. Field 12 contains only one expression in some cases, but in others there are more if variants exist, e.g. the same core of a fixed expression combined with different verbs. If one wants to, one can call this a lemma field. We do not; that would rather be field 1, called the core field, which is identical to Field 12 if there are no variants and contains only the words it has in common with the variants in Field 12 if there are variants. This field is used for automatic searches on the one hand, and on the other for items the user can use as links if search results are displayed as a list, or if synonyms or antonyms are provided. The field contains key words with all the lexical words, including irregularly conjugated forms which occur in the fixed expression(s) of a particular card. In Field 22, we use the term 'collocations' in the sense of combinations of words in which the fixed expression occurs. A collocation is never a complete sentence, unlike the data in Field 25, where 'example' refers to a full sentence. Field 9 contains a brief history behind the full expression; if there are two different histories and it is not clear which one is correct, both are given. In addition, in some cases reference is made to background histories which are given in various textbooks and dictionaries (but are not necessarily correct). Lastly, field 34 contains associated concept(s). This refers to concepts which can be associated with the meaning and use of the fixed expression. Finding such concepts could be very time consuming if some sort of semantic system were applied. That is not how it is done here. In fact, the editor's lexicographic instruction is to write down up to five such associated concepts within 30 seconds (but never more).

4. Six dictionaries with fixed expressions

At present the concept provides for six dictionaries – five monofunctional and one polyfunctional. It is a model that can be used not only for these languages and this language combination, but in principle also for at least all Indo-European languages and probably also for other language families, for example the Austronesian languages.

4.1 MEANING OF FIXED EXPRESSIONS

Access to the first dictionary is gained by pressing the button "I am reading a text, but do not understand the meaning of a fixed expression". Here the user enters an expression or part of an expression in the search field and obtains the desired information, i.e. the meaning of the fixed expression. This dictionary is called MEANING OF FIXED EXPRESSIONS. When a search is done in this dictionary, the program looks in two of the fields in the database in the order indicated by figures[1] (see column 1 in Table 3 below). For this dictionary a maximising search is done. The user obtains one or several articles with the content of three of the fields in the database (see column 3). In other words, the user receives only a small part of the database asarticle, but it is exactly the part that is needed to solve a reception problem. If more than 10 articles are found they are displayed as a list where the content of the core field and the first line of the meaning are shown.

The following tables show only those fields which are used for a search and those fields from which data are presented in the dictionary article for the specific dictionary. Because of space limitations in the headings of the tables, "Search" is used as an indication of the fields that are searched and the numbers indicate the order in which the search is carried out. Similarly, "Entry" refers to the fields that are shown to the user and the numbers indicate the ordering sequence of the fields in the specific dictionary article. "List" is used as an indication of which data are displayed when a list is needed (i.e., when more than 10 articles are found).

Search	Field	Entry	List
1	1. Core field		1
	2. Meaning in Afrikaans	1	2 1st line
	5. Further meaning item in Afrikaans	2	
2	12. Fixed expression(s) in Afrikaans	3	
3	35. Key word(s)		

Table 3: Search and data fields in the dictionary MEANING OF FIXED EXPRESSIONS

4.2 USE OF FIXED EXPRESSIONS

The second dictionary is activated by pressing the button "I am writing a text with a specific fixed expression". Here the user enters a fixed expression or part of it in the search field and obtains information about the use of the fixed expression, including its meaning, grammar, collocations, example sentences and synonymous or antonymous fixed expressions. In other words, the search is expression-specific. We call this dictionary USE OF FIXED EXPRESSIONS. When this dictionary is activated, four fields of the database are searched; however, this is a minimising search, where the search is terminated after one field type has been searched and other fields are therefore not searched. The items relevant to text production are reflected as figures in column 3; if there are more than 10 articles, a list is shown.

Search	Field	Entry	List
1	1. Core field		1
	2. Meaning in Afrikaans	3	2 1st line
	3. Internet link to meaning	5	
	4. Further meaning item in Afrikaans	4	
	6. Grammar	8	
	7. Comment on grammar	9	
2	12. Fixed expression(s) in Afrikaans	1	
	13. Remarks on the fixed expression(s)	2	
	17. Style	6	
	18. Comment on style	7	
4	22. Collocation(s)	10	
	23. Comment on collocation	11	
5	25. Example(s)	12	
	26. Comment on examples	13	
	28. Synonym(s)	14	
	29. Comment on synonyms	15	
	31. Antonym(s)	16	
	32. Comment on antonyms	17	
3	35. Key word(s)		

Table 4: Search and data fields in the dictionary USE OF FIXED EXPRESSIONS

[1] In a maximising search the order does not really matter, as all subresults for each individual search are added up in the overall result. In a minimising search this is different. In this case the search ends after searching one field if one or more results are found. Therefore the next fields are not searched.

4.3 Fixed Expressions with a Specific Meaning

The third dictionary is activated by pressing the button "I am writing a text and am looking for a fixed expression with a specific meaning". Here the user can enter one or several words with a specific meaning and find expressions with this meaning or part of this meaning. The user then receives information about the use of the expression, including its meaning, grammar, collocations, example sentences and synonymous or antonymous fixed expressions. In other words, the point of departure is a meaning, which can be very wide and can therefore yield many hits. If a more restricted meaning is used as the search string, fewer hits may be found or even none at all. This dictionary is called Fixed Expressions with a Specific Meaning. When a search is done in this dictionary, the program looks in three of the fields in the database, in the case of a maximising search. The data are presented as in the dictionary mentioned above (Use of Fixed Expressions), as the function is the same, i.e. assistance with text production problems.

Search	Field	Entry	List
	1. Core field		1
1	2. Meaning in Afrikaans	3	2 1st line
	3. Internet link to meaning	5	
2	4. Further meaning item in Afrikaans	4	
	5. Grammar	8	
	6. Comment on grammar	9	
	12. Fixed expression(s) in Afrikaans	1	
	13. Remarks on the fixed expression(s)	2	
	17. Style	6	
	18. Comment on style	7	
	22. Collocation(s)	10	
	23. Comment on collocation	11	
	25. Example(s)	12	
	26. Comment on examples	13	
	28. Synonym(s)	14	
	29. Comment on synonyms	15	
	31. Antonym(s)	16	
	32. Comment on antonyms	17	
3	34. Associated concept(s)		

Table 5: Search and data fields in the dictionary Fixed Expressions with a Specific Meaning

One can then click on the core expression to get to the dictionary article which gives a meaning that fits the context. An article will be displayed with a set of corresponding data, as was illustrated above in the Use of Fixed Expressions dictionary. Although the data presentation of the two dictionaries is identical, the dictionaries are not. In the dictionary Fixed Expressions with a Specific Meaning access is gained by means of a meaning-oriented search, as in a printed dictionary with a systematic macrostructure and with one or more registers, whereas the dictionary Use of Fixed Expressions corresponds to a dictionary with an alphabetic macrostructure without registers. But the information the user is looking for to assist him/her with the production of a text is the same for both dictionaries.

4.4 Knowledge about Fixed Expressions

For the Danish dictionaries with fixed expressions mentioned above there are four dictionaries, the first three like those presented here and a fourth which shows all fields in the database. Here we found that the two text production dictionaries accounted for only about 9% of all user actions during the second period in Table 6. A comparison with the log file analysis from the earlier period (2007), when there was only one production dictionary, shows that this share is relatively stable. Compared with the polyfunctional dictionary, which shows everything, the reception dictionary showed a substantial shift in the user actions between the two periods, which are presented in Table 6 as absolute figures and as percentages.

27 February 2007 until 17 December 2007		
Understanding a text	51 242	60.33%
Writing a text	5 294	6.23%
All data	28 405	33.44%
17 December 2007 until 1 December 2008		
Understanding a text	154 239	29.57%
Writing a text with a known expression	19 386	3.72%
Writing a text with a known meaning	27 052	5.19%
All data	320 865	61.52%

Table 6: Usage statistics for the Danish dictionaries of fixed expressions

Feedback from a random selection of users showed that the change is explained by the fact that many users are looking particularly for the historical (= generic) background to the fixed expression and selected the dictionary that displayed all data for this reason. In view of this experience, we therefore offer a separate dictionary that supplies such historical data as well as meaning items. It is therefore a cognitive dictionary in which a maximising search is performed (left column) and the items of the respective fields are shown in the third column in the order indicated. We call this dictionary Knowledge about Fixed Expressions.

Search	Field	Entry	List
1	1. Core field		1
	2. Meaning in Afrikaans		2 1st line
	9. Background remark(s)	6	
	10. Comment on background remark(s)	7	
	11. Internet link to background remark(s)	8	
2	12. Fixed expression(s) in Afrikaans	1	
	13. Remark(s) on the fixed expression(s)	2	
	14. References to fixed expression(s)	3	
	20. Classification of the fixed expression	4	
	21. Comment on classification	5	
3	35. Key word(s)		

Table 7: Search and data fields in the dictionary KNOWLEDGE ABOUT FIXED EXPRESSIONS

4.5 AFRIKAANS-ENGLISH DICTIONARY OF FIXED EXPRESSIONS

We call the fifth dictionary the AFRIKAANS-ENGLISH DICTIONARY OF FIXED EXPRESSIONS. It is a communication dictionary with the function of translation. It is not an ideal translation dictionary, however, as no grammatical information on the English equivalents is presented and no translations of collocations or examples are supplied in Afrikaans.

Search	Field	Entry	List
1	1. Core field		1
	2. Meaning in Afrikaans	2	
	3. Internet link to meaning	3	
	4. Further meaning item in Afrikaans	4	
	5. Meaning in English	6	
2	12. Fixed expression(s) in Afrikaans	1	
	16. Fixed expression(s) in English translated from Afrikaans	5	
3	35. Key word(s)		

Table 8: Search and data fields in the dictionary KNOWLEDGE ABOUT FIXED EXPRESSIONS

4.6 COMPREHENSIVE KNOWLEDGE ABOUT FIXED EXPRESSIONS

The sixth dictionary is activated by pressing the button "I want to know as much as possible about fixed expressions". We call it COMPREHENSIVE KNOWLEDGE ABOUT FIXED EXPRESSIONS. It is a traditional polyfunctional dictionary that shows all fields (except for the field for working notes). A minimising search is performed.

Search	Field	Entry	List
1	1. Core field	1	1
	2. Meaning in Afrikaans	12	
	3. Internet link to meaning	13	
	4. Further meaning item in Afrikaans	14	
	5. Meaning in English	15	
	6. Grammar	19	
	7. Comment on grammar	20	
	8. Internet link to grammar	21	
	9. Background remark(s)	16	
	10. Comment on background remark(s)	17	
	11. Internet link to background remark(s)	18	
2	12. Fixed expression(s) in Afrikaans	7	
	13. Remark(s) on the fixed expression(s)	8	
	14. References to fixed expression(s)	9	
	15. Internet link to variants, e.g. statistical	10	
	16. Fixed expression(s) in English translated from Afrikaans	11	
	17. Style	2	
	18. Comment on style	3	
	19. Internet link to style	4	
	20. Classification of the fixed expression	5	
	21. Comments on classification	6	
	22. Collocation(s)	22	
	23. Comment on collocations	23	
	24. Internet link to collocations	24	
	25. Example(s)	25	
	26. Comment on examples	26	
	27. Internet link to examples	27	
	28. Synonym(s)	28	
	29. Comment on synonyms	29	
	30. Internet link to synonyms	30	
	31. Antonym(s)	31	
	32. Comment on antonyms	32	
	33. Internet link to antonyms	33	
	34. Associated concept(s)	34	
3	35. Key word(s)	35	
	36. Memo field		

Table 9: Search and data fields in the dictionary COMPREHENSIVE KNOWLEDGE ABOUT FIXED EXPRESSIONS

5. Forthcoming attractions

Ultimately, a database and the dictionaries extracted from it are never finished, as new cards can constantly be added and those that have already been made can also be expanded or corrected. Our aim is to build up a database of 10,000 to 15,000 cards. However, we will already offer the users the lexicographically recorded expressions when only 1,000 cards are ready. In the further course of the work we will, as explained with reference to the Danish dictionaries above, amend or add specific / additional data on the basis of log file analyses and user feedback, as well as on the basis of further research on and experimentation with different concepts and tools for manipulating data in the e-environment.

Provision has already been made for expansion. The intention is to give users the opportunity to define their profiles, to define their search criteria and to select fields and the order in which they are displayed. For some fields we intend providing the option of displaying more detailed information on request and access to advanced tools. We assume that only a small number of users will make use of these options. Nevertheless, when they do, even more dictionaries will be extracted from one and the same database. It may not be possible to give each of the new, user-defined dictionaries a functional description, as has been done here. However, such options will be "capable of meeting all the users' needs in specific types of situations" (Tarp 2009a: 292) by providing "dynamic articles […] structured in different ways according to each type of search criteria", "articles that are especially adapted", resulting in "the 'individualization' of the lexical product, adapting to the concrete needs of a concrete user" (Tarp, 2009b: 57-61).

5.1 User profiling

We intend providing users with the possibility to define a user profile at the beginning of a consultation session; see Bothma, 2011 for details about user profiling technologies. Users will be able to set up a persistent profile that will remain active across multiple user sessions, but will be able to either reset or change this profile at any stage. Profiles fill enable users to define the specific dictionary they intend consulting during a specific interaction session. For example, a user who is reading a text and regularly needs help only with the meaning of fixed expressions may set his/her profile to use the dictionary MEANING OF FIXED EXPRESSIONS as the default dictionary. A user will also be able to set personalised search options (as discussed below) as default.

5.2 Personalized search and display options

The six dictionaries discussed above are six different customised views on the database. Each of these dictionaries is defined in terms of a specific type of user need defined by the lexicographer. Each of the dictionaries is monofunctional in terms of a text reception, text production, text translation or a cognitive information need (in addition to a “traditional” polyfunctional dictionary). It is possible to provide any further number of monofunctional dictionaries in terms of the lexicographer’s analysis of perceived user needs. However, it is also possible to provide the user with the option to define his/her own search and therefore define his/her own personalised / customised dictionary. The principles are discussed in Bothma, 2011 and Bergenholtz & Bothma, 2011. We intend providing such customised advanced search facilities where the user can define exactly which data are to be displayed. The user will be able to display the data of only a single field or any combination of fields to satisfy unique information needs in a given situation.

5.3 Additional fields for more detailed information

Currently we assume that all users require the same amount of detail when accessing a dictionary article by means of any of the six dictionaries and / or the customisation options. However, this is not necessarily the case. Some users may require only a brief description whereas others may require a detailed exposition. This obviously does not apply to all fields, but could typically apply to, for example, background remarks (fields 9-11) and examples (fields 25-27).

A user may require only a few brief comments about the origin and/or history of a fixed expression, or, alternatively, could require a comprehensive exposition on the origins of an expression, alternative views about the origin, a discussion about erroneous or popularly held beliefs about the origins of the expression, etc. The database should make provision to satisfy these individualised user needs as well. The content required for these details can be provided by a member of the lexicographic team (probably a team member who has a background or interest in history, heritage and culture studies) or could be a link to external source(s) where the background of a fixed expression may have been discussed in detail. We intend providing such a facility for expansion. These data can be made accessible on demand, either by means of a “Read more” button when data of fields 9-11 are displayed or by adapting the user profile at the start of the consultation session.

The current database structure makes provision for examples with comments about and links to the original contexts of the examples. We provide a highly selective list of examples to illustrate meaning and use of a specific fixed expression. However, we foresee that in individual cases users may require either more examples or additional detail. For example, in a text production situation, a user writing a historical novel may require to know which of two current variants of a fixed expression was used (or was the more common variant) at the time the novel takes place. This requires access to data typically not within the database and tools for text manipulation that are not associated with a lexicographical database. (One of a number of dictionaries that does incorporate such a facility is the *Base lexicale du français* (*BLF*) (http://ilt.kuleuven.be/blf) which provides the user with the option of linking to various corpora, including a set of documents of the European Parliament and Wikipedia. The selection of the examples does not require any input from the lexicographer as the *BLF* and the corpora are linked automatically. These examples are displayed by the *BLF* only when the user requires this and the possible information overload is displayed on demand.) In the above example a user may require to see the actual examples in context, i.e., a concordance of examples in a keyword in context (KWIC) format; alternatively, a user may require to see a table that provides only a statistical analysis of the occurrence of variants at a specific time. The two options require two different types of tool, namely a tool that can present "raw" corpus data in a KWIC format as well as a tool that can do statistical analysis of the "raw" corpus data and present the results in statistical tables. We hope to incorporate such facilities in due course. This will, however, require a considerable amount of both theoretical and empirical research and depends on the availability of suitable corpora. Research issues that need to be taken into account to incorporate such a facility are, *inter alia*:

- How should the data in the external database(s) be marked up to enable access to specific data at a fine level of granularity? In terms of the above example, granularity may include mark-up for different time periods, different genres, etc.
- How are word form variants such as inflections and conjugations to be handled? For example, does the database require detailed tagging of morphological forms beforehand, or would it be possible to link to the "raw" text of the corpora on the fly without prior tagging?
- What type of tools will be required to make this type of searching/linking possible?

5.4 Multi-language databases

Currently, the database makes provision for Afrikaans and only a single field for English. It is feasible to use the concepts and database structures outlined here for other languages as well, as indicated above. It is therefore feasible to create multiple interlinked databases for fixed expressions in multiple languages. For translation purposes such multiple databases could be interlinked via a pivot language, for example English. Existing databases of fixed expressions could also be linked, even if the data fields in the different databases are not identical. The minimum requirement would be that there are at least a minimum set of corresponding fields, or that translation tables between different fields can be created.

6. Conclusion

Some of the envisaged expansions discussed above may not necessarily currently be commercially feasible since the time required to do the programming or to write / collate / select the data may simply be too much to complete the dictionary in a reasonable time. In addition, some of these expansions may not be what individual users may require. However, if researchers do not experiment with concepts and technologies that currently do not seem commercially realistic or feasible, innovation in e-information tools will be stifled. Such "blue sky" research could eventually lead to e-information tools that are not only incrementally better than those that are currently available, but provide different tools through disruptive innovation. The current project therefore has two aims:

- To create a database of fixed expressions, as well as to develop the necessary database tools, administrative backend, user interface and search functions, that enable users to have access to a number of monofunctional and one polyfunctional dictionaries. To result in a useful product this database and set of tools has to be completed in a limited timeframe (even though further extensions and updates need to be added regularly).
- To provide a platform to experiment with disruptive technologies and see to what extent any of these technologies can add value for the user in providing access to information in terms of the user's specific information need in a given user situation.

Such "blue sky" research is absolutely essential to ensure that not only better but different types of e-tools are developed. After all, the development of new cars is not left up to the drivers. One can ask drivers about which aspects of their cars they are not quite satisfied, and the designers and manufacturers of cars can then make the required improvements. However, drivers do not possess the know-how and the technical creativity that is necessary to design and develop cars that are totally new, much better and also manufactured quite differently. As Henry Ford allegedly said, "If I had asked people what they wanted, they would have said faster horses." e-Dictionaries are no different. Users may help to improve e-dictionaries incrementally, but only fundamental research in metalexicography, user needs, database technologies and principles of information organisation, access and retrieval will result in different types of e-tools.

7. References

Bergenholtz, H. (2010). Needs-Adapted Data access and data presentation. In *Doctorado Honoris Causa del Excmo. Sr. D. Henning Bergenholtz*. Valladolid, pp. 41-57.

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

Bergenholtz, H., Bergenholtz, I. (2011). A dictionary is a tool, a good dictionary is a monofunctional tool. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.)

e-Lexicography: The Internet, Digital Initiatives and Lexicography. 2011. London & New York: Continuum, pp. 188-207.

Bergenholtz, H., Bothma, T.J.D. (2011). Needs-adapted data presentation in e-information tools. *Lexikos*, in press.

Bergenholtz, H., Johnsen, M. (2005). Log files as a tool for improving Internet dictionaries. *Hermes*, 34, pp. 117-141.

Bergenholtz, H., Johnsen, M. (2007). Log files can and should be prepared for a functionalistic approach. *Lexikos*, 17, pp. 1-20.

Bergenholtz, H., Tarp, S. (2002). Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen. *Lexicographica*, 18, pp. 253-263.

Bergenholtz, H., Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes*, 31, pp. 171-196.

Bergenholtz, H., Tarp, S. (2005). Wörterbuchfunktionen. In I. Barz, H. Bergenholtz & J. Korhonen (eds.) *Schreiben, Verstehen, Übersetzen und Lernen: Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*. 2005. Frankfurt a.M./Bern/New York/Paris: Peter Lang, pp. 11-25.

Botha, R.P., Kroes, G. & Winckler, C.H. (1994). *Afrikaanse idiome en ander vaste uitdrukkings*. Halfweghuis: Southern.

Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. 2011. London & New York: Continuum, pp. 71-102.

De Villiers, M., Gouws, R.H. (1988). *Idiomewoordeboek*. Cape Town: Nasou.

Malherbe, D.F. (1924). *Afrikaanse spreekwoorde en verwante vorme*. Bloemfontein: Nasionale Pers.

Prinsloo, A.F. (1997). *Afrikaanse spreekwoorde en uitdrukkings*. Pretoria: J.L. van Schaik.

Prinsloo, A.F. (2009). *Spreekwoorde en waar hulle vandaan kom*. Cape Town: Pharos.

Tarp, S. (2002). Translation dictionaries and bilingual dictionaries. Two different concepts. *Journal of Translation Studies*, 7, pp. 59-84.

Tarp, S. (2007). Lexicography in the Information Age. *Lexikos*, 17, pp. 170-179.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learners' lexicography*. (Lexicographica. Series Maior 134). Tübingen: Max Niemeyer.

Tarp, S. (2009a). Reflections on lexicographical user research. *Lexikos*, 19, pp. 275-296.

Tarp, S. (2009b). Reflections on data access in lexicographic works. In S. Nielsen, S. Tarp (eds.) *Lexicography in the 21st Century. In Honour of Henning Bergenholtz*. (Terminology and Lexicography Research and Practice, Volume 12). 2009. Amsterdam: John Benjamins, pp. 43-65.

Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.) e-*Lexicography: The Internet, Digital Initiatives and Lexicography*. 2011. London & New York: Continuum, pp. 55-70.

Vrang, V., Bergenholtz, H. & Lund, L. (2003-2005). *Den danske Idiomordbog*. Database and layout: Richard Almind. www.idiomordbogen.dk.

Wiegand, H.E. (1977). Nachdenken über Wörterbücher. In *Nachdenken über Wörterbücher*. Mannheim/Wien/ Zürich, pp. 51-102.

Describing Linde's Dictionary of Polish for Digitalisation Purposes

Joanna Bilińska
Formal Linguistics Department, University of Warsaw,
Browarna 8/10, 00-927 Warsaw, Poland
E-mail: j.bilinska@uw.edu.pl

Abstract

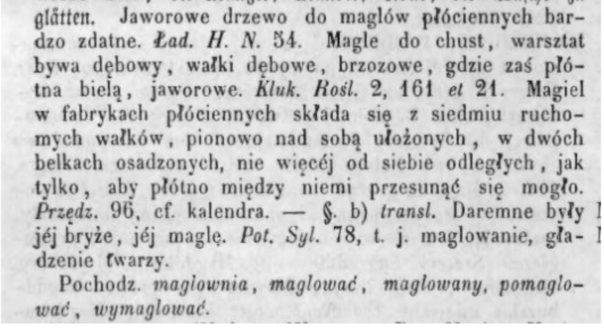
The present paper describes the attempts at digitalising the so called Linde's dictionary of Polish published in 6 volumes between 1807 and 1814 by Samuel Bogumił Linde. We are working on a formal description of the dictionary's structure, whose purpose will be to allow programmers to design a tool for automatic tagging of the text. The dictionary is multilingual, so performing OCR with good quality is a difficult task. The paper also describes the indexes that are going to be added. Compiling an a tergo index and indexes of abbreviations, qualifiers and the names of quotation authors would improve the quality and usefulness of the digitalised version. Our work with the 2nd edition of the dictionary (1854-1861) allows us to test several tools (in different stages of development) that are being developed within the framework of a Polish government grant directed by Janusz S. Bień.

Keywords: digitalisation; old dictionaries; Linde's dictionary of Polish

1. Linde's dictionary

The paper will demonstrate the attempts made to digitalise the so called Linde's dictionary of Polish (Linde, 1807-14) published in 6 volumes between 1807 and 1814 by Samuel Bogumił Linde. It was the first work of such kind for Polish and it met with excellent reception in Poland and abroad. Being a part of Polish cultural heritage, it ought to be represented in digital form to allow more people to get acquainted with it and to enable more advanced usage of it.

The author's intention was to present the Polish language extensively. The dictionary contains as much Polish vocabulary as the author was able to find in available texts. Every word was supplied with all typical pieces of information, such as grammatical properties, definition, quotations from source texts. Moreover, translations are provided into German (in Gothic), Slavic languages, and sometimes also other languages (e.g. Hebrew), as in the author's opinion they were useful to understand older Polish words. The dictionary is both descriptive and normative, because it includes additional information if a word is not used anymore or whether it is more likely to be encountered in poetry or in speech.



MAGIEL, glu, *m.*, *et* gli, *ż.*, *Boh.* mandl; *Vind.* likaunik, likalu, roloviſhè, povalaviſhe, povalilu; *Ross.* катокъ; §. a) narzędzie do gładzenia płótna lub chust pranych, wałkownica, die Mangel, Mandel, Rolle, die Wäſche zu glätten. Jaworowe drzewo do maglów płóciennych bardzo zdatne. *Ład. H. N.* 54. Magle do chust, warsztat bywa dębowy, wałki dębowe, brzozowe, gdzie zaś płótna bielą, jaworowe. *Kluk. Rośl.* 2, 161 *et* 21. Magiel w fabrykach płóciennych składa się z siedmiu ruchomych wałków, pionowo nad sobą ułożonych, w dwóch belkach osadzonych, nie więcéj od siebie odległych, jak tylko, aby płótno między niemi przesunąć się mogło. *Przędz.* 96, cf. kalendra. — §. b) *transl.* Daremne były jéj bryże, jéj maglę. *Pot. Syl.* 78, t. j. maglowanie, gładzenie twarzy.

Pochodz. *maglownia*, *maglować*, *maglowany*, *pomaglować*, *wymaglować*.

Figure 1: An example entry from Linde's dictionary

Due to its multilingualism, the dictionary's usefulness as a research resources is not only limited to Poland. It could be used for research purposes by any historian, librarian, or lexicographer interested in other (mainly Slavic) languages and cultures. Moreover, it can be interesting for other scholars studying old books, especially dictionaries, and people interested in digitising them.

We are trying to create a formal description of the dictionary's structure, as this could be used by programmers to tag the text automatically in terms of entry names and abbreviations, especially those naming the languages.

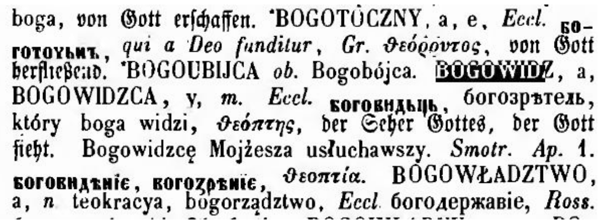
2. Digitalisations

Several digital versions of this dictionary exist in digital libraries, for example Google Books and Kujawsko-Pomorska Biblioteka Cyfrowa[1]. Both the 1st and the 2nd edition are freely available on the Internet but their quality is not perfect. Generally, they are available in scanned form with OCR that is far from perfect. As such, they are great examples of care of the Polish heritage but at the same time they are useless for research purposes.
However, since the dictionary contains a lot of vocabulary in other languages than Polish, different alphabets and fonts, it is very difficult to perform good OCR. Unfortunately, FineReader does not work sufficiently well with multilingual texts and does not recognise texts that are written in Gothic. While the OCR of Polish parts in the dictionary is good, the parts in other languages are virtually useless.

Furthermore, the book is old and even the 2nd (and last) edition, which was later only reprinted, comes from the 19th century (1854-1861). This results in print errors,

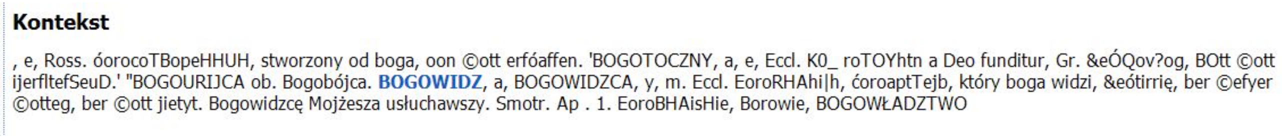
[1] http://books.google.pl/books?id=rs0GAAAAQAAJ&printsec=frontcover#v=onepage&q&f=false (Google Books) and http://kpbc.umk.pl/publication/8173

such as variable position of words when they are typed in other alphabets, etc. For example:



boga, von Gott erſchaffen. ˙BOGOTÓCZNY, a, e, *Eccl.* БО-ГОТОЧЬНЪ, *qui a Deo funditur*, *Gr.* *θεόῤῥοντος*, von Gott herfließend. ˙BOGOUBIJCA *ob.* Bogobójca. BOGOWIDZ, a, BOGOWIDZCA, y, *m.* *Eccl.* БОГОВИДЬЦЬ, богозрѣтель, który boga widzi, *θεόπτης*, der Seher Gottes, der Gott ſieht. Bogowidzcę Mojżesza usłuchawszy. *Smotr.* *Ap.* 1. БОГОВИДѢНІЄ, БОГОЗРѢНІЄ, *θεοπτία*. BOGOWŁADZTWO, a, *n* teokracya, bogorządztwo, *Eccl* богодержавіе, *Ross.*

Figure 2: Variable position of words



Kontekst

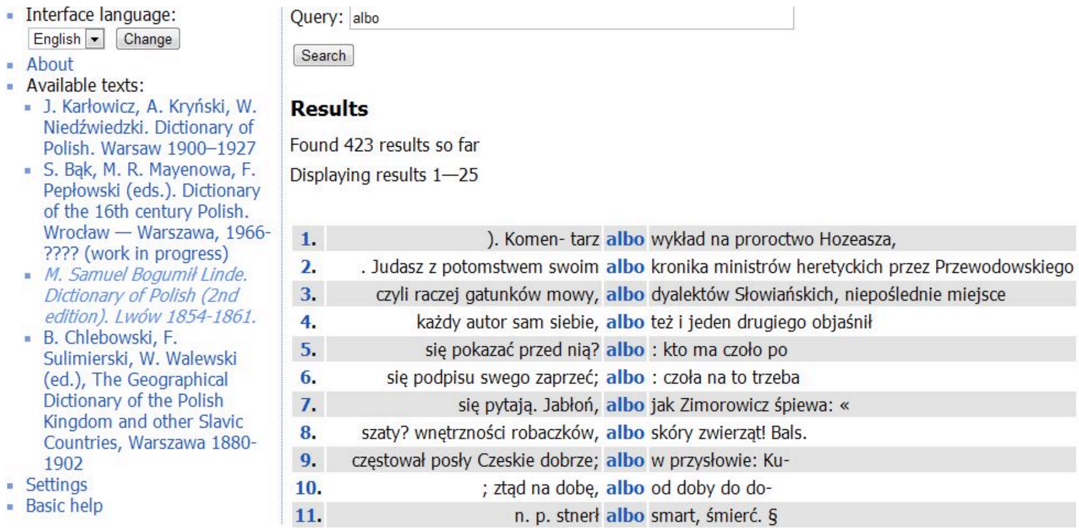
, e, Ross. óorocoTBopeHHUH, stworzony od boga, oon ©ott erfóaffen. 'BOGOTOCZNY, a, e, Eccl. K0_ roTOYhtn a Deo funditur, Gr. &eÓQov?og, BOtt ©ott ijerfltefSeuD.' "BOGOURIJCA ob. Bogobójca. **BOGOWIDZ**, a, BOGOWIDZCA, y, m. Eccl. EoroRHAhi|h, ćoroaptTejb, który boga widzi, &eótirrię, ber ©efyer ©otteg, ber ©ott jietyt. Bogowidzcę Mojżesza usłuchawszy. Smotr. Ap . 1. EoroBHAisHie, Borowie, BOGOWŁADZTWO

Figure 3: OCR quality

This is why this type of digitalisation needs to be done with several specialised tools in addition to a standard OCR program. Some of the tools used will be presented in the paper.

3. Linde's dictionary as a corpus

Searching the hidden text layer in large DjVu files, e.g. dictionaries, is not really efficient as it demands loading the whole file. It was decided that it would be much easier to treat the dictionaries' texts as corpora (Bień, 2011) and to use a specialised search engine for corpora. Therefore, recently a digital version of the 2nd edition of Linde's dictionary was made available at University of Warsaw, with a preliminary OCR (*SJPL,* 2010). This version of OCR was prepared with FineReader 10 (with 300 dpi resolution) and then converted from PDF to DjVu. It was then converted into a text corpus which now consists of ca. 7 million segments (http://poliqarp.wbl.klf.uw.edu.pl/slownik-lindego/).



- Interface language:
  English ▾ [Change]
- About
- Available texts:
  - J. Karłowicz, A. Kryński, W. Niedźwiedzki. Dictionary of Polish. Warsaw 1900–1927
  - S. Bąk, M. R. Mayenowa, F. Pepłowski (eds.). Dictionary of the 16th century Polish. Wrocław — Warszawa, 1966-???? (work in progress)
  - *M. Samuel Bogumił Linde. Dictionary of Polish (2nd edition). Lwów 1854-1861.*
  - B. Chlebowski, F. Sulimierski, W. Walewski (ed.), The Geographical Dictionary of the Polish Kingdom and other Slavic Countries, Warszawa 1880-1902
- Settings
- Basic help

Query: albo

[Search]

Results

Found 423 results so far

Displaying results 1—25

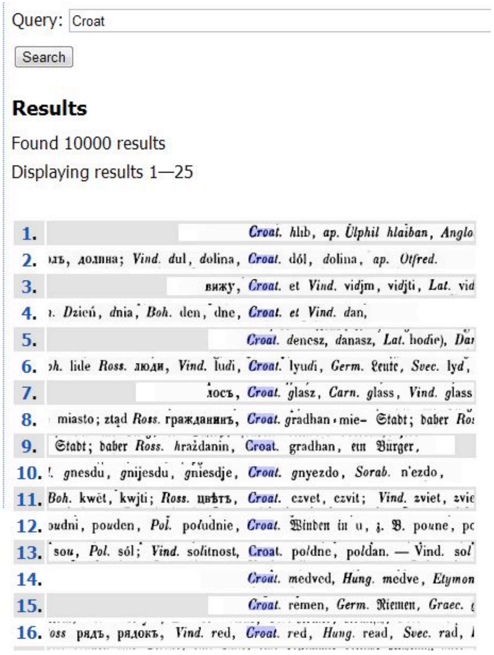
1.	). Komen- tarz	albo	wykład na proroctwo Hozeasza,
2.	. Judasz z potomstwem swoim	albo	kronika ministrów heretyckich przez Przewodowskiego
3.	czyli raczej gatunków mowy,	albo	dyalektów Słowiańskich, niepoślednie miejsce
4.	każdy autor sam siebie,	albo	też i jeden drugiego objaśnił
5.	się pokazać przed nią?	albo	: kto ma czoło po
6.	się podpisu swego zaprzeć;	albo	: czoła na to trzeba
7.	się pytają. Jabłoń,	albo	jak Zimorowicz śpiewa: «
8.	szaty? wnętrzności robaczków,	albo	skóry zwierząt! Bals.
9.	częstował posły Czeskie dobrze;	albo	w przysłowie: Ku-
10.	; ztąd na dobę,	albo	od doby do do-
11.	n. p. stnerł	albo	smart, śmierć. §

Figure 4: Lexicographic browser

The current version of the dictionary can be searched with the Poliqarp for DjVu browser search engine and a concordancer (also called marasca)[2], which allows users to browse the dictionaries as if they were corpora, returning lists of concordances as search results.

One of the most useful features of the search engine is that the query results can showed as graphical concordances (see figure 5). And as the results are linked with the scans, one can see the original page of the dictionary with the marked result. An example can be seen on figure 6.

Regular expressions can be used for searching, as it is one of the standard features of the Poliqarp concordancer[3], which Poliqarp for DjVu is based on. The query syntax is thus the same as in the Poliqarp version used for the National Corpus of Polish, which makes the tool easy to use for people familiar with the earlier version.



Query: Croat

Search

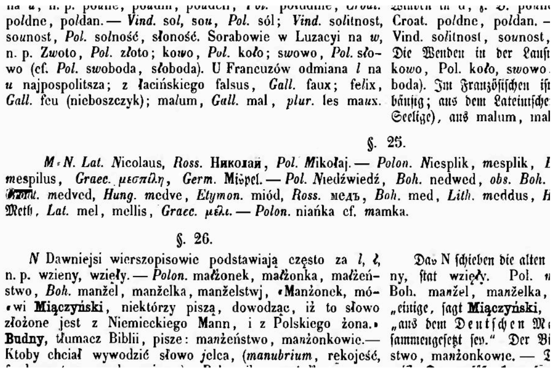
Results

Found 10000 results

Displaying results 1—25

1. Croat. hlib, ap. Ulphil hlaiban, Anglo
2. ɪлъ, долина; Vind. dul, dolina, Croat. dól, dolina, ap. Otfred.
3. вижу, Croat. et Vind. vidjm, vidjti, Lat. vid
4. ı. Dzień, dnia, Boh. den, dne, Croat. et Vind. dan,
5. Croat. dencsz, danasz, Lat. hodie), Dai
6. oh. lide Ross. люди, Vind. ludi, Croat. lyudi, Germ. Leute, Svec. lyd,
7. лосъ, Croat. glasz, Carn. glass, Vind. glass
8. miasto; ztąd Ross. гражданинъ, Croat. gradhan, mie- Stadt; daher Ro
9. Stadt; daher Ross. hraždanin, Croat. gradhan, ein Bürger,
10. !. gnesdu, gnijesdu, gniesdje, Croat. gnyezdo, Sorab. n'ezdo,
11. Boh. kwět, kwjti; Ross. цвѣтъ, Croat. czvet, czvit; Vind. zviet, zvie
12. oudni, pouden, Pol. południe, Croat. Winden in u, z. B. poune, pc
13. sou, Pol. sól; Vind. solitnost, Croat. poldne, poldan. — Vind. sol
14. Croat. medved, Hung. medve, Etymon
15. Croat. remen, Germ. Riemen, Graec.
16. oss рядъ, рядокъ, Vind. red, Croat. red, Hung. read, Svec. rad,

Figure 5: Graphical concordances



poldne, poldan. — Vind. sol, sou, Pol. sól; Vind. solitnost, sounost, Pol. solność, słoność. Sorabowie w Luzacyi na w, n. p. Zwoto, Pol. złoto; kowo, Pol. koło; swowo, Pol. słowo (cf. Pol. swoboda, sloboda). U Francuzów odmiana l na u najpospolitsza; z łacińskiego falsus, Gall. faux; felix, Gall. feu (nieboszczyk); malum, Gall. mal, plur. les maux.

Croat. poldne, poldan. – Vind. solitnost, sounost, Die Wenden in der Lauſi kowo, Pol. koło, swowo boda). Im Franzöſiſchen iſt häufig; aus dem Lateiniſche Seelige), aus malum, mal

§. 25.

M·N. Lat. Nicolaus, Ross. Николай, Pol. Mikołaj. — Polon. Niesplik, mesplik, mespilus, Graec. μεσπιλη, Germ. Miſpel. — Pol. Niedźwiedź, Boh. nedwed, obs. Boh. Croat. medved, Hung. medve, Etymon. miód, Ross. медъ, Boh. med, Lith. meddus, Meth, Lat. mel, mellis, Graec. μελι. — Polon. niańka cf. mamka.

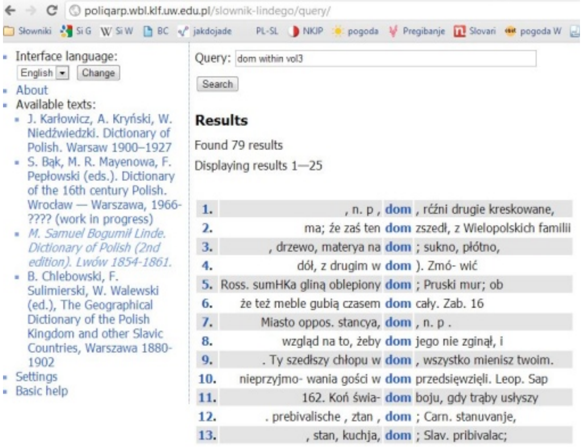
§. 26.

N Dawniejsi wierszopisowie podstawiają często za l, ł, n. p. wzieny, wzięły. — Polon. małżonek, małżonka, małżeństwo, Boh. manžel, manželka, manželstwj, «Manżonek, mó- «wi Miączyński, niektórzy piszą, dowodząc, iż to słowo złożone jest z Niemieckiego Mann, i z Polskiego żona.» Budny, tłumacz Biblii, pisze: manżeństwo, manżonkowie.— Ktoby chciał wywodzić słowo jelca, (manubrium, rękojeść,

Das N ſchieben die alten ny, ſtat wzięły. Pol. Boh. manžel, manželka, „einige, ſagt Miączyński, „aus dem Deutſchen M ſammengeſetzt ſey.“ Der Bi stwo, manżonkowie. —

Figure 6: Highlighted query result on a DjVu file

For the time being, in the official digital version of Linde's dictionary it is possible to limit the search to individual volumes and the main sections of the dictionary (such as introductory texts, errata, main text, etc.) using the clause "within". Below there is an example of searching for the word "dom" (house) within the 3rd volume.



- Interface language: English Change
- About
- Available texts:
  - J. Karłowicz, A. Kryński, W. Niedźwiedzki. Dictionary of Polish. Warsaw 1900–1927
  - S. Bąk, M. R. Mayenowa, F. Pepłowski (eds.). Dictionary of the 16th century Polish. Wrocław — Warszawa, 1966-???? (work in progress)
  - *M. Samuel Bogumił Linde. Dictionary of Polish (2nd edition). Lwów 1854-1861.*
  - B. Chlebowski, F. Sulimierski, W. Walewski (ed.), The Geographical Dictionary of the Polish Kingdom and other Slavic Countries, Warszawa 1880-1902
- Settings
- Basic help

Query: dom within vol3

Search

Results

Found 79 results

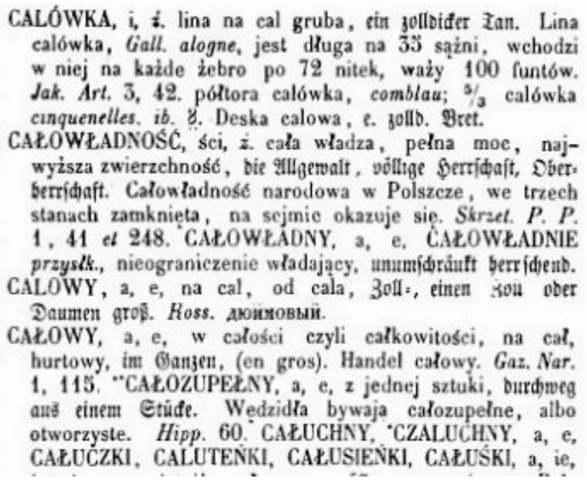
Displaying results 1—25

1.	, n. p ,	dom	, rćźni drugie kreskowane,
2.	ma; że zaś ten	dom	zszedł, z Wielopolskich familii
3.	, drzewo, materya na	dom	; sukno, płótno,
4.	dół, z drugim w	dom	). Zmó- wić
5.	Ross. sumHKa gliną oblepiony	dom	; Pruski mur; ob
6.	że też meble gubią czasem	dom	cały. Zab. 16
7.	Miasto oppos. stancya,	dom	, n. p .
8.	wzgląd na to, żeby	dom	jego nie zginął, i
9.	. Ty szedłszy chłopu w	dom	, wszystko mienisz twoim.
10.	nieprzyjmo- wania gości w	dom	przedsięwzięli. Leop. Sap
11.	162. Koń świa-	dom	boju, gdy trąby usłyszy
12.	. prebivalische , ztan ,	dom	; Carn. stanuvanje,
13.	, stan, kuchja,	dom	; Slav. pribivalac;

Figure 7: Usage of the "within" clause

4. Preparing Linde's dictionary for the users

The dictionary is mainly alphabetical but in the entries one can find also derivates, diminutives, etc. which means that the word order is not strict. Moreover, the Polish characters are ignored which makes the paper dictionary quite difficult to use.



CALÓWKA, i, ż. lina na cal gruba, ein zolldicker Tau. Lina calówka, Gall. alogne, jest długa na 33 sążni, wchodzi w niej na każde żebro po 72 nitek, waży 100 funtów. Jak. Art. 3, 42. półtora calówka, comblau; 5/3 calówka cinquenelles. ib. 8. Deska calowa, e. zolld. Bret.
CAŁOWŁADNOŚĆ, ści, ż. cała władza, pełna moc, najwyższa zwierzchność, die Allgewalt, völlige Herrſchaft, Oberherrſchaft. Całowładność narodowa w Polszcze, we trzech stanach zamknięta, na sejmie okazuje się. Skrzet. P. P. 1, 41 et 248. CAŁOWŁADNY, a, e, CAŁOWŁADNIE przysłk., nieograniczenie władający, unumſchränkt herrſchend.
CALOWY, a, e, na cal, od cala, Zoll-, einen Zoll oder Daumen groß. Ross. дюймовый.
CAŁOWY, a, e, w całości czyli całkowitości, na cał, hurtowy, im Ganzen, (en gros). Handel całowy. Gaz. Nar. 1, 115. *CAŁOZUPEŁNY, a, e, z jednej sztuki, durchweg aus einem Stücke. Wędzidła bywaja całozupełne, albo otworzyste. Hipp. 60. CAŁUCHNY, *CZALUCHNY, a, e, CAŁUCZKI, CALUTEŃKI, CAŁUSIEŃKI, CAŁUŚKI, a, ie,

Figure 8: Ordering of lemmas

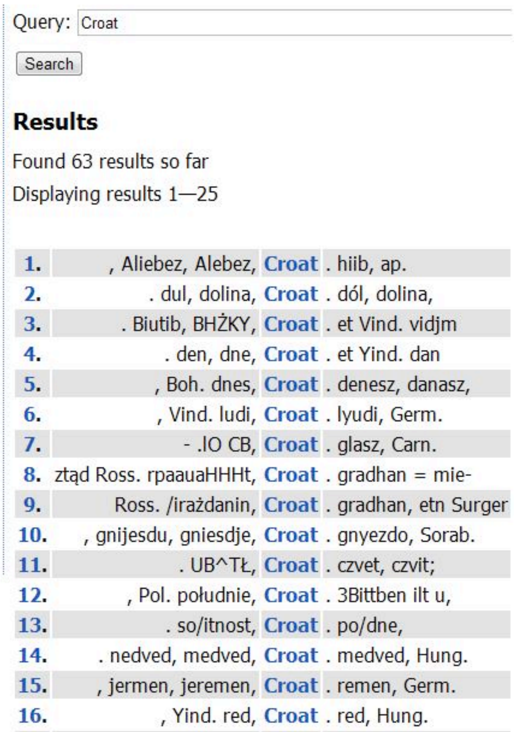
Considering this, it would be useful to link with it the *a tergo* index (Grzegorczykowa, 1965), after reverting it to the alphabetical order. It would become a kind of a list of content for the dictionary. The index is already scanned and the OCR was proofread.

[2] http://poliqarp.wbl.klf.uw.edu.pl/
[3] http://korpus.pl/index.php?lang=en&page=poliqarp

It is important to note that as a result of being multilingual and historic, the dictionary can be valuable especially for Slavonic scholars. For instance, Linde represented the language spoken on the territory of contemporary Slovenia (non-codified at that time) as two separate languages, which he labelled *Crn., Carn. Carniolice* for western dialect and *Vind., Vd.* for eastern dialect. However, the author did not provide a complete list of abbreviations and he used various abbreviations for the same thing (eg. *Boh., Bh.* and *Cz.* for Czech language).

Thus an index of abbreviations is being prepared for the language names. The list could be applied to the OCR version of a dictionary to tag the sections with specific language labels. It could then be used to narrow the search to the specific sections.

At the moment one can simply search the dictionary for the language labels (just typing the abbreviation for the language) and hence get a list of words in this language contained in Linde's dictionary (as the right context in concordances). However, it would be much easier and more comfortable to search within the dictionary parts labelled with language names.



Results

Found 63 results so far

Displaying results 1—25

1.	, Aliebez, Alebez,	Croat	. hiib, ap.
2.	. dul, dolina,	Croat	. dól, dolina,
3.	. Biutib, BHŻKY,	Croat	. et Vind. vidjm
4.	. den, dne,	Croat	. et Yind. dan
5.	, Boh. dnes,	Croat	. denesz, danasz,
6.	, Vind. ludi,	Croat	. lyudi, Germ.
7.	- .IO CB,	Croat	. glasz, Carn.
8.	ztąd Ross. rpaauaHHHt,	Croat	. gradhan = mie-
9.	Ross. /irażdanin,	Croat	. gradhan, etn Surger
10.	, gnijesdu, gniesdje,	Croat	. gnyezdo, Sorab.
11.	. UB^TŁ,	Croat	. czvet, czvit;
12.	, Pol. południe,	Croat	. 3Bittben ilt u,
13.	. so/itnost,	Croat	. po/dne,
14.	. nedved, medved,	Croat	. medved, Hung.
15.	, jermen, jeremen,	Croat	. remen, Germ.
16.	, Yind. red,	Croat	. red, Hung.

Figure 9: Searching for "Croat" label (Croatian)

It does not seem as if there were any strict rules for the language order within the entries, but we are trying to find regularities in it. With the browser it is possible to search for the language labels and observe whether they appear in the same order all the time. Of course, the same is possible with the paper version of a dictionary, but an electronic version combined with a specialised browser considerably accelerates and facilitates the task.

Preparing a list of other abbreviations is also planned. It could be implemented into the electronic version of the dictionary, enabling the reader to see the meaning of an abbreviation by clicking on it or pointing the mouse on it.

The author did not prepare such a list as he found it obvious and understandable to everybody. Some information on it was published in (Matuszczyk, 2006), but it is not complete and is not aimed to be a list but just a description of the kinds of labels in the dictionary.

Linde's dictionary is well known also because the author included a great number of examples from literature, adding the authors' names as abbreviations. A detailed list of such abbreviations and text titles was provided by the author himself and in the book of Hrabec and Pepłowski (Hrabec & Pepłowski, 1963). It would be very useful if those texts were linked, as there are many researchers working on describing how the texts of the authors they are interested in are exemplified in dictionaries. In the future it would be possible to link also the abbreviations with the digitalised texts in digital libraries.

The abbreviations (language names, quotations authors, qualifiers) are mainly in italics, and once again the prototype of a newer lexicographic browser with font tagging (see page 5) is of a great help, as it allows for searching for a particular font shape.

Furthermore, it would be useful to work out the meaning of punctuation marks which were used by Linde. He used commas, semicolons and ellipsis extensively and generally the usage is clear. However, there are three types of section marks too (two normal and one mirrored), and it would be of much help if we were sure what is the meaning of them. Maybe it is just for the lack of the same fonts but maybe there is a deeper sense.

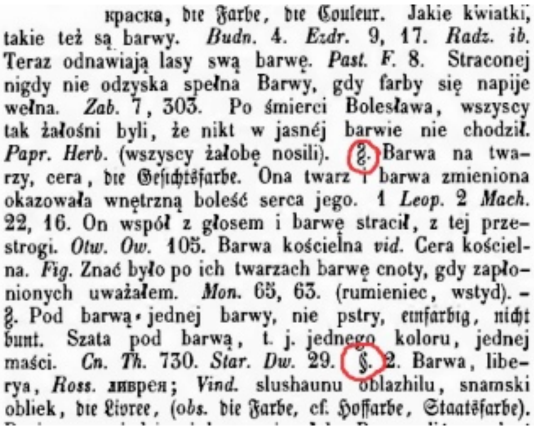

краска, die Farbe, die Couleur. Jakie kwiatki, takie też są barwy. *Budn.* 4. *Ezdr.* 9, 17. *Radz. ib.* Teraz odnawiają lasy swą barwę. *Past. F.* 8. Straconej nigdy nie odzyska spełna Barwy, gdy farby się napije wełna. *Zab.* 7, 303. Po śmierci Bolesława, wszyscy tak żałośni byli, że nikt w jasnéj barwie nie chodził. *Papr. Herb.* (wszyscy żałobę nosili). §. Barwa na twarzy, cera, die Geſichtsfarbe. Ona twarz i barwa zmieniona okazowała wnętrzną boleść serca jego. 1 *Leop.* 2 *Mach.* 22, 16. On współ z głosem i barwę stracił, z tej przestrogi. *Otw. Ow.* 105. Barwa kościelna *vid.* Cera kościelna. *Fig.* Znać było po ich twarzach barwę cnoty, gdy zapłonionych uważałem. *Mon.* 65, 63. (rumieniec, wstyd). – §. Pod barwą · jednej barwy, nie pstry, einfarbig, nicht bunt. Szata pod barwą, t. j. jednego koloru, jednej maści. *Cn. Th.* 730. *Star. Dw.* 29. §. 2. Barwa, liberya, *Ross.* ливрея; *Vind.* slushaunu oblazhilu, snamski obliek, die Livree, (*obs.* die Farbe, cf. Hoffarbe, Staatsfarbe).

Figure 10: Two of three types of section marks

5. Tools used for digitising

Several tools were used for the purposes of a new digital version of Linde’s dictionary.

For scanning the 2nd edition of Linde’s dictionary in 600 dpi resolution we used Scanhelper[4], a command-line scanning tool prepared by Jakub Wilk. These scans are available at the digital library: http://eprints.wbl.klf.uw.edu.pl/view/creators/Linde=3A Samuel_Bogumi==0142=3A=3A.html.

The preliminary version of the digitalisation which is available for searching with a lexicographic browser (http://poliqarp.wbl.klf.uw.edu.pl/slownik-lindego/)[5] was prepared in 300 dpi resolution using Kofax VRS software.

The scans were then converted into DjVu which is thought to be the best format for scanned texts, as it allows for instance to store the OCR results in a hidden text layer. Moreover, the documents can be stored as individual pages, which facilitates the usage of the files, especially those containing large texts.

In the framework of the project mentioned earlier, Grzegorz Chimosz prepared a prototype of a DjVu files viewer based on DjView4 which is, at the same time, a Poliqarp client and thus is an alternative for traditional usage of Poliqarp through a web browser.

The results of a query in the form of graphical concordances are shown in the left panel, and in the main panel there is the DjVu file with highlighted results. An example is provided in Figure 11.

When pointing on a word, one can see underneath the hidden text layer (OCR). Moreover, the viewer enables the user to mark the result and copy its URL (Figure 13).

Later, thanks to Tomasz Olejniczak’s pdfa2hocr[6] converter, Linde’s dictionary in PDF/A files from FineReader was converted into hOCR files. This was done because Poliqarp for DjVu is now being prepared to enable the user to search for fonts. For this the font-sensitive corpus builder by Marcin Zając (Janusz S. Bień’s student) is being used[7]. The program enables converting hOCR files into a corpus for Poliqarp concordancer.

Currently this browser is only used for experiments but it looks very useful for our research purposes. Figures 14, 15, and 16 below show the possible queries and the results.

As you can see on the figures, the OCR program recognizes the little differences between the fonts, although they generally seem the same when looked at by humans and are of no importance within a text. However, with regular expressions one can search for example for all the text typed in a specified font size range (for example all words in font size 12 and more).

Apart from searching for font sizes, users can search for font shape. Figure 16 shows an example of a query for all the words in bold italic.

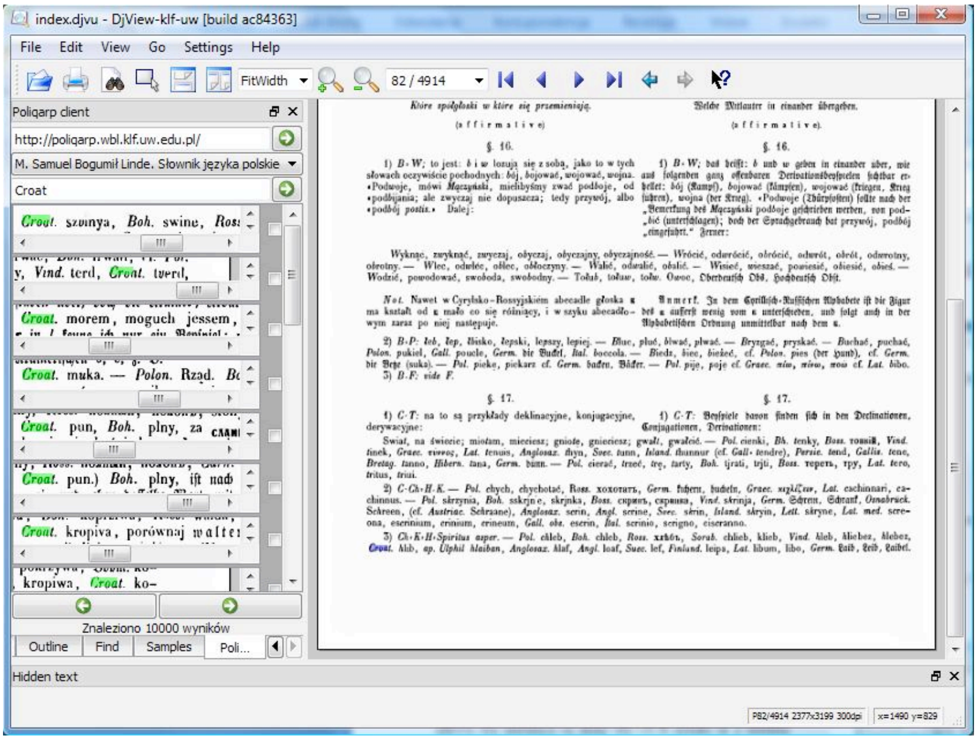
This feature of the experimental version of the lexicographic browser is of great usefulness for preparing indexes for the dictionary. For example, searching for a text in italic we will get generally language names’ abbreviations, qualifiers and quotation authors abbreviations.

[4] http://jwilk.net/software/scanhelper

[5] http://poliqarp.wbl.klf.uw.edu.pl/extra/linde/index.djvu

[6] http://students.mimuw.edu.pl/~to236111/PDFAUtilities/

[7] https://github.com/mzajac/FSCB



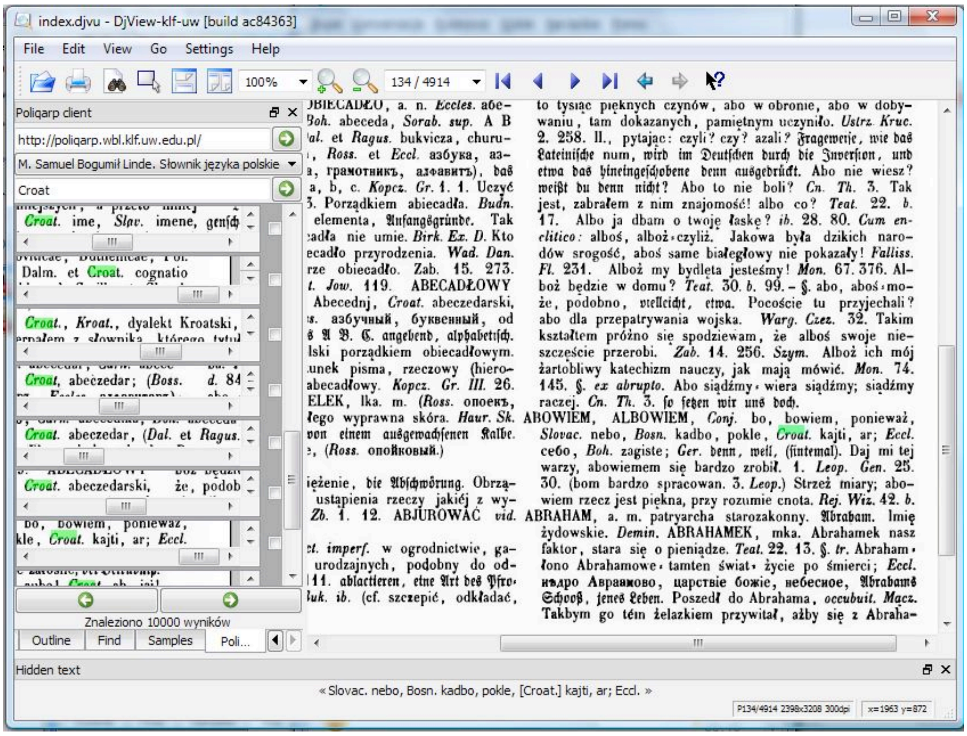
http://poliqarp.wbl.klf.uw.edu.pl/

M. Samuel Bogumił Linde. Słownik języka polskie

Croat

Znaleziono 10000 wyników

Figure 11: Poliqarp client



http://poliqarp.wbl.klf.uw.edu.pl/

M. Samuel Bogumił Linde. Słownik języka polskie

Croat

Znaleziono 10000 wyników

« Slovac. nebo, Bosn. kadbo, pokle, [Croat.] kajti, ar; Eccl. »

Figure 12: Hidden text

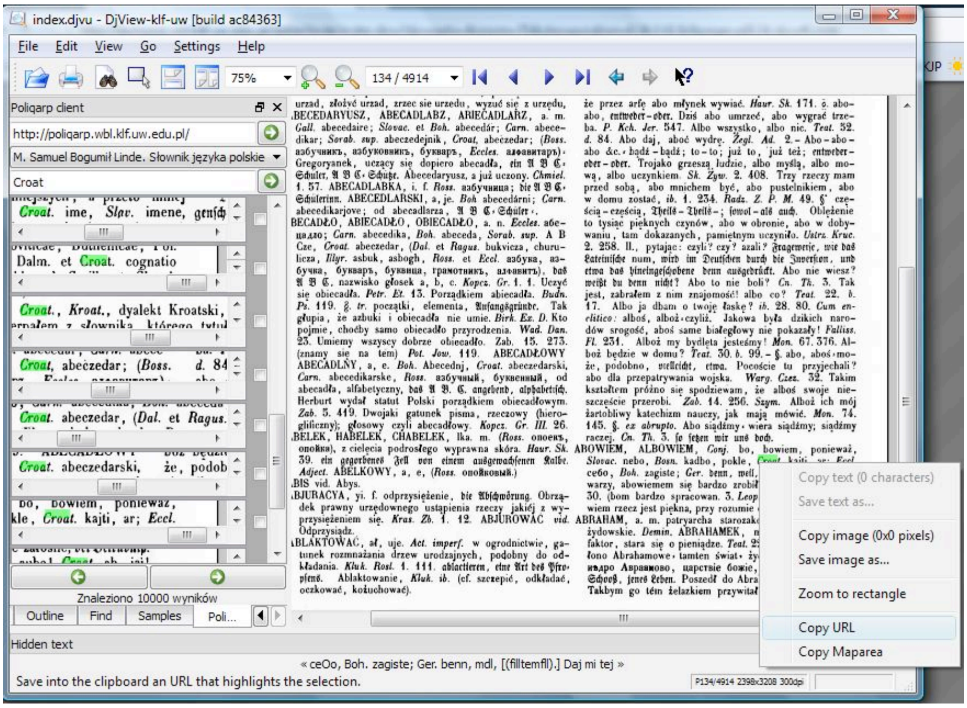
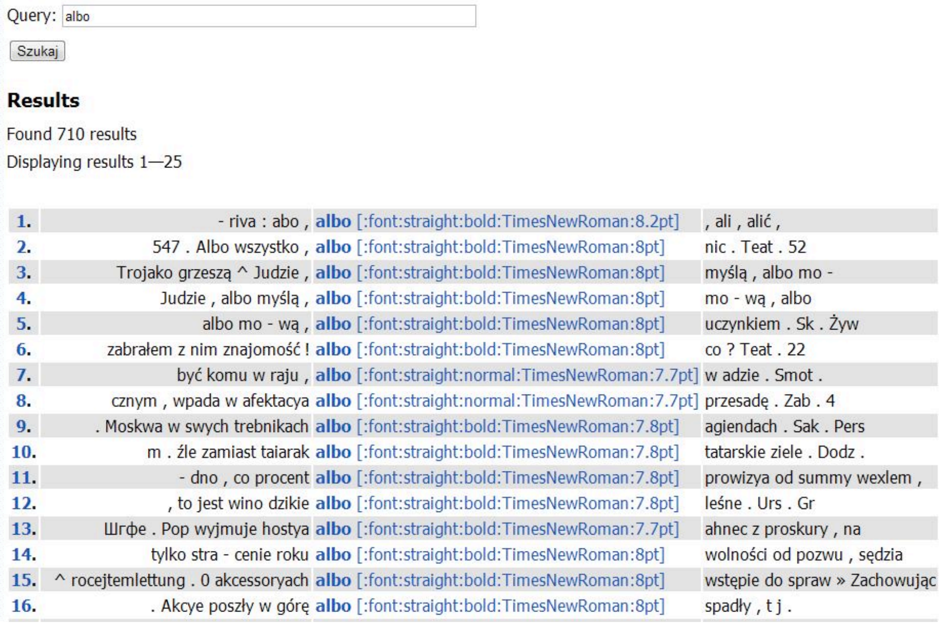

index.djvu - DjView-klf-uw [build ac84363]
File Edit View Go Settings Help
75%
134 / 4914
Poliqarp client
http://poliqarp.wbl.klf.uw.edu.pl/
M. Samuel Bogumił Linde. Słownik języka polskie
Croat
Znaleziono 10000 wyników
Outline Find Samples Poli...
Hidden text
Copy text (0 characters)
Save text as...
Copy image (0x0 pixels)
Save image as...
Zoom to rectangle
Copy URL
Copy Maparea
« ceOo, Boh. zagiste; Ger. benn, mdl, [(filltemfl).] Daj mi tej »
Save into the clipboard an URL that highlights the selection.
P134/4914 2398x3208 300dpi

Figure 13: Copy URL feature


Query: albo

Szukaj

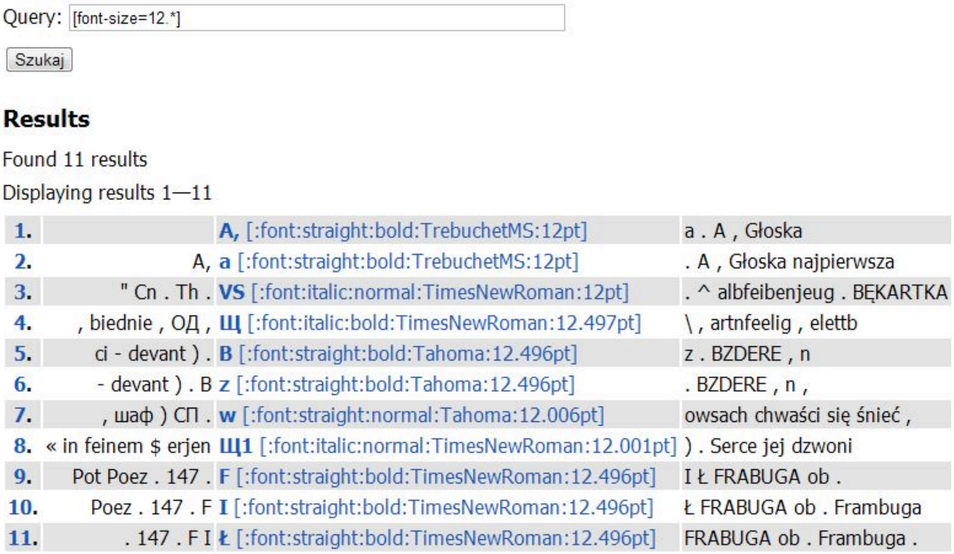
Results

Found 710 results

Displaying results 1—25

1.	- riva : abo ,	**albo** [:font:straight:bold:TimesNewRoman:8.2pt]	, ali , alić ,
2.	547 . Albo wszystko ,	**albo** [:font:straight:bold:TimesNewRoman:8pt]	nic . Teat . 52
3.	Trojako grzeszą ^ Judzie ,	**albo** [:font:straight:bold:TimesNewRoman:8pt]	myślą , albo mo -
4.	Judzie , albo myślą ,	**albo** [:font:straight:bold:TimesNewRoman:8pt]	mo - wą , albo
5.	albo mo - wą ,	**albo** [:font:straight:bold:TimesNewRoman:8pt]	uczynkiem . Sk . Żyw
6.	zabrałem z nim znajomość !	**albo** [:font:straight:bold:TimesNewRoman:8pt]	co ? Teat . 22
7.	być komu w raju ,	**albo** [:font:straight:normal:TimesNewRoman:7.7pt]	w adzie . Smot .
8.	cznym , wpada w afektacya	**albo** [:font:straight:normal:TimesNewRoman:7.7pt]	przesadę . Zab . 4
9.	. Moskwa w swych trebnikach	**albo** [:font:straight:bold:TimesNewRoman:7.8pt]	agiendach . Sak . Pers
10.	m . źle zamiast taiarak	**albo** [:font:straight:bold:TimesNewRoman:7.8pt]	tatarskie ziele . Dodz .
11.	- dno , co procent	**albo** [:font:straight:bold:TimesNewRoman:7.8pt]	prowizya od summy wexlem ,
12.	, to jest wino dzikie	**albo** [:font:straight:bold:TimesNewRoman:7.8pt]	leśne . Urs . Gr
13.	Шrфe . Pop wyjmuje hostya	**albo** [:font:straight:bold:TimesNewRoman:7.7pt]	ahnec z proskury , na
14.	tylko stra - cenie roku	**albo** [:font:straight:bold:TimesNewRoman:8pt]	wolności od pozwu , sędzia
15.	^ rocejtemlettung . 0 akcessoryach	**albo** [:font:straight:bold:TimesNewRoman:8pt]	wstępie do spraw » Zachowując
16.	. Akcye poszły w górę	**albo** [:font:straight:bold:TimesNewRoman:8pt]	spadły , t j .

Figure 14: Lexicographic browser with font tagging



Query: [font-size=12.*]

Szukaj

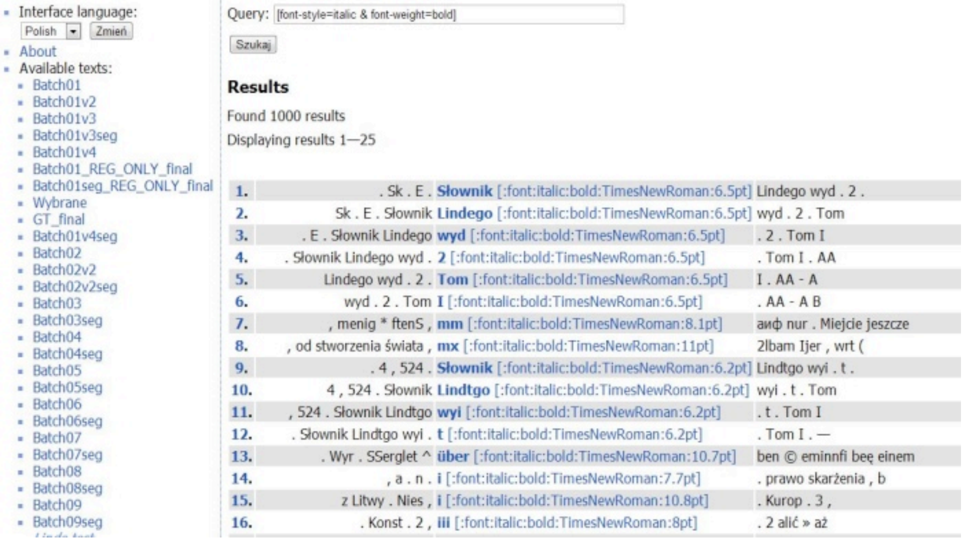
Results

Found 11 results

Displaying results 1—11

1.		**A,** [:font:straight:bold:TrebuchetMS:12pt]	a . A , Głoska
2.	A,	**a** [:font:straight:bold:TrebuchetMS:12pt]	. A , Głoska najpierwsza
3.	" Cn . Th .	**VS** [:font:italic:normal:TimesNewRoman:12pt]	. ^ albfeibenjeug . BĘKARTKA
4.	, biednie , ОД ,	**Щ** [:font:italic:bold:TimesNewRoman:12.497pt]	\ , artnfeelig , elettb
5.	ci - devant) .	**B** [:font:straight:bold:Tahoma:12.496pt]	z . BZDERE , n
6.	- devant) . B	**z** [:font:straight:bold:Tahoma:12.496pt]	. BZDERE , n ,
7.	, шаф) СП .	**w** [:font:straight:normal:Tahoma:12.006pt]	owsach chwaści się śnieć ,
8.	« in feinem $ erjen	**Щ1** [:font:italic:normal:TimesNewRoman:12.001pt]	) . Serce jej dzwoni
9.	Pot Poez . 147 .	**F** [:font:straight:bold:TimesNewRoman:12.496pt]	I Ł FRABUGA ob .
10.	Poez . 147 . F	**I** [:font:straight:bold:TimesNewRoman:12.496pt]	Ł FRABUGA ob . Frambuga
11.	. 147 . F I	**Ł** [:font:straight:bold:TimesNewRoman:12.496pt]	FRABUGA ob . Frambuga .

Figure 15: Searching for words in font size 12.



- Interface language:
  Polish ▾ Zmień
- About
- Available texts:
  - Batch01
  - Batch01v2
  - Batch01v3
  - Batch01v3seg
  - Batch01v4
  - Batch01_REG_ONLY_final
  - Batch01seg_REG_ONLY_final
  - Wybrane
  - GT_final
  - Batch01v4seg
  - Batch02
  - Batch02v2
  - Batch02v2seg
  - Batch03
  - Batch03seg
  - Batch04
  - Batch04seg
  - Batch05
  - Batch05seg
  - Batch06
  - Batch06seg
  - Batch07
  - Batch07seg
  - Batch08
  - Batch08seg
  - Batch09
  - Batch09seg

Query: [font-style=italic & font-weight=bold]

Szukaj

Results

Found 1000 results

Displaying results 1—25

1.	. Sk . E .	**Słownik** [:font:italic:bold:TimesNewRoman:6.5pt]	Lindego wyd . 2 .
2.	Sk . E . Słownik	**Lindego** [:font:italic:bold:TimesNewRoman:6.5pt]	wyd . 2 . Tom
3.	. E . Słownik Lindego	**wyd** [:font:italic:bold:TimesNewRoman:6.5pt]	. 2 . Tom I
4.	. Słownik Lindego wyd .	**2** [:font:italic:bold:TimesNewRoman:6.5pt]	. Tom I . AA
5.	Lindego wyd . 2 .	**Tom** [:font:italic:bold:TimesNewRoman:6.5pt]	I . AA - A
6.	wyd . 2 . Tom	**I** [:font:italic:bold:TimesNewRoman:6.5pt]	. AA - A B
7.	, menig * ftenS ,	**mm** [:font:italic:bold:TimesNewRoman:8.1pt]	анф nur . Miejcie jeszcze
8.	, od stworzenia świata ,	**mx** [:font:italic:bold:TimesNewRoman:11pt]	2lbam Ijer , wrt (
9.	. 4 , 524 .	**Słownik** [:font:italic:bold:TimesNewRoman:6.2pt]	Lindtgo wyi . t .
10.	4 , 524 . Słownik	**Lindtgo** [:font:italic:bold:TimesNewRoman:6.2pt]	wyi . t . Tom
11.	, 524 . Słownik Lindtgo	**wyi** [:font:italic:bold:TimesNewRoman:6.2pt]	. t . Tom I
12.	. Słownik Lindtgo wyi .	**t** [:font:italic:bold:TimesNewRoman:6.2pt]	. Tom I . —
13.	. Wyr . SSerglet ^	**über** [:font:italic:bold:TimesNewRoman:10.7pt]	ben © eminnfi beę einem
14.	, a . n .	**i** [:font:italic:bold:TimesNewRoman:7.7pt]	. prawo skarżenia , b
15.	z Litwy . Nies ,	**i** [:font:italic:bold:TimesNewRoman:10.8pt]	. Kurop . 3 ,
16.	. Konst . 2 ,	**iii** [:font:italic:bold:TimesNewRoman:8pt]	. 2 alić » aż

Figure 16: Searching for words in bold italic

6. Conclusion

As it was mentioned, it is not an easy task to digitise such a dictionary, but we hope that the ongoing formal analysis and lexicographical description of the book will help to better understand the structure of the dictionary and therefore prepare better tools to work with it.

The combination of 1) the indexes to the dictionary that are being prepared with 2) DjVu files of Linde's dictionary and 3) the lexicographic browser with many features valuable for researchers will considerably improve the quality and usefulness of digital dictionary.

While studying the structure of the dictionary (both macro- and microstructure), we are using computer tools already prepared for digitalisation purposes. This both facilitates the analysis, and allows to test the tools and develop them further. They could be later used for other digitalisation purposes.

In our work we are going to follow the best available examples of the old text digitalisations. We hope that our experience will be of some help for other researchers working on early dictionaries.

7. Acknowledgements

Development of most of the tools described in the present paper is supported by the Polish Ministry of Science and Higher Education's grant no. N N519 384036 *Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych* directed by Janusz S. Bień. The grant supports also my participation in the conference *Electronic lexicography in the 21st century: new applications for new users (eLEX2011)*.

The paper was proofread by Radosław Moszczyński.

8. References

Bień, J.S. (2009). Digitalizing dictionaries of Polish. In K. Bogacki, J. Cholewa & A. Rozumko (eds.) *Methods of Lexical Analysis: Theoretical Assumption and Practical Applications*. Wydawnictwo Uniwersytetu w Białymstoku, Białystok, pp. 37-45. Accessed at: http://bc.klf.uw.edu.pl/71/.

Bień, J.S. (2011). Efficient search in hidden text of large DjVu documents. In *Advanced Language Technologies for Digital Libraries. Lecture Notes in Computer Science (Theoretical Computer Science and General Issues)* (6699). Springer, pp. 1-14. Accessed at: http://bc.klf.uw.edu.pl/177/.

Grzegorczykowa, R., Kurzowa, Z., Puzynina, J. & Doroszewski, W. (1965). *Indeks a tergo do Słownika języka polskiego S.B. Lindego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Hrabec, S., Pepłowski, F. (1963). *Wiadomości o autorach i dziełach cytowanych w Słowniku Lindego*. Warszawa : Wiedza Powszechna.

Linde, S.B. (1807-1814). *Słownik języka polskiego.* 1st ed., Warszawa.

Linde, S.B. (1854-1860). *Słownik języka polskiego, 2nd ed.* Lwów: Zakład Narodowy im. Ossolińskich. Accessed at: http://kpbc.umk.pl/publication/8173.

Linde, S.B. (1994). *Słownik języka polskiego, 2nd ed., reprint*. Warszawa: Wydawnictwo «Gutenberg-Print». Accessed at: http://kpbc.umk.pl/publication/8173.

Matuszczyk, B. (2006). *Słownik języka polskiego S.B. Lindego*. Warsztat leksykografa, Lublin: Wydawnictwo KUL.

Ptaszyk, M. (2007). *Słownik języka polskiego Samuela Bogumiła Lindego*. Szkice bibliologiczne, Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

SJPL (2010) Linde, S.B., Słownik języka polskiego, digital version under search engine, Accessed at: http://poliqarp.wbl.klf.uw.edu.pl/extra/linde/index.djvu.

Hooking up to the corpus: the *Viennese Lexicographic Editor's* corpus interface

Gerhard Budin[1,2], Karlheinz Mörth[1]
1 Institute for Corpus Linguistics and Text Technology (ICLTT), Austrian Academy of Sciences
Sonnenfelsgasse 19/8, 1010 Vienna, Austria
2 Center for Translation Studies, University of Vienna
Gymnasiumstraße 50, 1190 Vienna
E-mail: karlheinz.moerth@oeaw.ac.at, gerhard.budin@oeaw.ac.at, gerhard.budin@univie.ac.at

Abstract

The paper addresses the issue of interfacing between digital corpora and a new dictionary writing application being developed at the ICLTT (Institute of Corpus Linguistics and Text Technology of the Austrian Academy of Sciences). It deals with issues of dictionary creation, software design, usability and interoperability in relation to the example of this fairly new piece of software, the *Viennese Lexicographic Editor*. In addition, it outlines the ICLTT's projects which are using the new tool and explains the role of these undertakings as part of the ICLTT's involvement in the Austrian CLARIN-AT initiative. The focus of the discussion will be on the implementation of efficient workflows designed to streamline the transfer of corpus examples stemming from distributed online corpora into dictionary entries. An important additional topic is the access mode of digital corpora available on the internet and the issue of service-oriented software design. As a last point, we will also touch on the important issue of standards and de-facto standards used in these projects.

Keywords: dictionary editor; dictionary software; CLARIN; LRT standards; TEI

1. Introduction

The Institute for Corpus Linguistics and Text Technology (ICLLT), which was founded in early 2010, is among the youngest departments of the Austrian Academy of Sciences. It is the successor of the Austrian Academy Corpus (AAC) and has been designed to pursue research in a number of cross-disciplinary projects which cover a wide range of interests. Some of these projects focus on issues of corpustechnology, some on computational lexicography, yet others belong to the sphere of humanities computing, conducting investigations into digital editing and text encoding. Traditionally, research questions have been dealt with in an interdisciplinary manner, as the staff of the department is made up of scholars with a varied background in social sciences and humanities. As the department's name suggests, its primary mission is to conduct empirically based research into human language, in particular written language. Many of the department's projects are driven by lexicographic and terminological interests.

2. Print dictionaries

Over the past decades, the department has been involved in longstanding lexicographic projects which – in the recent past – have transformed into smaller, more diversified projects. Products of the earlier endeavours have been, for the most part, print dictionaries. The two "*Fackel* dictionaries"[1] came into existence as results of a large-scale text lexicographic experiment which was, in part, carried out in cooperation with the Academy's Commission for the Publication of a *Fackel*-Dictionary.

Another dictionary product to which the ICLTT contributed NLP and corpus technology is the hitherto largest German-Russian dictionary, which was published as a cooperative project of the Austrian and the Russian Academies of Sciences.[2]

3. Computational lexicography

The following paragraphs are meant to give a short overview of our lexicographic endeavours and to showcase some of the experiments in eLexicography that are currently being undertaken at the ICLTT.

3.1 Digitised historical dictionaries

Apart from the traditional print dictionary line described above, the department's activities in eLexicography derive from a second source: smaller historical dictionaries, which are part of the Austrian Academy Corpus[3], a digital German language corpus comprising of approximately half a billion tokens. The texts contained in this corpus were collected with a both literary and a lexicographic perspective in mind. The corpus also contains a considerable number of functional and informational texts. Roughly half of the data is made up of periodicals, not large-size daily newspapers but rather medium- and small-size weekly and monthly publications. There are many collective publications such as yearbooks, readers, commemorative volumes, almanacs, and anthologies covering a wide range of writers, topics, types of texts, and genres. While at a first glance the collection might appear heterogeneous, it actually represents a unique collection of historical German texts, many of which cannot be found elsewhere in digital form.

[1] W. Welzig (ed.): *Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift »Die Fackel«* (Austrian Academy of Sciences Press, Vienna 1998) and *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift »Die Fackel«* (Ibid. 2008).

[2] D. O. Dobrovol'skij (ed.) Neues Deutsch-Russisches Grosswörterbuch. Moskau 2008-2010.

[3] Henceforth, AAC is used in the sense of corpus, not as the institutional name of the predecessor of the ICLTT.

As the corpus was built on a wide concept of literature, which principally included anything reduced to written form, a few monolingual and bilingual dictionaries were also incorporated in the collection. These constituted another starting point for our digital dictionary ambitions.

3.2 Digitally created lexicographic data

Furthermore, the ICLTT also holds some data that came into existence in the digital medium. These were the result of experiments with dictionary creation and dictionary enhancement through automatic and semi-automated procedures. All of the dictionary activities have been closely intertwined with the department's numerous corpus activities.

So far, the largest amount of manually created digital dictionary data produced at the department stem from a project investigating contemporary Arabic varieties. In this project, the *Viennese Corpus of Arabic Varieties*, four small dictionaries have been created so far and a number of others are to follow. These electronic dictionaries are meant to serve a twofold purpose: first, they will furnish the basis of comparatistic dialectological research and will be used to set up a specialised interface (in the department's nomenclature this kind of software component is called a resource viewer) to visualise particular salient linguistic features across a number of linguistic varieties. Second, these dictionaries will also be put to didactic use in language teaching courses at the university. To our knowledge, there are no other machine readable dictionaries of these linguistic varieties so far.

Another project aims at the creation of a machine readable dictionary of Early Modern German (EMG). This is an undertaking that is being carried out on the basis of a small corpus, which has been compiled and annotated at the department. The texts–all of them of Austrian provenance–were automatically tagged with POS and lemmas. In a second step, this data was manually verified. The list of lemmas which were enriched with automatically extracted corpus data will serve as the basis of a small machine-readable dictionary of Austrian EMG. It is planned to complement this dataset with data from other available corpora of the same period in a second step in order to obtain a larger basis for studies of historical variational linguistics.

While the amount of currently available data at the ICLTT is not very large, the number of entries in our dictionary database keeps growing. Yet, it goes without saying that projects such as those described before need a great deal of cooperation and cannot be carried out by individuals. For this reason, we are increasingly focusing on strengthening ties with other interested institutions and intensifying our efforts towards setting up infrastructures that allow collaborative working on dictionaries.

3.3 TEI for dictionaries

In digital text encoding, the guidelines of the Text Encoding Initiative[4] (TEI) have long been considered the de-facto standard and are widely used. While the digital collections of the Austrian Academy Corpus were encoded in a system that might at best be described as TEI inspired, the ICLTT has started to convert all its holdings to TEI P5 as of this year. This also includes the few historical dictionaries contained in this corpus of written German.

When looking for an encoding standard for machine readable dictionaries, LMF (Lexical Markup Framework, ISO 24613:2008) is probably the first thing one might think of. While using the TEI dictionary module to encode digitised print dictionaries has become a fairly uncontested and very common standard procedure, using the very same system for NLP purposes is quite another story. The ICLTT has attempted to make use of TEI's dictionary module to create machine readable dictionaries by imposing a number of constraints on the comparatively flexible structure of the original specifications[5].

The adaptability of digitised dictionary data to TEI P5 is currently being tested in several smaller and larger projects. In one of these experiments, we are converting the above mentioned German-Russian dictionaryinto a machinereadable TEI conformant version.

Another lexicographic experiment is being conducted on the German language version of the collaborative dictionary project Wiktionary. The scarcity of freely available digital multilingual lexical data makes such resources a valuable treasure trove for computational linguists and lexicographers to experiment with. Regrettably, the content of this steadily growing resource, which is not being produced by professional lexicographers but enthusiastic volunteers, is formatted with a lightweight markup system used in different Wiki applications. It is neither standardised nor very structure-oriented. Attempts at preparing Wiktionary for use in NLP applications have been made before, but we have created a freely available tool furnished with a graphical user interface–the first such application to our knowledge–that converts the Wiktionary database dump into a technically reusable XML format, i.e. TEI P5.

4. Digital corpora and dictionary writing

While digital corpora keep growing, they are playing an increasingly important role in modern linguistics. Although representatives of many fields of these

[4] The current version of the guidelines is usually cited with the suffix P5 (i.e. proposal number 5) and can be accessed at http://www.tei-c.org/Guidelines/P5/

[5] We presented a paper entitled *Creating lexical resources in TEI P5. Experiences from building multi-purpose digital dictionaries* at the TEI Members' Meeting 2011 in Würzburg (Germany).

disciplines have remained reluctant to use digital corpora as basis for their investigations, making use of native speaker intuition has become outmoded in contemporary lexicography. Quite to the contrary, lexicographers have long since adopted modern corpus technology; and lexicography has become something like a prototypical field of application for digital corpora.

When creating dictionaries, lexicographers often rely on one particular corpus, very often a collection that has been set up for the particular purpose. The many relevant issues regarding the type of corpus needed for a certain kind of dictionary, the appropriate size of corpus for a particular dictionary, and the corpus features required for a particular lexicographic project will not be touched upon here. It is, however, important to stress the fact that dictionary makers of the 21st century rely heavily on corpus data to build and improve their products. To achieve this end, they need software that allows them to access digital corpora while working on their projects.

When creating dictionaries, lexicographers often rely on one particular corpus, very often a collection that has been set up for the particular purpose. Unfortunately, creating high quality corpora is still costly and time-consuming. It is therefore the most natural thing for those working in the field to look for existing available resources instead. The number of usable corpora has increased considerably over the past years. However, there are still a number of issues that need to be resolved. First of all, freely available does not necessarily imply being ready to be used for lexicographic projects; accessing such corpora often involves some troublesome procedures. Furthermore, federated search in more than one corpus is usually not feasible in optimised dictionary creating workflows.

5. Viennese Lexicographic Editor (VLE)

There are a number of well-established dictionary editing applications. Some of the best-known products include:

- ABBYY Lingvo Content[6],
- DEBII,[7]
- IDM DPS[8],
- Shoebox and the Field Linguist's Toolbox[9],
- iLex[10],
- Lexique Pro[11],
- LEXUS[12],
- TshwaneLex[13]

This list is by no means meant to be exhaustive. Actually, one could make it much longer. Some of these products provide a wide range of functionalities which can be put to use for collecting, refining and enhancing lexicographic data. Some packages are fully integrated systems, others are built in a modular way. Some are being used for particular purposes such as preserving endangered languages, some offer specialised multi-media support. Technically, dictionary writing software is often built around RDBM systems, and often makes use of client-server or multi-tier architecture.

The module presented in this paper is part of a fairly new piece of software that first came into existence as a by-product of an entirely different development activity: the creation of an interactive online learning system for university students. It was first used in a collaborative glossary editing project carried out as part of university language courses at the *University of Vienna*. As the tool proved to be remarkably flexible and adaptable, it was put to work in other projects and is now being used in this research effort designed to fathom out the potential of a more direct integration of corpus data in the dictionary creating process.

At the heart of our dictionary writing system is a dictionary writing client, a standalone application[14], which for the time being has been dubbed – in default of a more adequate name – *Viennese Lexicographic Editor* (VLE). Over the past few months, the client and the associated server scripts have been continually adapted and improved. The whole system relies heavily on XML and cognate technologies such as XSLT and XPath.

5.1 Architecture

The module discussed in this paper is integrated into the above mentioned VLE. The current version only supports web-based editing; the dictionary entries are stored on a web-server. All additional software components (PHP and MySQL) are open source and freely available. On many operating systems, in order to setup the dictionary, simply copying four PHP scripts will suffice to get a working installation of the dictionary server. PHP and MySQL are usually part of the basic installation of such systems.[15] Communication between the dictionary client and the server has been implemented as a RESTful web service.

The distributed architecture has a number of obvious advantages. Being able to work on the data wherever one has access to the internet is unquestionably a useful feature. Lexicographers can then work when they are on vacation without having to carry all data around with

[6] http://www.abbyy.com/lingvo_content/

[7] v. Horak 2006

[8] http://www.idm.fr/products/

[9] http://www.sil.org/computing/shoebox/mdf.html, http://www.sil.org/computing/toolbox

[10] v. Erlandsen 2004

[11] http://www.lexiquepro.com/

[12] v. Ringersma 2007

[13] v. Joffe 2004 and http://tshwanedje.com/tshwanelex/

[14] The software was written in Delphi 2010. Writing software in Pascal dialects is part of a long-standing tradition in Humanities computing. Over the past two decades, many programs have been written at our department making use of this high-level programming language. The large number of reusable libraries allowed us to keep programming overhead to a minimum.

[15] Currently, our main dictionary server is running on *openSuse* 11.3.

themselves. In addition, this system also allows for collaborative working on the dictionary data.

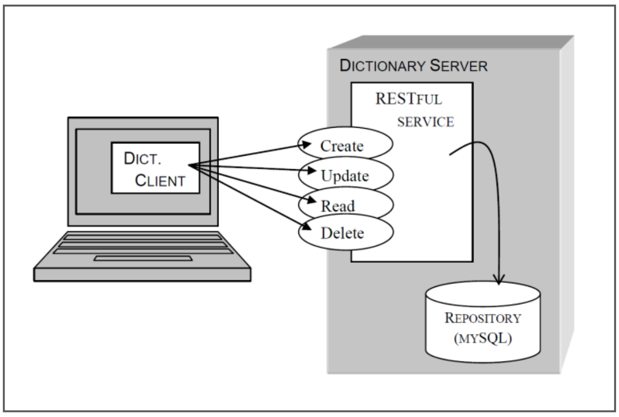


Figure 1: System architecture

5.2 Input validation

One of the main reasons for working with XML is the possibility of ensuring the formal correctness of input. VLE offers the usual two levels of input control: well-formedness, which can be described as the basic compliance of an XML document or data snippet with the syntax of the XML recommendation. When validating a document, the data is checked against a so-called document type definition.

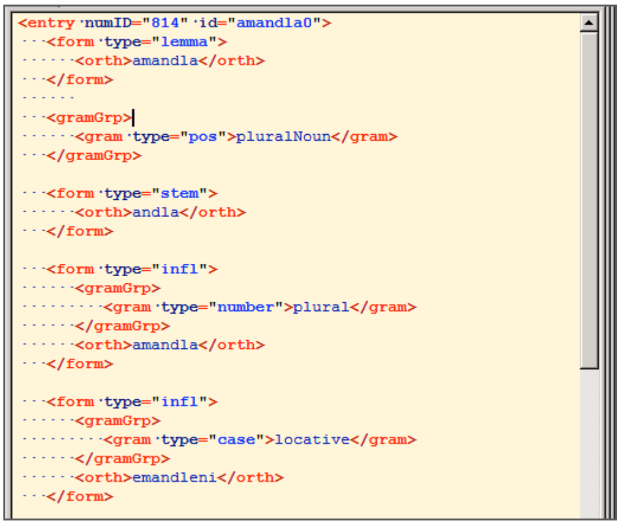
The well-formedness of data in an entry is verified by the VLE tool every time the dictionary entry is saved. Users can also trigger the process manually or make the program check this status with every modification of the entry.

Validation is the process of matching the data on a higher level. When validating the structure of a document, its contents are checked against another document which contains definitions of permissible elements and information as to where these elements may appear in the document. Currently, our tool expects document type definitions in form of an XML Schema which is, like XML, a W3C recommendation. On the to-do-list of the dictionary tool, there is also the implementation of an option to validate against RELAX NG, which is an ISO standard and has found much support in the TEI and OpenDocument communities.

5.3 Editor modes

The user of VLE has two basic options for editing the dictionary data: working in XML mode, which may be considered the expert mode, or working in an editor form with predefined entry controls that function like traditional database input fields. While working on an entry, it is possible to switch between the two modes at random. The second option, i.e. making use of edit controls for particular XML elements, is especially useful when working on the same field across a number of dictionary entries. Navigating in the XML expert mode is more cumbersome than in the edit controls mode, since lexicographers have to position themselves in every entry they work on.

When working on large dictionary entries, keeping track of the entry details or even just the current position within an article can, at times, be a troublesome undertaking. Actually, XML encoded data is of great advantage in this respect, as the structure of the entries can help the software in tackling these problems. In particular, the TEI system with its intuitive and not too verbose element names eases the task. The software allows lexicographers to navigate to particular cognate points within the text.



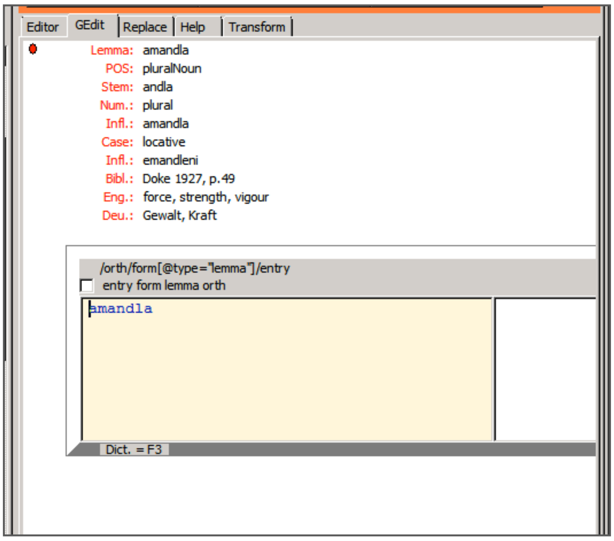
```
<entry·numID="814"·id="amandla0">
···<form·type="lemma">
······<orth>amandla</orth>
···</form>
······
···<gramGrp>
······<gram·type="pos">pluralNoun</gram>
···</gramGrp>

···<form·type="stem">
······<orth>andla</orth>
···</form>

···<form·type="infl">
······<gramGrp>
·········<gram·type="number">plural</gram>
······</gramGrp>
······<orth>amandla</orth>
···</form>

···<form·type="infl">
······<gramGrp>
·········<gram·type="case">locative</gram>
······</gramGrp>
······<orth>emandleni</orth>
···</form>
```

Figure 2: XML mode



Lemma: amandla
POS: pluralNoun
Stem: andla
Num.: plural
Infl.: amandla
Case: locative
Infl.: emandleni
Bibl.: Doke 1927, p.49
Eng.: force, strength, vigour
Deu.: Gewalt, Kraft

/orth/form[@type="lemma"]/entry
□ entry form lemma orth

amandla

Dict. = F3

Figure 3: Database-like input mode

5.4 Visualisation of entries

An additional useful feature of the tool is its capability to visualise the data. This task is achieved by means of XSLT stylesheets which are freely configurable. While this functionality is quite commonplace in many applications today, our tool proves to be particularly versatile. Making use of different styles, allows switching between different views of the same set of data. When working on very large entries, stylesheet transformations tend to be restrictively slow. This is not the case, however, when they are only applied to particular parts of an entry.

Automatically applied links in the output data (HTML) allow navigation from these visualisations back into the editor control, which again makes navigating copious dictionary entries a considerably more agreeable task.

5.5 Data export

VLE stores all data on a server. In addition, it has also been provided with the functionality to store output on the local machine. Making use of the export control, all data can be saved into one document. Usually, all metadata and production related data such as the configuration profile are inserted in these documents.

5.6 Web-publishing

Dictionary entries created by this tool can be published on the internet through a simple PHP script. Adapting an HTML template to create a new dictionary web-site is a matter of minutes. The resulting web-page has a query control and is able to display the results of dictionary queries.

6. From *corpus* to *dictionary entry*

When digital corpora are used to compile new dictionaries or to enhance existing ones, more often than not interfacing between the dictionary writing software and the respective corpus poses considerable problems. When accessing corpora in tandem with producing dictionary entries, the issue at hand is transferring the results of corpus queries in an acceptably comfortable manner. Very often this process involves rather cumbersome steps that require a series of manual manipulations. The focus in this project has been on streamlining this process, on speeding up the import of corpus data into dictionary entries.

When accessing digital corpora, lexicographers might be interested in a broad spectrum of data including, but not limited to:

- lists of collocations,
- multiword units,
- statistical information on lemmas, particular word forms or any of the afore mentioned categories,
- corpus examples.

Although this particular module of the dictionary writing editor can perform many other tasks, this paper focuses on the issue of corpus examples. The principal idea when preparing this module was optimizing access to digital corpora in order to allow lexicographers to glean sample sentences and to integrate them into dictionary entries in a reasonably comfortable manner. The focus of our work was on ease of use and direct access to the data. The corpus interface of the new dictionary writing application presented in this paper was supposed to enable lexicographers to launch corpus queries and to offer functionalities for inserting them into existing dictionary entries without needing to use the clipboard to copy-and-paste, which inevitably results in a lot of inefficient typing or clicking.

The new corpus browser module was designed to be a principally universal web-interface and to allow lexicographers to query not just one particular corpus, but any digital resources accessible via a web-browser. One of the important perspectives of the new corpus browser was its integration with evolving CLARIN[16] infrastructures, in particular *federated content search* facilities, a project which was initiated by a CLARIN working group this year. Researchers of the ICLTT have taken a keen interest in these activities and have actively contributed to this ongoing project.

For the purposes of our research, the VLE's corpus interface needed to enable lexicographers to copy the selected data into the dictionary writing editor using a single click or keystroke. In addition, the scripting of processes had to be possible and the transfer of data needed to be achieved through a transformer component that could automatically perform predefined modifications of the text and translate the HTML text received by the browser into the target formats required by the dictionary system.

The process of enriching dictionary data with data from digital corpora as performed by our new tool can be described as a workflow made up ofsix basic steps:

- Querying a corpus / corpora
- Optional pre-selecting of data in the browser
- Analysing the data
- Selecting data from a list of candidates
- Converting to the target format of the dictionary
- Inserting the data into an entry

6.1 Querying corpora via the Internet

VLE's integrated web browser allows lexicographers to access and search the internet. It works very much like other such tools, but it does not have some of the extra features such as bookmark management, download management, or a search-engine toolbar, which are unnecessary for our purposes. The component used to realise this part of the programme is a common wrapper, which was placed around Windows' native Internet Explorer component.[17]

[16] CLARIN stands for *Common Language Resources and Technology Infrastructure* (http://www.clarin.eu/external/) and was initiated as an ESFRI project.

[17] *TWebBrowser* is a visual component that allows programmers to create simple web-browser applications in just a few minutes.

The current version of the module can work with any corpus that delivers data through the HTTP protocol. It can not only work with text collections structured as corpora but with any data delivered as HTML or raw text. It cannot deal with PDF documents at the moment, however. When the module's browser receives XML data, they are transformed into HTML using XSLT style sheets. An additional interesting feature of the tool, albeit of lesser importance for the particular purpose being discussed here, is the capability to access other online available dictionaries.

Ideally, lexicographers should be able to trigger queries directly from the entry edit control, the part of the tool where the dictionary text is edited by the lexicographer. Unfortunately, this direct approach is not an option with many corpora or text collections available on the internet, as they offer access exclusively through their own web-interfaces. This means, that users have to navigate to the respective corpus entry points first and then input their queries manually.

6.2 Pre-selecting data

When search results have been retrieved from a corpus and appear in the browser, these data have to be dealt with in some way or another before they can be integrated into a dictionary entry. With our tool, users have two options at this stage: they can either accept the received data in their entirety or they can manually select only part of it.

6.3 Analysing the data

Manual intervention is practically unnecessary after the data have been pre-selected, as the program has the ability to perform a first analysis of the imported data. Usually, the results of corpus queries are delivered as concordance lines, typically in form of Key-Word-in-Context (KWIC) lines, which, when sent over the internet, are commonly transformed into HTML tables. These structures can easily be identified by the software of our tool.

6.4 Selecting data from a list of candidates

Having performed this initial analysis of data, the program passes it on to the selector control, which presents the data to the user in a listbox control. Here, the user can make the final selection for the dictionary entry. This is the only point in the process where manual intervention on part of the lexicographer is inevitable.

6.5 Converting data

After the selection of data, data snippets are passed into the entry editor through a template, in which the exact XML structure of the data to be inserted can be defined. In addition, the tool has also the capability of carrying out data conversions through a service based mechanism. This mechanism allows actions to be performed in a distributed manner. This might make it possible in the future to enrich parts of the data being worked on through services offered elsewhere on the internet.

7. Corpus access

The steps described above can each be performed separately. However, the point of systematically defining this workflow was to allow lexicographers to automate as many of the intermediary steps as possible, thus avoiding any redundant key strokes or clicks in order to stave off carpal tunnel syndrome as long as possible and to make work on external sources more efficient.

The most critical step in the importing process is the query. Circumventing the step "pre-selecting of data" as described in 6.2 is only possible through direct access to the corpora one wants to query. The user (with the help of the software) must be able to launch queries directly via the HTTP protocol. Relieving the user of manually initiating the communication with the remote corpus can only be achieved through a service that allows machine-to-machine communication. The establishment of service based access points for corpora is a fundamental prerequisite for the smooth integration of dictionary client and corpus.
VLE is capable of performing the above described process sufficiently well when accessing our own data servers at the Academy. It allows lexicographers to launch queries directly from the editor control, simply by selecting a string and triggering a function. The problem arises on the other side of the communication, as most other corpora do not offer service based interfaces that allow outside software to interact with them directly. Web-interfaces of corpora are usually geared towards the needs of human users. As a result, queries can only be triggered when text is manually entered into edit controls in web forms.

8. LRT standards

Many activities of the ICLTT have been characterised by a strong commitment to standards and de-facto standards. This awareness of the relevance of standards has been largely motivated by the department's involvement in interdisciplinary projects which involved heterogeneous resources and a wide range of methodologies and tools. The need for harmonising divergent environments has heightened our awareness for issues of interoperability, reusability and LRT standards. This also accounts for the extensive use of XML, Unicode and related technologies in all applications. The AAC's first XML encoded digital objects – a digital version of the Austrian historical magazine "Die Fackel" (6 million tokens) – date back as far as 1998, the year in which the World-Wide-Web-Consortium passed its first XML recommendation.[18]

[18] At that time, our work was based on Extensible Markup Language (XML) 1.2 (http://www.w3.org/TR/1998/REC-xml-19980210).

Our current experiments with the dictionary writing software have been conducted using a combination of TEI's dictionary module (P5) and ISOCat[19]. Other standards relied on in these projects include MAF (Morphosyntactic Annotation Framework, ISO/DIS 24611) and ISO 639 (Language codes).

Since the architecture of the dictionary writing system is built on XML, schemata other than TEI can also be implemented with ease. In addition to systems like OLIF (Open XML Language Data Standard), one might consider using formats such as OWL (Web Ontology Language), RDF (Resource Description Framework), TBX (TermBase eXchange) or LMF (Lexcial Markup Framework, ISO 24613)[20] in future projects.

Actually, the first experiments with VLE were undertaken on the basis of LMF encoded data, as our endeavours have been directed towards creating machine readable dictionaries. While we have not discarded the use of LMF for future projects, the less verbose and for human lexicographers more easily readable structure of TEI (P5) has so far tipped the balance in favour of using this encoding system in our projects. However, LMF continues to plays an important role; in contrast to TEI (P5), it is a full-fledged ISO standard. What has remained of the early LMF experiments is a function in the VLE's entry editor control capable of converting the ICLTT's TEI dictionary entries into LMF entries. However, this TEI to LMF converter is not a universally applicable tool, as it only works with the ICLTT's TEI dictionary format. In the future, this part of the editor might be extended as LMF, probably, will gain more importance.

9. Current status and availability

As we have shown above, our dictionary writing system is made up of easily distributable, easy to set up components: All that is needed is a client, a server (in our case Apache) with a mySql database and four PHP scripts to run the RESTful service. The system has been optimised for ease of use and interoperability, everything is based on XML.

The system has been intended for use by individual lexicographers and small groups of researchers. Currently, it is being tested in several small to medium sized projects and numerous amendments are constantly being applied. We have started to work on a small user guide and there are plans to make a first version of the client available for interested researchers in the course of 2012. The package is still in beta stage, and no decision has been made yet as to the license under which the client software will be available, but the four server scripts can be downloaded from the ICLTT Showcase website (http://corpus3.aac.ac.at/showcase).

10. Conclusions

In view of all the dictionary software that already exists, one could rightfully ask: why produce yet another tool? To answer this question, one has to consider the fact that software lifecycles have shortened considerably in recent years. Reusable components and libraries allow new products to be created with comparatively small overhead. In addition, there are hardly any state-of-the-art applications that are open source, extensible, comfortably manageable by non-technicians and free of charge. With the creation of the *Viennese Lexicographic Editor,* such reusable components have been combined with a state-of-the-art application that can potentially solve some of the problems of streamlining workflows at the interface between corpus and dictionary writing systems.

There are several simple answers to the question above: because it was possible to do it, because it did not cost much and because it might motivate others to muster courage to go ahead with their own lexicographic ambitions. Researchers are often wary of going digital, individual researchers are particularly confronted with problems which could be remedied to a certain degree by more easily attainable and usable software. As the national coordinator of the two projects CLARIN-AT and DARIAH-AT, the ICLTT sees its role also as a facilitator to enable more researchers and scholars in the Humanities and the Arts to take the digital path.

In conducting these experiments, we have been guided by a vision of a densely knit web of dictionaries, where datasets created by human editors are enhanced by automatically created data, where lexical resources created by automatic routines serve as the basis of an ever renewed and growing lexicographic web. In building this new application, we are envisaging more reusable, standards-based and ideally open-source LRTs being developed by ever growing communities of both individual and groups of researchers.

11. References

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

Baumann, S., Burnard, L. & Sperberg-McQueen, C.M. (eds.) (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford - Providence – Charlottesville - Nancy. Accessed at: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Breiteneder, E. (2003). *Austrian Academy Corpus.* In Thomas Städtler (ed.) *Wissenschaftliche Lexikographie im deutschsprachigen Raum*.

[19] A web-application, offering access to ISO TC37's Data Category Registry of widely accepted linguistic concepts.

[20] It is worth mentioning here that we also use the dictionary writing system to manage other types of data such as bibliographies, prosopographic databases and feature catalogues for a project involving comparatistic studies in linguistic varieties. All of this data is TEI conformant XML.

Heidelberg, Universitätsverlag Winter, pp. 447-448.

Breiteneder, E. (2003). *Wörterbuch der Fackel.* In T. Städtler (ed.) *Wissenschaftliche Lexikographie im deutschsprachigen Raum*. Heidelberg, Universitätsverlag Winter**,** pp. 295-297.

Erjavec, T., Evans, R., Ide, N., & Kilgarriff, A. (2003). From machine readable dictionaries to lexical databases: the Concede Experience. In *Proceedings of the 7th International Conference on Computational Lexicography. COMPLEX'03*. Budapest.

Erjavec, T., Tufiş, D., & Varadi, T. (1999). Developing TEI-conformant lexical databases for CEE languages. In *Proceedings of the 4th International Conference on Computational Lexicography. COMPLEX'99.* Pecs, Hungary.

Erlandsen, J. (2004). iLEX – an ergonomic and powerful tool combining, effective and flexible editing with easy and fast search and retrieval. In *EURALEX 2004, Lorient, France*.

Horák, A., Pala, K., Rambousek, A. & Rychlý, P. (2006). New clients for dictionary writing on the DEB platform. In *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems.* Torino, Italy: Lexical Computing Ltd., U.K., pp. 17-23.

Joffe, D., de Schryver, G.M. (2004). TshwaneLex – professional off-the-shelf lexicography software. In *Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts,* Brno, Czech Republic, Masaryk University, Faculty of Informatics.

Kilgarriff, A., Kovár, V. & Rychlý, P. (2009). Tickbox Lexicography. In S. Granger, M. Paquot (eds*.) eLexicography in the 21st century: New Challenges, New Applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 411-418.

O'Keeffe, A., McCarthy, M. (eds.) (2010). *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge.

Ringersma, J., Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme, R. van Son (eds.) *Proceedings of Interspeech 2007*. Baixas, France: ISCA-Int. Speech Communication Assoc., pp. 65-68.

Spohr, D. (2008). Requirements for the design of electronic dictionaries and a proposal for their formalisation. In *Proceedings of the EURALEX International Congress 2008*. Barcelona.

Walter, E. (2010). Using corpora to write dictionaries. In A. O'Keeffe, M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics.* London and New York: Routledge, pp. 428-443.

It's Not Impossible: Bringing Derived Words Out of the Shadows in an Electronic Dictionary

Jean V. Callahan
Wordsmyth
Ithaca, NY
E-mail: jean@loudeacstudio.com

Abstract

The paper proposes that expanded and improved treatment of morphological information is both needed and newly possible in electronic English dictionaries, including dictionaries for native speakers. The focus is dictionary selection, treatment, and presentation of derived words. Inconsistencies and inadequacies can be attributed not only to print's legacy, but also to assumptions about native speakers' automatic acquisition and application of word formation rules that are challenged in this paper, and to the more philosophical problem of how the lexicalized word relates to the morphologists' word. Word Formation Rules developed in linguistics and the properties of electronic media enable lexicographers to produce and display new entry information and new navigational pathways through dictionary data.

Keywords: English dictionaries; morphology; derivation; electronic dictionaries

1. Introduction

Is it a word? This is one of the questions for which a dictionary can provide answers. In many cases, however, a dictionary search will not return an answer. We can all think of numerous reasons why that would be, some of which might be overridden by online aggregation or integration of dictionaries, some of which would not. The present study focuses on one reason in particular, and that is the inadequacy of morphological, or word formation, information in the typical English monolingual dictionary. For "inadequacy," one might substitute the word "latency," for it has been estimated that in a 100,000-word dictionary, 80% of the words are complex words formed from other words by means of derivation. Morphologists use dictionary data in formulating theories and rules of word formation. Yet the only explicit dictionary treatment of words *as* derived words occurs in run-ons of a headword entry, where a limited type of derived words are listed, minimally with part of speech, sometimes with syllabification and pronunciation, and rarely, as in *Collins Cobuild*, an example sentence.[1] Users who look up *latent* may or may not find the noun *latency* listed in that entry, and, if they do find *latency*, there will necessarily be no headword entry for it. A search on the word *squashable* will yield no results in any dictionary, nor will *crushworthy*, but *crushproof*, *crushable*, and *crusher* are listed under *crush* in *AH4* and a search on them will turn up that entry. Morphological information, allied with electronic media and computation, can be used both to guide users in the deciphering and production of words that do not appear in a given dictionary or, perhaps, any dictionary, and to lay new navigational pathways by which users can more easily find relevant existing words in a dictionary.

Now many English dictionaries do include information about word formation in headword entries for affixes and combining forms. *Merriam-Webster Online*'s (2011) explanatory notes remark that such entries "make understandable the meaning of many undefined run-ons which for reasons of space would be omitted if they had to be given etymologies and definitions; and to make recognizable the meaningful elements of new words that are not well enough established in the language to warrant dictionary entry."

The notion of using affix entries to make undefined run-ons understandable is not entirely consistent with what explanatory notes tell the user about run-ons themselves. Of its selection of run-on derived words, *Webster's New World Dictionary* says: "We included words one might reasonably expect to encounter in literature or ordinary usage, and then only when the meaning of such derived words can be immediately understood from the meanings of the base word and the affix." (1982:xiv) The meaning of derived word run-ons has been variously described by dictionaries as "self-explanatory," "immediately understood," and "readily derivable," but the user may need to compose the meaning herself by consulting an affix entry.

2. Morphological Information in Dictionaries

Along with the two locations most often referenced in discussions of how dictionaries present morphological information about derived words, 1) run-on derived words and 2) entries for affixes, a third location can be added which broadens the scope of the problem and the solution, namely, 3) headword entries for lexicalized derived words: the 80%.

[1] Notable exceptions: some online dictionary sites, notably wordnik.com and thefreedictionary.com, return examples from a corpus of the word form entered, even if the word is a derived word sub-entry. Wordnik search, further, returns examples for words that have no representation in its dictionary sources. Vocabulary.com displays derived word data visually, with frequency information, upon look-up of a member of that word family.

2.1 Derived Word Run-ons

Because lexicographers have followed the principle that derived word run-ons have self-explanatory or readily derivable meaning and do not need definitions, the best, most useful derivations of a base often are not listed with their base because they are headword entries. They may be nearby alphabetically in English but are not purposefully so. Being able to see words around or near the word a user has looked up, while it has an important role in the translation of print dictionaries into electronic dictionaries, serves as poor compensation for the lack of structured navigational pathways connecting words that are closely related structurally. Thus under the adjectives *mediate*, *ferocious*, and *atrocious*, we find the derived noun run-ons *mediateness*, *ferociousness*, and *atrociousness*, but not *mediacy*, *ferocity*, and *atrocity*, let alone the prefixed derived adjectives *immediate* or *unmediated*.

word	**headword**	**run-on**	**occurrence COCA**
mediation	3/4	1/4	2355
mediator	2/4	2/4	1298
mediate	4/4	0/4	1139
mediational	0/4	1/4	67
mediately	0/4	3/4	16
mediacy	1/4 (AH4)	1/4 (MW)	8
mediatorial	0/4	1/4	1
mediateness	0/4	1/4	0
mediatorially	0/4	1/4	0

Table 1: status of derived words

Inconsistency in dictionaries' treatment of derived forms has of course been noticed before and often. In 1993, Bauer and Nation stated that lexicographers are not "deciding in a principled and consistent way on what derived forms to include as full entries, defined sub-entries, and non-defined sub-entries, and what forms not to include" (Bauer & Nation, 1993:255). Table 1 offers just a glimpse of this all-too-familiar inconsistency. The four dictionaries considered are *Merriam-Webster Online*, *American Heritage 4, Collins English Dictionary*, and *Longmans Dictionary of Contemporary English*. More recently, in 2010, *AH4*'s inconsistent treatment of derived words is noted with reference to the real estate given to headwords *retrench* and *retrenchment,* despite the latter's transparency, while much more frequent derived words under retaliate have run-on status (Delahunty & Garvey, 2010:240-1). Jackson and Amvela write: "The morphological aspects of lexical description are not systematically covered by dictionaries. Where morphemic relationships are indicated, they are evident more from the alphabetical ordering and nesting practices of dictionaries, rather than from any consciously explicit treatment." (Jackson & Amvela, 2000:168) Although more general-purpose dictionaries now include headword entries for affixes and combining forms, the "analytical work" is still left to the user (Jackson & Amvela, 2000:168).

2.2 Lexicalized Derived Words

Lexicalized derived words, such as *mediation*, are headwords with full entries, often with their own run-on derived words but no structured reference back to their base, thus completing the absence of explicit relationship that begins in the undefined run-ons. However, although a structured relation (metadata) is missing, morphology nevertheless often creeps back in via the definitions, when the base is used to define the word. Sometimes the definition of a derived headword is almost entirely morphosemantic and,. from the point of view of the user, frustratingly circular. As seen earlier in Table 1, *mediacy* appears in *Merriam-Webster* as a derived word under *mediate*, but as a headword in *American Heritage*:

mediacy
the state or quality of being mediate. (AH4)

Table 2: morphosemantic definition

The success with which dictionary definitions handle or juggle the distinction between morphosemantic information on the one hand and word sense on the other is uneven, and even when approached according to a set of coherent editorial guidelines, can be confusing for users (and lexicographers).[2] For native speaker and learners dictionaries, avoiding circularity in definitions is a valid consideration, indeed of utmost importance; however, defining a derived word without any reference to the base word means we lose morphological information and prevent morphological awareness.

We can attribute a significant portion of conventions related to derived words to the impressive legacy of print dictionaries of English. Still, it is important to examine the message in the medium, as well as the medium's incontrovertible material disadvantages (and advantages). Print's space, storage, and graphical display limitations are paralleled in a principle of non-redundancy that is not entirely materially-based. The same principle of non-redundancy that in part prohibits defining run-on derived words and also prohibits the appearance of a lexicalized derived headword as (also) a run-on under its base (*atrocity* under *atrocious*) loses its hold in the writing of definitions themselves. In addition to the problem of unsystematic selection of derived words and unrationalized delegation of them to their proper slot, dictionaries often do not manage to adequately distinguish morphosemantics (of the latent kind) from lexical meaning. This "latent" representation of morphology in dictionaries creates interference in the communication of lexical information**.** An approach to a possible solution might lie in creating a fuller, more explicit, more independent and self-contained layer of morphological data that would be integrated with but distinct from dictionary data.

[2] For a vivid, amusing documentation of the difficulty, see Gove's advice to the lexicographers working on *NID3* with regard to entering "self-explanatory words" (1966).

2.3 The Morphological Word vs. the Lexicon

One reason for the kinds of inconsistency we see in dictionaries' treatment and presentation of derived words is that, as has been already touched on, morphological words are not the kind of words a dictionary is all about. A dictionary provides users with information about the lexicon, the properly attested mappings of words with their arbitrary, idiosyncratic, acquired sense. The lexicon is for the "lawless." (De Sciullo & Williams, 1987:3) These words are the nodes around which information clusters. The goal of morphology, on the other hand, is "the enumeration of the class of possible words of a language" (Aronoff, 1976:17-18). "It is the task of morphology," Aronoff (1976:19) writes, "to tell us what sort of new words a speaker can form."

The morphologist's word is a different creature from the lexicographer's word. The words morphology treats of are words formed by consistent rule-based processes applied to lexical units, not only describable by syntax or syntactic demands. The word formation rules of morphology produce words whose meaning is, must be, compositional and predictable. Scalise and Guevara (2005:62) explain that: "The meaning of a complex word is always compositional when it has been created by a (synchronically) productive WFR. With time, a complex word may acquire unexpected or idiosyncratic meanings, i.e. meanings that cannot be derived from its constituents," and they cite the standard example of the word *transmission*. Thus, words as once and future rule-bound, synchronic formations, and their structural relation to other such words, fall almost by definition outside the lexicon. And yet, as we have seen, morphology is latently and often necessarily present in definitions (as well as in lexicalized headwords), creating interference in the communication of lexical sense.

2.4 Affix entries and Word Formation Rules

Affix entries are as isolated from run-ons and headwords in electronic dictionary displays as they are in print, but it is here that dictionaries provide information about word formation most explicitly.

As Table 3 makes evident, dictionaries can differ significantly in their treatment of affix entries. Dardano et al., considering monolingual Italian dictionaries, propose that "information contained in dictionaries on affixes and combining forms must include not only the meaning of these elements, but also the way in which they form new words." (2006:1117) By these criteria, *Collins English Dictionary* exceeds *The American Heritage College Dictionary* (Table 3) on several points, because the affix is presented as functioning in a process that forms words of one category from words of another category and makes a regular semantic change to the base.

AHCD3	CED
-al^1 *suff.* Of, relating to, or characterized by: *parental.* [ME< OFr. < Lat. *–alis,* adj. suff.]	-al^1 *suffix forming adjectives* of; related to; connected with: *functional sectional tonal* [from Latin *-ālis*]
-al^2 *suff.* Action; process: *retrieval.* [ME -aille<OFr. < Lat. *–alia,* neut. pl. of *alis*.]	-al^2 *suffix forming nouns* the act or process of doing what is indicated by the verb stem: *rebuttal recital renewal* [via Old French *-aille, -ail,* from Latin *-ālia,* neuter plural used as substantive, from *-ālis* -al^1]
-al^3 *suff.* Aldehyde: *citronellal* [<al(DEHYDE)]	-al^3 *suffix forming nouns* **1.** indicating an aldehyde *ethanal* **2.**(Medicine/Pharmacology) indicating a pharmaceutical product *phenobarbital* [shortened from aldehyde]

Table 3: suffix entries in two dictionaries

The briefest look at morphologists' work on Word Formation Rules (WFR), however, suggests how much further dictionaries have to go--or could go, as motivation and resources allow. For the most ambitious undertakings, lexicography will need to take advantage of the work morphologists have done on WFR, which, with the advent of electronic dictionaries and other lexical databases, has gained new potential for practical applications. [3]

"Word Formation Rule" is in a sense the morphologists term for affix. A WFR involves process and uncovers and explicates regularity at a fine level. A WFR for

[3] Here it should be noted that significant projects based in Europe are under way which provide work or user interfaces that access morphological databases: MuLexFor (Cartoni & Lefer, 2010), elexico (Klosa et al., 2006; Storjohann, 2005), and Word Manager (Domenig & ten Hacken, 1992). Canoo.com uses Word Manager to develop software products, one category of which is "Unknown Word Tools," which can "*analyze* unknown (i.e. not lexicalized) words based on word formation rules" and "*recognize* unknown (i.e. not lexicalized) words based on word formation rules." (Canoo.com).

suffixation is the suffix morpheme itself and rules regarding the input and output. Word Formation Rules encompass:

1. the part of speech of the word a suffix forms (*-ness, -ment, -ion, -al*^{2} form nouns)

2. the features of the base it "selects": part of speech (*-ness* selects adjectives, *-ive* selects verbs, *-al*^{2} selects verbs); bound or free morphemes (or both); Latinate or native bases (or both), and even register.

3. the suffixes' position in relation to the base and to other suffixes.

4. phonological and orthographical changes a suffix effects in the base, if any (Table 4), such as stress patterns, pronunciation, and spelling.

5. semantic effects. For example, *-al* selects verbs to forms "abstract nouns denoting an action or the result of an action" (Plag, 2003:109).

6. restrictions on output. For example, semantic restrictions entail that meaning must be compositional (synchronically). A phonological restriction on output prevents, for example, *candidity* and *obsoletity* from being possible words (Plag, 2003:115).

7. productivity and distribution. For example, information on present and historical productivity of affixes. Questions of whether an affix is productive, and how productive, and in which registers and domains, are practically untouched by English dictionaries for native speakers.

suffixes triggering alternation		suffixes not triggering alternation	
-(at)ion	alternation	-ness	religiousness
-y	candidacy	-less	televisionless
-al	environmental	-ful	eventful
-ize	hypothesize	-ship	editorship
-ive	productive	-ly	headmasterly
-ese	Japanese	-ish	introvertish

Table 4: Phonological effects of suffixes (Plag, 2003:101)

Even a small sampling of the information about affixes uncovered by Word Formation Rules will hint at their untapped potential for dictionary development and use. For example, of words formed with *-ity*, "Words belonging to this morphological category are nouns denoting qualities, states or properties usually derived from Latinate adjectives (e.g. *curiosity, productivity, profundity, solidity*)" (Plag, 2003:115). All adjectives suffixed with *-able*, *-al* and *-ic* or ending in the [Id] sound can take *-ity* as a nominalizing suffix (*readability, formality, erraticity, solidity*) (Plag, 2003:115). And of *-ity*'s phonological features: "All words formed with this suffix have their main stress on the antepenultimate syllable" (Plag, 2003:119). Of the suffix *–al*, we learn that it only selects verbs with final stress (*arrival, accrual, reappraisal, overthrowal, recital, referral, renewal, abettal*) (Plag, 2003:76).

One approach to utilizing work on WFR to improve dictionary treatment of morphological information is to re-evaluate and expand information provided on affixes in dictionaries. Prcic, for example, through his examination of the big four monolingual ELL dictionaries, proposes ten categories of information that should be represented in affix entries. Prcic's categories: Spelling, Pronunciation, Input/Output units, Sense distinctions, Definitions, Cross-references, Usage labels, Productivity, Examples and Terminology (Prcic, 1999, cited in Lefer, 2010:1).

Another tack would be to make this expanded information on affixes (and concomitantly bases) available at or from individual headword entries, thus meeting the user at the point where she seeks and obtains information while engaged in a particular use case (de Caluwe, 2011). De Caluwe builds on the writings of ten Hacken and is one of the few sources that explicitly makes a case for the usefulness of word formation information in dictionaries for native speakers: "Providing the user looking for the meaning of a word with information on the paradigmatic, *in casu* morphological relations, of that word with other items in the lexicon really constitutes an added value to the user, on the condition of course that it will not lead to *information stress*" (de Caluwe, 2011).

Online dictionaries have not yet exploited the granularity of data and metadata now possible to allow users to "look inside this word." Using a set of word formation rules compiled with "users and uses" in mind, a rough draft of classifying, tagging, and indexing dictionary data can be automatically generated and then manually edited. Derived words, whether undefined run-ons or headwords, can be graphically marked to convey their word parts with a simple asterisk. Hyperlinked affixes and bases would allow users to click or hover and open a pop up box containing at least affix entries and at most full morphological dictionary entries. The ability to customize display, to show or hide fields, or to choose to see only lexical or only morphological information can be applied to counteract clutter or "information stress."

3. Morphological Awareness in the Curriculum

A particular use case scenario for which there are evidence-based user needs and potential market demand has emerged quite recently from education research, specifically studies in literacy, vocabulary acquisition, and vocabulary teaching in K-12 for native speakers. The case for the importance of word formation information in dictionaries for second-language learners has been well represented. Ten Hacken (2006:243), for example, writes:

"In second language acquisition, word formation is important for the decoding of words the learner does not know, for the production of regular new words when the learner has not acquired the standard word, and for the creation of a tighter network structure in the mental lexicon, which facilitates vocabulary acquisition. "

Research by Bauer and Nation (1993), Nagy et al. (1984; 1989), and others has shown the effectiveness of teaching vocabulary in "word families" rather than as individual words. The vocabulary load of a reading text is reduced significantly when the unit of learning and measurement of text difficulty is word families, a set of words related by derivation, rather than the individual word. According to Nagy and Anderson (1984), "The less aware a student is of word relations, the more distinct words need to be learned." As students progress from learning to phonologically decode and encode in writing high-frequency words to encountering longer, more morphologically complex lower-frequency words, "knowledge of word-formation processes becomes necessary for reading and spelling words" (Nagy & Anderson, 1984:). More than half of the unfamiliar words students encounter in middle school and beyond will be words whose meaning they will be able to deduce from context, if those students are equipped to discern morphological structure (Nagy et al., 1989). These findings were published more than 20 years ago, and their message seems to be edging closer to the threshold of standardization. A 2009 study states that, "To date, national attention in the United States has focused on evidence based practices related to phonological decoding, but not to evidence-based practices related to word formation, which may be critical for fostering literacy achievement in fourth grade and beyond" (Berninger et al., 2010:156). Joanne F. Carlisle (2010:3) writes that "(m)orphological awareness, defined as the ability to reflect on, analyze, and manipulate the morphemic elements in words, can be considered one form of students' developing linguistic awareness. Morphological awareness develops gradually, as students come to understand complex relations of form and meaning." These recent conclusions come out of a psycholinguistically-inflected reading research. As well as representing a welcome pendulum-swing away from the emphasis on the meaning-less decoding of "phonics," they reflect interestingly and critically on assumptions that native speakers acquire and apply word formation rules unconsciously or automatically to decipher and produce unfamiliar derived words.

4. Conclusion

In Atkins and Rundell (2008:48), the lexicographer is said to deal with "the probable, not the possible." This phrase appears when the authors ask how we "cope as lexicographers" with the "individual departure from 'normal' modes of expression" that generate the countless words that do not appear in dictionaries. "As always," they write, the answer will depend to some extent on 'users and uses': the kinds of people the dictionary is designed for and the reference needs which the dictionary aims to cater for. But a good basic principle is that… the job of the dictionary is to describe and explain linguistic *conventions*… Our focus in other words, must be the probable, not the possible." (Atkins & Rundell, 2008:48). Neither every word that has ever been used by individual members of a language community, nor every linguistically legitimate word in a language can or should be included in a dictionary. These fall into the category of "possible or potential words," not "probable words."

Nevertheless, the boundary between word probability and word possibility, where documentation of attested words and their frequency and usages stops and the range and likelihood of possible words begins, is a shifting boundary depending on "users and uses," on the parts of the lexicon and of the language possessed by the intended user, on how well and in what way those parts are possessed, and on the uses intended to be served by a given dictionary.

Native speaker students of English constitute a set of users for whose individual mental lexicons lexicalized words may have the status of merely "possible words." Bauer and Nation's classification of affixes into seven levels of increasing difficulty and complexity acknowledges and provides stepping stones for the mental lexicon to convert possible words into vocabulary (1993).

Even outside the context of educational institutions, dictionary entries can provide answers to questions all kinds of users might ask about words: is it a word? is it a possible word? is there an adjective/verb/noun form of this word? Can I add prefix X or suffix Y to this word? "For a long time," Dardano et al. (2006:125) write, "lexicographers have not acknowledged the importance of explaining the mechanism of Word Formation and educating users to create their own neologisms and apply them to everyday life." There is a range of possible responses to this observation. We might create search capabilities that can recognize user queries as possible or nonpossible derived words and that offer users explanation of relevant Word Formation Rules, and provide data that suggest frequency, domain, and register, from corpus examples of the queried word. Or we might map derivational morphology onto dictionary entries in order to guide users to navigate more effectively among existing headwords related by morphological structure. Both of these directions expand the usefulness of dictionaries and allow the human subject of language to harvest the potential of technology's transformation of dictionaries.

5. References

American Heritage Dictionary of the English Language, 4th ed. (2003). Accessed at http://www.thefreedictionary.com

American Heritage College Dictionary, 3rd ed. (1993). Boston: Houghton Mifflin.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

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

Bauer, L., Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), pp. 253-279.

Bauer, L. (1988). *Introducing Linguistic Morphology.* Edinburgh: Edinburgh University Press.

Berninger, V., Abbott, R., Nagy, W. & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in grades 1 to 6. *Journal of Psycholinguistic Research*, 39, pp. 141-163.

Carlisle, J. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly*, 45(4), pp. 464-487.

Carstairs-McCarthy, A. (1992). *Current Morphology.* London & New York: Routledge.

Cartoni, B., Lefer, M-A. (2010), The MuLeXFoR database: representing word-formation processes in a multilingual lexicographical environment. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta.*

Collins English Dictionary – Complete and Unabridged. (1991, 1994, 1998, 2000, 2003). Accessed at http://www.thefreedictionary.com.

Dardano, M., Frenguelli, G. & Colella. G. (2006). What Lexicographers Do with Word Formation. In *Proceedings of Euralex 2006, Torino*, pp. 1115-1127.

De Caluwe, J. (forthcoming). Dictionary entries as windows on the onomasiological aspects of word formation. Forthcoming in *Proceedings of the Word formation in Electronic Dictionaries Conference, 26-27 May 2011*. Institut für Deutsche Sprache, Mannheim.

Delahunty, G., Garvey, J. (2010). *The English language: From sound to sense*. *Perspectives on Writing*. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press.

DiSciullo, A.M., Williams, E.S. (1987). *On the Definition of Word*. Cambridge: MIT Press.

Domenig M., ten Hacken P. (1992). *Word Manager: A System For Morphological Dictionaries*, Hildesheim: Olms Verlag.

Fellbaum, C., Miller, A. (2003). Morphosemantic links in WordNet. *Traitement Automatique des Langues*. 44(2), pp. 69-80.

Gove, P. (1966). Self-explanatory words. *American Speech.* 41(3), pp. 182-198.

Jackson, H., Zé Amvela, E. (2000). *Words, Meaning and Vocabulary: an Introduction to Modern English Lexicology*. London: Continuum.

Klosa, A., Schnörch, U. & Storjohann, P. (2006). ELEXIKO-A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim. In *Proceedings of the 12th EURALEX International Congress, EURALEX 2006, Turin, Italy, September 6th-9th, 2006*, pp. 425-430.

Lefer, M.-A. (2010). Word-formation in English-French bilingual dictionaries: the contribution of bilingual corpora. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Fryske Academy: Leeuwarden, pp. 810-823.

Merriam-Webster.com. (2011) Accessed at http://www.merriam-webster.com.

Nagy, W.E., Anderson, R.C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3), pp. 304-330.

Nagy, W.E., Anderson, R.C, Schommer, M., Scott, J.A. & Stallman, A.C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly* 24(3), pp. 263-282.

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.

Prcic, T. (1999). The treatment of affixes in the 'Big Four' EFL dictionaries. *International Journal of Lexicography*, 12(4), pp. 263-267.

Scalise, S., Guevara, E. (2005). The lexicalist approach to word-formation and the notion of the lexicon. In P. Stekauer, R. Lieber (eds.) *Handbook of word formation.* Dordrecht: Springer.

Storjohann, J. (2005) elexiko: A corpus-based monolingual German dictionary. *Hermes, Journal of Linguistics* 34, pp. 6-13.

ten Hacken, P. (1998). Word formation in electronic dictionaries. *Dictionaries*, 19, pp. 158-187.

ten Hacken, P. Abel, A. & Knapp, J. (2006). Word Formation in an Electronic Learners' Dictionary. *International Journal of Lexicography*, 19(3), pp. 243-256.

Webster's New World Dictionary, 2nd College ed. (1982). New York: Simon & Schuster.

Online specialised dictionaries: a critical survey

Valeria Caruso
University of Naples 'L'Orientale'
Largo S. Giovanni Maggiore, 30 – 80134 Napoli, Italy
Email: vcaruso78@gmail.com

Abstract

Online specialised free dictionaries offer assistance to everybody in need of information on the Web. More refined search tools are needed however in order to get as quickly as possible to the best terminological resource. At present Internet surfers can rely upon some lexicographical inventories and metasearch engines to acquire the information needed more easily. However, none of these resources offer evaluations of the collected vocabularies, or advice for a more efficient search. We present a critical inventory of 505 Web dictionaries, offering a broad overview of their main features. The analysis has been carried out using an evaluation form managed by a relational database (also published online) that assigns ratings and can also be used for analytical searches. The various fields of the form, and an adequate rating system allow us to define the specific parameters for three users' profiles (layman, semi-expert, expert) and four kinds of situations (cognitive, communicative, translation, learning) which are used for a qualitative analysis of dictionaries backed up by quantitative assessments.

Keywords: free online dictionaries; specialised lexicography; rating systems; lexicographical function theory

1. Free Dictionaries on the Web, an aid in terminology

Free specialised dictionaries have sprouted up everywhere on the Web, confirming the primacy of the Internet as a reference tool among the information media. These resources provide users with an extraordinary amount of data and no means to achieve the better information available in a reasonable searching time and effort. As Tarp (2010: 41) points out: "the risk of being suffocated by the overwhelming amount of data and suffering what could be called *information death* is omnipresent when browsing the Internet".

The great number of free dictionaries seems to be linked to the Web marketing strategies, which count dictionaries among the more reliable features to attract clients nowadays, since analysts (Lannoy, 2010) assure that browsing through dictionary pages warrants longer and in-depth visits of the host site.

Furthermore, the lemmatisation in alphabetic order is a fitting device to present contents on the Internet (Campoy Cubillo, 2002), and the 'Dictionary' appears to be the perfect text genera for the quick exchange of brief and thorough information in the digital space. In many cases 'dictionary' or, more often, the word 'glossary' labels summary pages of the industrial process relative to the products sold or advertised through the site[1].

Apart from their commercial implications, specialised dictionaries that users can access for free on the Web are useful sources for everybody needing readily available information. They can also be valuable orienteering tools at different levels of specialisation, depending on the kind of user in search of data, the specific needs that the dictionary should satisfy, and the context in which those needs are required.

These three factors constitute the basic parameters accounted for by the *functional lexicographic approach*, which is a prolific field of lexicographic research, proved to be useful both for dictionary writing and for their critical analysis (see, among others, Fuertes-Olivera, 2010; Nielsen & Mourier, 2007). This same frame of reference will be used here to evaluate the specialised free dictionaries available online, using a suitable judging form managed through a relational database, named *Web Linguistic Resources* [2], which is also published on-line to offer an effective guide to the Internet surfers in need of help with specialist terminology. The next pages will be devoted to the illustration of this inventory project of Internet dictionaries, which serves the 'democratic' aim of improving the potential reference function of the World Wide Web, since the reviewed vocabularies can all be accessed free of charge.

2. Guides and tools for surfers lost in definitions

The Web already offers valuable orienteering tools for quick and easy access to lexical resources. Metalexicographical sites present lists of vocabularies arranged by field[3] while metasearch engines display definitions taken from different glossaries[4] or, in other

[1] See the 'glossaire' of *Bojo Novo*, which is devoted to the illustration of the general maceration process of Beaujolaise wines ("Généralités sur la Macération Beaujolaise") in comparison with that used by the owners of the site ("Processus de vinification à la Cave Beaujolaise de Saint Vérand"). Both processes are described in paragraphs headed by titles which are not set in alphabetical order.

[2] The database is accessible at: www.weblinguisticresources.org, and also at: http://www.cila.unior.it/index.php/it/risorse-linguistiche.

[3] See for example *metadictionary dot com*, *Glossarist*, or the "Speciality and Language Dictionaries" page of *YOURDICTIONARY*.

[4] *MetaGlossary*.

cases, find links to the pages of various dictionaries containing the definitions required. The latter search option, available through *OneLook*, scans the most credited lexicons on the Web for 'major fields' such as business and law, while other domains, for example oenology, are not as well supplied and users can access a more limited number of definitions, which can be read on separate pages, since the browser shows only links, not contents.

Definitions from different dictionaries are instead displayed on a single page at *MetaGlossary*, which is suitable for quick comparisons but has a less valuable assortment of source dictionaries. However both browsers do not supply users with comprehensive information, if we take into account that the clearest definition attainable for a particularly counterintuitive term such as 'extra dry' (or 'extra-dry'), used in oenology for champagne classification, says: "A term used on Champagne labels to indicate not-quite-dry; not as dry as Brut" [5] . The matter would be more comprehensible if the definition initially clarified that champagne is classified according to a sweetness-level scale, where extra dry means 'more than dry', even though the champagne in question tastes sweet, since the scale is thus made: brut zero[6], extra brut, brut, extra dry ('extra sec' in French), dry ('sec'), medium dry ('demi-sec'), and sweet ('doux'). Only a whole explanation would made it clear that 'extra', in 'extra-dry', refers to a lower degree of sweetness in comparison to that found in dry champagnes. This information is nevertheless available in some dictionaries not listed by the metaresources quoted above, two of which [7] figure among the highest rated oenological dictionaries of the *Web Linguistic Resources* database.

The dictionary collection is designed to be extensive and evaluative at the same time. To this end an appraisal form, managed through a relational database, assigns ratings on the basis of each given field. The form also functions as a search device, since every evaluation field is also given as a search criteria. In this way the database users can look not only generically for medical dictionaries, but more specifically for those reporting, for examples, etymological notes too, such as the following:

> «Anoxia: 1. Strictly speaking, the absence of oxygen. 2. The near absence of oxygen. 3. Sometimes used loosely as a synonym for hypoxia. From an- (without) + -ox- (oxygen) + -ia == the state of being without oxygen»[8].

Started as a collection of rated vocabularies, which were initially judged by simply adding grades as the features of the lexicons increased (Caruso & Pellegrino, in press), the project has been improved for a more accurate analysis of the collected material in relation to its utility for Internet surfers, and therefore adopting the standpoint of the *lexicographical function theory* (particularly Tarp, 2008). Within this framework, dictionaries are «social and cultural products made by human beings in order to satisfy certain needs» (Tarp, 2009: 22), therefore lexicography is concerned with situations in which a dictionary is expected to be used and with the variables that make it better suited to its predictable functions. In fact, the lexicographical analysis is not concerned with direct inquiries of real vocabulary use, since lexicographers are already provided with the basic types of needs a dictionary is expected to satisfy. These are called "lexicographically relevant users situations" and represent the more general needs that dictionaries must satisfy.

For lexicographers a 'situation' is nevertheless something different from what a pragmaticists could have in mind, as they are generic circumstances under which a user needs to: a) increase or acquire new knowledge (therefore called a *cognitive* situation); b) manage communicative issues (a *communicative* situation); c) acquire a skill and know how to do something (the *operative* situation); or d) *interpret* symbols (the *interpretative* situation). Though the order of the different situations listed here reflects that used by Tarp (2009: 25) himself, it is clear that the four situations satisfy two basic functions, such as to provide encyclopedic knowledge and to fulfil linguistic needs, both of which can be seen from an active or passive point of view. Knowledge is achieved in *cognitive* situations, while it is transformed into action or skills in an *operative* environment; whereas language requires a symbolic interpretation before active production can be realized in communication. For a qualitative evaluation of Internet specialized free dictionaries in relation to their potential users, here we have concentrated on the two basic situations[9].

Since the World Wide Web is such an immense universe, used for all kinds of aims by all kinds of people, it is impossible to account for detailed users profiles and situations. It is more reasonable to adopt general categories and scan the free lexicographical material in order to offer valuable resources to everybody accessing the Web.

3. From ratings to users needs

The inventory of specialised dictionaries on the Web collects 505 resources and is realised through a judging form addressing both the dictionary macro- and

[5] *Sally's Place - A Wine Taster's Glossary*.

[6] The full scale is reported in the *D'Araprì Srl - Piccolo Dizionario dello Spumante e del Vino*, where many synonyms are given for brut zero: «brut integral, brut natur, brut nature, brut non dose, brut sauvage, brut zero, non dosage, pas dose, dosage zero, pas operé, nature».

[7] *WebFinance Inc. - Wine Define*, and *D'Araprì Srl - Piccolo Dizionario dello Spumante e del Vino*. Dictionary names are always preceded by the name of the host site.

[8] *MedicineNet.com - MedTerms medical dictionary*, entry: "Anoxia".

[9] Tarp (2008) himself outlined initially only two situations (the *cognitive* and the *communicative*) which were more recently doubled (Tarp, 2009: 25-26).

microstructure, which serves the purpose of collecting data analytically, but also of assigning scores, since fields are associated with grades. Grades assignment and the whole collection of data are managed through a relational database. Many features listed in the form can be evaluated as always present (the judging criteria is assigned a ‘yes’ answer and receives 5 points), not present (‘no’, assigns no grades), sometimes present (‘sometimes’ answer, 3 points). As we will see in the brief inventory that follows, the main characteristic of online dictionaries is their unsystematic nature, as they lack strict lexicographical organisation. It isn’t surprising than that the majority of them have a title containing the word ‘glossary’ – instead of ‘dictionary’. Their analysis is thus an exercise in careful reading and patient evaluation of lexical data hidden in the definitions.

In the evaluation form, the dictionary macrostructure is addressed by the *General Organisation*[10] field – whether it is arranged by alphabetical listings of the words, or by concepts or both; the *Number of Entries* (1-50, 50-100, more than 100); the *Access Structure* – whether by browsing or through a simple search engine or a smart one; the kind of *Word List* – if it has a one-word or a multi-word entry list; and the *Kind of Dictionary*. For this parameter the choice is among a *Multilingual Word List*, where only translational equivalences are given, a *Monolingual* dictionary or a *Multilingual* one. Besides these, a special Internet typology has been added, the *Plurilingual*, a dictionary comprised of many languages, whose glosses are written in the tongue of the entries, while cross-references between these languages is completely missing (see below, § 4).

If plurilingual dictionaries are in no way *Bidirectional* (another field of the evaluation form), this feature should characterise the multilingual ones. However less than a half (20) of the 41 bilingual dictionaries collected are equally accessible in both the languages involved. Moreover *Translation Equivalences* can also be found in monolingual dictionaries explaining specialist foreign words whose meaning could be obscure. See for example ‘terroir’ in two English dictionaries. While the first gives a translation, the second denies that English equivalences exist:

> Terroir: A term for 'soil', terroir refers to the set of geographic factors, such as soil composition, altitude, topography, position relative to the sun, sunlight hours and water drainage, that combine to create a unique taste and wine quality representative of a particular location.[11]
>
> Terroir: French term with no translation to describe the characteristics of a defined area. That includes soil, underground, exposure, climate and local traditions.[12]

If translation equivalences are the main feature of multilingual word lists, some of them also provide some *Encyclopedic Information*, such as the *Grabungswörterbuch (Eine Sammlung von Fachbegriffen für Grabungstechnik und Archäologie)*, which gives brief definitions for a small number of its headwords; or *Watson's wine glossary*, which sometimes provide brief ‘descriptions’[13] of terms, or the *Pescheria Gallina - I pesci*, a rich inventory of fish names in many different languages and also in many regional Italian language varieties. At the bottom of every page there’s always an encyclopedic piece of information, consisting in a list of the suggested months in which is it preferable to eat each species.

Another classification feature is the kind of *Lexicon* selected, which can have only specialist words or also ordinary language entries. For example, in *Investor Dictionary* users can find terms like *arrive*, *get back*, *add*, which are useful verbs in finance but have no special meaning in this sector. However the dictionary provides Explanatory *Phrases* (another field of the evaluating form) which show how the verb can be used in financial texts:

> **add**, verb, to put figures together to make a total
> Examples: If you add the interest to the capital you will get quite a large sum. • Interest is added monthly.

Referring to the microstructure, the form lists many fields. For *Cross References* and *Related Terms* separate fields are given: the former are direct cross-links among the dictionary entries, the latter are explicit references to other terms for a more detailed explanation of the subject (see figure 1).

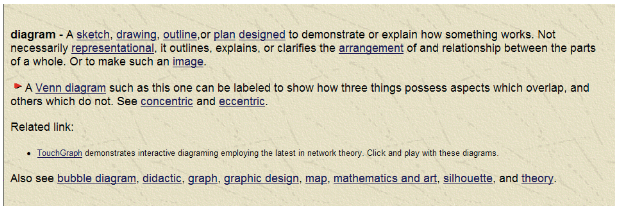


diagram - A sketch, drawing, outline,or plan designed to demonstrate or explain how something works. Not necessarily representational, it outlines, explains, or clarifies the arrangement of and relationship between the parts of a whole. Or to make such an image.

A Venn diagram such as this one can be labeled to show how three things possess aspects which overlap, and others which do not. See concentric and eccentric.

Related link:

- TouchGraph demonstrates interactive diagraming employing the latest in network theory. Click and play with these diagrams.

Also see bubble diagram, didactic, graph, graphic design, map, mathematics and art, silhouette, and theory.

Figure 1: Cross-links and related terms in the *ArtLex. Art Dictionary for artists, collectors, [...] and education*

Grammatical Category is also a separate field from *Morphological Indications*, which give evidence of every inflectional and derivational variation registered by the dictionary. Of the 505 dictionaries collected, 6% always give indications of the grammatical category, and 2% only report this information occasionally. Sporadic notes on morphology are given by 4% of dictionaries, even though they don’t resemble what is usually listed in a standard lexicographical work. In bilingual resources, morphological indications are used to give minimal

[10] Labels of the evaluative dictionary form fields are given here in Italics.

[11] *Jack's wine - Wine Glossary.*

[12] *French Wine a Day - Wine Glossary.*

[13] For example the description for ‘aroma’ is «especially referring to aromatic whites», that for ‘corredo aromatico’ is «referring to aroma».

advice on the inflectional variations of the second language, such as in the following:

> **Tonel (toneis)** (Portuguese) Large wooden cask(s) which lie on their sides, usually over 1000 litres capacity […].[14]

while in monolingual vocabularies they mainly serve to explain terms underlying their derivational formation:

> SUIDECIDE - (sui-decide) RATIONAL SUICIDE by a TERMINALLY ILL individual (Schmerl).[15]

Only five resources give systematic evidence of the morphological component of terms, two of them are etymologically oriented, such as the *Bio-Top - Lexique de Terminologie Médicale*, organised as an etymological dictionary, with entries grouped by word bases, and entries provided with a detailed account of etymological roots. The other, the *Dictionary of Botanical Epithets*, which is also a valuable learning resource for the Latin language, is made up of explanatory tables giving separate columns for 'Definition', 'Stem', 'Type/Gender' (reporting the grammatical category of the word), and 'Meaning' of the single morphemes into which the source Latin word has been analysed. The following is an extract relative to the headword 'acridens':

Epithet	Definition			
	Derivation	Stem	Type/Gender	Meaning
acridens	sharp teeth			
acridentes	*acer*	acr	adj	sharp, irritating, pungent; some spellings, espcially modern, give acris as the masculine instead of acer
	i	i	cnct	connective vowel used by botanical Latin
	dens	dent	noun/m	tooth
	Rubus acridens Bailey			

Systematic notes on morphology are also provided by *The Debenhams fashion dictionary*, since it lists 20 blends, followed by their source words, which represent the latest fashion neologism explained to assist disorientated customers with terms they cannot understand, such as:

> **Blurt** [Blouse/Skirt] all in one blouse and skirt combo.
> **Cardigown** [Cardigan/Dressing Gown] Cardigan, usually long and belted like a cardigan.

A significant inventory also for morphologists, still not at ease with blends (see Bauer, 1983).

Another resource which always gives etymological indications (*Etymology* is another field form), is the *Glossario Enologico* compiled at the University of Genoa collecting pieces of specialised vocabularies already published, so the contents given are perfectly in line with the standards of lexicographical tradition:

> olfactory examination […] Etimologia
> Olfactory: Latin. *Olfactōrius*, adjective, from latin *olfactor*, one who smells, *undefined*, a smellin […].
> Examination: late 14c., "action of testing or judging, judical inquiry," from O. Fr. *Examinacion*, from Latin *examinationem* (nom. *Examinatio*) […].

Furthermore, name explanations, which correspond to simplified etymologies for children, are always present in the *Enchanted Learning - Dinosaur and Paleontology Dictionary* in the form of short paraphrases: «Tyrannosaurus rex (meaning "tyrant lizard king")», «Tylosaurus (meaning "swollen lizard")», «Lesothosaurus, "Lizard from Lesotho, South Africa"». An extra page[16] (called *Hypertext* in the evaluation form) contains also notes on how dinosaurs' names are built, and offers a list of their roots and affixes.

Affixes are also listed in *aly-abbara.com - Lexique des affixes (préfixes et suffixes)*, an inventory of productive morphemes – and also lexemes – in the medical sector.
It is interesting as well to note the function assigned to etymological notes in dictionaries which use them occasionally. For example, the "Origins of Terms in International Economics" is a page in the *Deardorffs' Glossary of International Economics* dedicated to the first time some specialist terms were introduced, while almost all chocolate glossaries give notes on the origin of the word 'chocolate'. Particularly surprising are the examples of dictionaries which sometimes lack etymologies altogether, revealing in this way the different degrees of adaptation of the loanwords in specialised lexicons. In the *Wein.de - Weinglossar*, 'Amabile' is recognised as the Italian designation for a lovely taste, 'Aroma' and 'Aperitif' are explained through their Latin etymologies, however, for the adapted German borrowing 'Bukett' (from French 'bouquet') no etymological explanation is provided, while the term 'Finess' is simply considered as a synonym for 'Feinheit', 'subtil' or 'vornehm' in the wine jargon[17].

Other fields in the form register phonological information: *Phonetic Transcription*, *Pronunciation Annotation*, *Stress Information*, and *Syllabification*, since

[14] *Graham's Port – Glossary*.

[15] *The Vocabulary of Loss: A Glossary of Suicide-related Terminology*.

[16] "Dinosaur Name Roots: What Do Dinosaurs' Names Mean?".

[17] «Finesse: Synonym für "Feinheit", "subtil" oder "vornehm" in der Weinansprache».

not all of them might be given at the same time. Generally speaking, pronunciation notations signal the foreign words and their eventual degree of adaptation in the specialist lexicon of one language. Even though ‘ullage’ is provided with a translational equivalence, the suggested pronunciation shows that it is and adapted loanword in English:

> **Ullage** (UL-ij)—The space in a bottle between the wine and the cork. Also called “headspace”. If there is too much, the bottle has obviously leaked[18].

Many dictionaries however don’t give accent annotations with pronunciation:

> **AGIORGITIKO**
> Pronounced “Ah-jee-or-jee tee-koh”. (a.k.a St. George)[19].

while some others give full written indications and also audio files, as in the *National Cancer Institute - Dictionary of cancer terms*.

A noteworthy use of *Audio Files* (label of another field form) is made by the English version of the *NHGRI - Talking Glossary of Genetics*, a learning project of the National Human Genome Research Institute (NHGRI), which gives audio definitions: spoken explanations given by the Institute scientists, as well as illustrations and many 3-D animations as extra *Learning Resources* (label of a field in form).

Other main linguistic features accounted for in the inventory are the *Frequency of Use*, *Linguistic Variation*, *Idioms*, *Collocations*, and *Examples*. The variations registered, relative to the *Linguistic Variation* field, are relative to geography and space (i. e. dialects, regional varieties), style:

> **Aroma**: Olor agradable cuyo bouquet es la expresión más refinada./Aroma: Pleasant smell. Its bouquet is the most refined expression[20]

context:

> **alopecia**
> **English**: Technical term: alopecia/ Popular term: baldness
> **Danish**: Technical term: alopecia/Popular term: skaldethed […]
> **Spanish**: Technical term: alopecia (nf)/Popular term: caída general o parcial de cabellos o pelos […][21]

but also to other differences over time and history:

> **Basque**: Section of bodice below waist, shaped to hips; late c20th name for corset[22]

or in spelling:

> **log** […] (Note the verbs can be spelled log on, log-on, or logon; log off, log-off or logoff)[23].

Annotations about the frequency of use are rare, and are conveyed through generic statements of the kind of «palabra muy empleada»[24], in Spanish, or «sometimes called…» [25] to specify the difference between two synonyms.
Some indications about collocations have been found in a small percentage of dictionaries (around 10%). They needed to be carefully detected since they are given in the body of the texts without any explicit signalling:

> **aroma**: [...] One might speak of the “floral aroma” of a Riesling, for example[26].

and sometimes they are provided in a prescriptive form:

> **inoculate**: […] (Note You inoculate someone with or against a disease.)[27]

Even more complicated is the search for idioms, since it requires a comprehensive reading of all the vocabularies collected, in order to distinguish idiomatic expressions[28] correctly. At present we can only give one example, but we cannot provide any quantitative analysis:

> **Catch a Falling Knife**:
> To catch a falling knife is an idiomatic expression which is used in investments. It is a phrase that refers to a dangerous investment strategy such as stocks that drop tremendously resulting to worthless investments[29].

In addition, dictionaries are valued as technical or non-technical on the base of their *Definitions*. These could be self-explanatory and non-technical, in the sense that no specific subject knowledge or only little effort is required in order to understand them. On the other end they could necessitate a certain degree of previous knowledge, being written for experts in the field – so they are defined as technical. Of the following definitions, the first is considered to be technical, the

[18] *Napa Now - Glossary of Wine Terms*.

[19] *Epicurus.com - Wine Glossary*.

[20] *Welcome Argentina - El Lenguaje del Vino/The language of wine*.

[21] This is an excerpt from the entry ‘alopecia’ in *Multilingual Glossary of technical and popular medical terms in nine European Languages*.

[22] *Dictionary of Corset-related Words and Terms*.

[23] *WebFinance - Computing-Dictionary*.

[24] From the entry ‘Cuerpo’ in *Welcome Argentina - El Lenguaje del Vino/The language of wine*.

[25] From the entry ‘Parole revocation hearing’ in *Crime Victims Services-Criminal Justice System - Glossary of Terms*.

[26] *Top Side Wine and Spirits - Wine Tasting Terms*.

[27] From *Medical-Glossary.com*.

[28] We refer to the definition of *idiom* given by Ayto (2006): "The term idiom may be defined as an institutionalized multiword construction, the meaning of which cannot be fully deduced from the meaning of its constituent words, and which may be regarded as a self-contained lexical item".

[29] From *Investor Dictionary*.

second non-technical:

> **Gametes:** A collective term for haploid reproductive cells (germ cells; male sperm cells or spermatozoon, and female egg cells (oocytes, ovum) that fuse to form a diploid cell, the zygote from which multicellular organisms develop.
> For related information see also: Cell types[30].

> **Gamete**: Mature male or female reproductive cell (sperm or ovum) with a haploid set of chromosomes (23 for humans)[31].

As seen in the first entry above, *Examples* are additional features for a better understanding, but also *Synonyms*, *Antonyms*, *Hypernyms*, *Hyponims*, and the indication of the *Domain Field* can serve this purpose. About a half of the dictionaries collected (42%) have synonyms, a small number give antonyms (13%), and only one[32] gives also hyponims and hypernyms, while these conceptual relations are necessary to clarify generic or specific terms in definitions:

> Action: […] parlando di mercati finanziari, la tipica Action è quella di comprare (Buy) o di vendere (Sell) dei titoli quotati[33].

Other features listed for the digital dictionaries collected, are *Video Files* and *Pictures*. However, only 2% of the present inventory offers video files, while 12% has images. The dictionaries can also be hosted by different *Kind Of Sites*: *Amateur*, *Commercial*, *Institutional*, *Collective*, *Specialised*. While *Amateur* and *Collective* refer to the dictionary authors, the others specify the main character of the host site. In particular *Collective* dictionaries are user-made resource such as WIKITIONARY (see Fuertes-Olivera, 2009), while *Institutional* sites are only those belonging to public or private Institutions, such as foundations, universities, and research organisations.

The various fields of the evaluation form are associated with marks, and this allows us to assign a *General Rating*, which gives a rough judgment of the variety of information and languages present in a single dictionary – since every translation language warrants 5 points and each foreign language present as a foreign entry word gives 3 additional points.

However, in order to evaluate the collected dictionaries on the base of their usability for Internet surfers, a qualitative analysis has also been carried out. It will be discussed below in § 5 and 7, after a brief explanation of the new kind of dictionaries hosted by websites, that we have called plurilingual.

4. Lexicology between translation and marketing: plurilingual dictionaries

The plurilingual dictionaries are generally hosted by multilingual commercial sites whose contents have undergone what is technically called 'a localization process', the specialized translation activity with high-technological expertise that combines cultural needs with selling requirements. At present the most valuable strategy for trading on a global scale through the Internet is considered to be the capability of offering linguistically and culturally adequate contents. Products and their selling correlates, such as websites or packaging, must have the specific target culture requirements, so translators have become part of the industrial production. Dictionaries are no exception to this role, and major wine producers try to offer lexical resources on their websites to attract wine lovers and diffuse wine culture as much as possible. So far as we can see from the examples collected[34], *plurilingual* dictionaries are translations of a given vocabulary across the various language versions of the same website, with no cross-reference between them. Very often the translated dictionaries are reductions of the original, both in the number of entries and text contents, as in the wine glossary of Sapareta, a wine producer:

> *It*. CORPOSO: Si usa per indicare in un vino una piacevole ricchezza di componenti fra di loro equilibrati, soprattutto di estratti e alcol, che presenta colore e sapore in armonia.
> *En*. FULL-FLAVOURED: used to indicate a wine with a delightful richness of balanced components.
> *Fr*. ETOFFÉ: s'emploie pour indiquer un vin dont on apprécie la richesse des composants bien équilibrés entre eux[35].

The Italian 'corposo' in the plurilingual dictionary of Cavit has a different English correspondent term, 'full-bodied', while the French and German version seem to provide no translation at all for that, since the dictionary versions of these languages are shorter than the Italian (one fewer headword) and the English one (two fewer entry words). In other cases it is possible to browse the entry lists of different languages of this same dictionary and try to find linguistic correspondences, such as:

> *It*. MATURO: vino che ha raggiunto lo stadio ottimale di maturazione./*En*. MATURE: wine which has reached an optimum stage of maturation./*Ge*. ROBUST: Ein Wein, der reich an Alkohol und Körper ist./ *Fr*. VIN FAIT: vin qui a atteint son vieillissement optimal.

[30] *C O P E - Cytokines & Cells Online Pathfinder Encyclopedia*.

[31] *BERIS - DOE Human Genome Program - Genome Glossary*.

[32] *C O P E - Cytokines & Cells Online Pathfinder Encyclopedia*, and *Terminologie zur Lichtplanung und Lichtsimulation / Lighting Design and Simulation Terminology*.

[33] *Il Faro Finanziario - Glossario. Dizionario dei termini in uso nei mercati anglosassoni*.

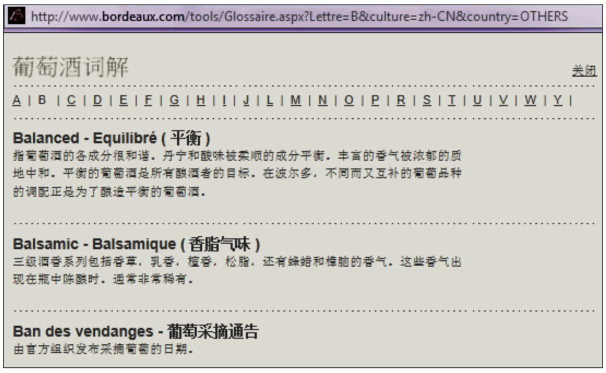
[34] Only 20 of the 505 dictionaries collected belong to the plurilingual type.

[35] *Azienda Agricola Sapereta – Glossario del vino*.

Though not particularly reliable as cross-linguistic reference tools, the offspring of this kind of dictionary is worth the notice, as it is the product of a wider Internet process (*Content Localization*), involving lexicographical resources.

The same marketing process has given birth to hybrids, dictionaries which are conceived as plurilingual, with no cross-reference system between the different languages, but giving many translational equivalences in brackets anyway, such as in the German and English versions of the *Bordeaux.com* lexicon.

However in Korean, Chinese, and Japanese this same dictionary is actually multilingual, indexing French terms, followed by English equivalents, and giving the Chinese, Korean, or Japanese correspondence in brackets. The following is the Chinese one (figure 2):



http://www.bordeaux.com/tools/Glossaire.aspx?Lettre=B&culture=zh-CN&country=OTHERS

葡萄酒词解

关闭

A | B | C | D | E | F | G | H | I | J | L | M | N | O | P | R | S | T | U | V | W | Y |

Balanced - Equilibré (平衡)
指葡萄酒的各成分很和谐．丹宁和酸味被柔顺的成分平衡．丰富的香气被浓郁的质地中和．平衡的葡萄酒是所有酿酒者的目标．在波尔多，不同而又互补的葡萄品种的调配正是为了酿造平衡的葡萄酒．

Balsamic - Balsamique (香脂气味)
三级酒香系列包括香草，乳香，檀香，松脂，还有蜂蜡和樟脑的香气．这些香气出现在瓶中陈酿时．通常非常稀有．

Ban des vendanges - 葡萄采摘通告
由官方组织发布采摘葡萄的日期．

Figure 2: Chinese section of *Bordeaux.com*

5. A fitting analysis for users needs

As previously mentioned (§ 3), the fields used in the evaluation form can serve also for qualitative analysis. Referring to the *lexicographically relevant users situations* already discussed, we can attribute ratings to each specific field taken into consideration as characterising features for the *cognitive* and *communicative* situations and, in addition to these, for *translation* and *learning*, other two situations that could be useful to Internet surfers.

We can also estimate the kind of user accessing the specialized Web dictionaries in relation to their level of field expertise. For this aspect we refer to the distinction made by Bergenholtz & Kaufmann (1977: 101-102) among laymen, semi-experts and experts. In particular semi-experts are «experts from other related subject fields» who are confronted daily with other sectors, for example journalists writing about scientific issues, or political advisors or workers in the public administration that are particularly familiar with sectors related to their professional activities. The experts are instead considered by Bergenholtz & Kaufmann as the kind of users that do not rely upon dictionaries, since they refer to other sources in order to acquire new knowledge. Though this assumption is unquestionable, the present aim is different from that of Bergenholtz and Kaufmann, and other professional lexicographers. They need clear instruction in order to write valuable dictionaries, while we offer orienteering tools for quicker and successful Internet surfing. A more refined scale of expertise (divided in layman, semi-expert and expert) will help to achieve better this purpose.

The rating scale is explicitly designed to obtain clear orienteering indications: 2 points are given to the most characterising features, 1 point to less important ones and penalizing marks (-1 and -2) help to avoid contradictory responses, such as dictionaries highly valuable for experts and laymen, or for cognition and communication at the same time. For this purpose the rating distribution of the *Definition* field is given as such:

Definitions	Layman	Semi-expert	Expert
Technical	-2	1	2
Non-Technical	2	-2	-2

All users' profiles may reach 13 points maximum, the cognitive and communicative functions can reach 16, while learning and translation 15. The following list shows how points are given[36]:

Layman profile
Cross-references: Yes, 2; Smt.[37], 1/ Definitions: Technical, -2; Non-Technical, 2/ Encyclopaedic information: Yes, 2/ Examples: Yes, 2; Smt., 1/ Kind of site: Institutional, 1; Specialised, 1/ Lexikon: Specialist & Ordinary Words/ Explanatory Phrases: Yes, 2; Smt., 1/ Pronunciation notation: Yes, 2; Smt., 1/ Quotations: Yes, -2; Smt., -1/

Semi-expert profile
Access structure: Search engine, 1; Smart search engine, 1/ Bibliographic resources: Yes, 2/ Definitions: Technical, 1/ Etymology: Yes, 1; Smt., 2/ Kind of site: Institutional, 2; Specialised, 1/ Lexicon: Only Specialist words, 2/ Linguistic variation: Yes, 2; Smt., 1/ Quotations: Yes, 2; Smt., 1/

Expert profile
Access structure: Browse, -2; Search engine, -1; Smart search engine, 1/ Bibliographic resources: Yes, 2/ Definitions: Technical, 2; Non-Technical, -2/ Entries number: 0-50, -2, 50-100, -1, over 100, 1/ Etymology: Yes, 2/ Etymology: Smt., 1/ Kind of site: Institutional, 1; Specialised, 1/ Phonetic transcription: Yes, 2; Smt., 1/ Quotations: Yes, 2; No, -2/

Cognitive situation
Antonyms: Yes, 2; Smt., 1/ Domain field: Yes, 2; Smt., 1/ Encyclopaedic information: Yes, 2; Smt., -1; No, -2/ Kind of site: Institutional, 2; Specialised, 2/ Pictures: Yes, 2;

[36] Since the inventory of idioms is still to be done, they couldn't be used as rated features at present.
[37] Smt.= sometimes.

Smt., 1/ Related terms: Yes, 2; Smt., 1/ Synonyms: Yes, 2; Smt., 1/ Video files: Yes, 2; Smt., 1/

Communicative situation
Audio files: Yes, 2; Smt., 1/ Collocations: Yes, 2; Smt., 1/ Frequency of use: Yes, 2; Smt., 1/ Kind of site: Institutional, 2; Specialised, 2/ Linguistic variation: Yes, 2; Smt., 1/ Stress information: Yes, 2; Smt., 1/ Syllabification: Yes, 2; Smt., 1/ Synonyms: Yes, 2; Smt., 1/

Translation
Bidirectionality: Yes, 2/ Cultural notes: Yes, 2/ Kind of dictionary: Multilingual, 2; Multilingual word list, 1; Plurilingual, 2/ Kind of site: Institutional, 2; Specialised, 2/ Explanatory Phrases: Yes, 2; Smt., 1/ Translation equivalences: Yes, 2; Smt., 1/

Learning
Audio files: Yes, 2; Smt., 1/ Grammatical category: Yes, 2; Smt., 1/ Hyperlinks: Yes, 2; Yes, 2/ Hypertexts (explanatory pages): Yes, 2/ Kind of dictionary: Multilingual, 2; Monolingual, 2/ Kind of site: Institutional, 2; Specialised, 2/ Learning resources, 2/ Morphological indications: Yes, 2; Smt., 1/

6. Conclusions

Using this grading system, we are able to present an evaluative estimation of the data collected (see Figure 3).

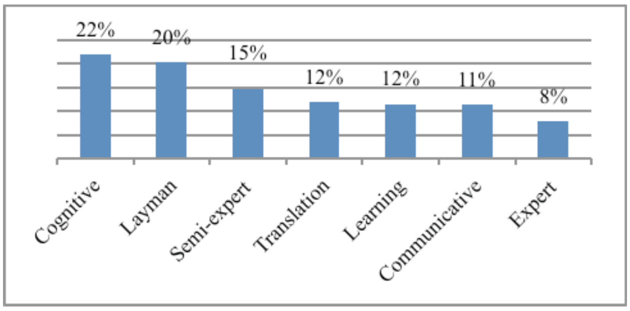


Figure 3: Percentage of users' profiles and lexicographical situations in the collected dictionaries.

The majority of dictionaries satisfy the parameter of the cognitive situation (22%), while the communicative obtains half this result (11%). Translation and learning obtain similar results, being equally represented by 12% of the collected resources while, referring to users' profiles, Laymen (20%) and semi-experts (15%) can quite similarly find adequate tools on the web. Only experts can count on fewer resources to refer to (8%).

In addition, only a few dictionaries highly satisfy the requirements for translation, learning, cognitive situation, expert and semi-expert profiles, since on average this parameters receive low scores (see Figure 4). On the contrary, a lot of Internet resources gain high marks for the layman profile, while communication is the lowest rated of all, so the average marks gained by dictionaries for this profile aren't much lower than the highest score obtained.

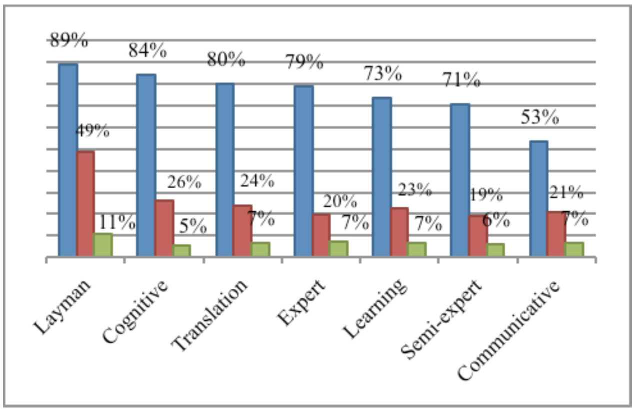


Figure 4: Highest (in blue), lowest (green) and average (red) scores for each user profile and lexicographical situation in the collected dictionaries.

This means that the Internet offers many useful resources for the layman, while only a few are well suited for experts and semi-experts. The same can be said for translation, learning and cognitive situation, which can only count on a few good dictionaries. Whereas for communicative needs, users are not provided with valuable free lexicographical tools yet.

The table below (table 1) lists the best and worst rated dictionaries for each category:

Situation/Profile	Dictionary Name	%
Cognitive	ArtLex - Art Dictionary for artists, collectors, ….	84%
Cognitive	Italian VI Trading Wine Glossary	5%
Communicative	Wine Lovers Page - Wine Lexicon	53%
Communicative	Wine Road - Glossary	7%
Expert	Glossario Enologico	79%
Expert	Whonamedit? A dictionary of medical eponyms	7%
Layman	ArtLex - Art Dictionary for artists, collectors, ….	89%
Layman	Watson's wine glossary	11%
Learning	Math Spoken Here! An Arithmetic and Algebra Dictionary	73%
Learning	Wein-plus - Translator for wine terms	7%
Semi-expert	WebFinance - Computing-Dictionary	71%
Semi-expert	Wineeducation.com - Wine Glossary	6%
Translation	DiCoInfo - Le dictionnaire fondamental de l'informatique et de l'Internet	80%
Translation	Winetasting.com - Wine Glossary	7%

Table 1: Highest and lower rated dictionaries for each user profile and lexicographical function.

While another report (table 2) can be given on the basis of the additive rating system described in §3, which gives particular evidence of the number of different languages used in the dictionaries:

Dictionary Name	rating
Scouting Dictionary	203
Glossario Enologico	118
Multilingual Glossary of technical and popular medical terms in nine European Languages	118
DermIS.net (Dermatology Information System)	117
AskPhil - Glossary of Stamp Collecting Terms	96
ArtLex - Art Dictionary for artists, collectors, ….	94
CILF - Dictionnaire Commercial	93
HON Foundation - List of rare diseases	92
Islamic Philosophy Online - Dictionary of Islamic Philosophical Terms	88
Wein-Plus Wein-Glossar Ehrenfelser	85

Table 2: The best rated sites.

However, while the general estimation of features provided might be valuable for quite a long time, the single dictionaries listed in the tables will probably disappear soon, since things change quickly on the Web. The present research started some months ago with an inventory of more than 700 Internet dictionaries, and 200 of them have since vanished. Some of the vocabularies collected are digital versions of printed books, while only one was created on the Internet and then subsequently printed[38]. We have tried to give as many examples as possible throughout our inventory, in order to give future evidence of this vanishing repertory of writing.

7. Acknowledgements

A special acknowledgement to Gianluca Monti for managing the database and providing the project with all the technical assistance needed.

The present research has been sustained by academic grants from the University of Naples ‘L'Orientale’.

8. References

Ayto, J. (2006). Idioms. In K. Brown (ed.) *Encyclopedia of language & linguistics*. Amsterdam: Elsevier, vol. 5, pp. 518-521.

Bauer, L. (1983). *English Word-Formation*. Cambridge: Cambridge University Press, pp. 234-237.

Bergenholtz, H., Kaufmann, U. (1997) Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. *Hermes, Journal of Linguistics*, 18, pp. 91-127.

Bergenholtz, H., Tarp, S. (eds.) (1995). *Manual of Specialised Lexicography*. Amsterdam, Philadelphia: John Benjamins.

Caruso, V., Pellegrino, E. (in press) Metadizionari digitali specialistici. In S. Ferreri (ed.) *Atti del XLIV Congresso della Società di Linguistica Italiana* (SLI). Università degli Studi della Tuscia, Viterbo, Italy.

Campoy Cubillo, M.C. (2002). General and specialised free online dictionaries. In *Teaching English with technology*, 2/3. Accessed at: http://www.iatefl.org.pl/call/j_review9.html.

Fuertes-Olivera, P.A. (2009) The Function Theory of Lexicography and Electronic Dictionaries: WIKTIONARY as a Prototype of Collective free Multiple-language Internet Dictionary. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a crossroads: dictionaries and encyclopedias today, Lexicographical Tools Tomorrow*. Bern: Peter Lang.

Fuertes-Olivera, P. A. (2010). 32- Lexicography for the third Millennium: Free Institutional Internet terminological dictionaries for learners. In P. Fuertes-Olivera (ed.) *Specialised Dictionaries for Learners*. Berlin-New York: De Gruyter, pp. 193-223.

Guinan, J. (ed.) (2009) *The Investopedia Guide to Wall Speak. The Terms You Need to Know to Talk like Cramer, think like Soros, and buy like Buffett*. New York: McGraw-Hill.

Lannoy, V. (2010). Free on-line dictionaries: why and how?. In S. Granger, M. Paquot, (eds.) *eLexicography in the 21st century. New challenges, new applications*, Louvain-la-Neuve: UCL - Presses Universitaire de Louvain, pp. 173-182.

Nielsen, S., Mourier, L. (2007). Design of a function-based Internet Accounting Dictionary. In H. Gottlieb, J.E. Mogensen (eds.) *Dictionary Visions, research and Practice*. Amsterdam: John Benjamins, pp. 119-135.

Piotrowski, T. (2009). Review. Sven Tarp. Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography. *International Journal of Lexicography*, 22(4), pp. 480-486.

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, volume 134. Tübingen: Max Niemeyer Verlag.

Tarp, S. (2010). Functions of Specialized Learners Dictionaries. In P. Fuertes-Olivera (ed.) *Specialised dictionaries for learners*. Lexicographica. Series Maior. Berlin-New York: De Gruyter, pp. 39-53.

Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp. 17-32.

8.1 Web Dictionaries

Abraham's Interactive Textile Dictionary. Accessed at: http://www.editeam.com/dictionary.asp.

Agricamping.it - Dizionario Enologico. Accessed at: http://www.agricamping.it/agricamping-vino/dizionario-enologico-a.htm.

aly-abbara.com - Lexique des affixes (préfixes et suffixes). Accessed at: http://www.aly-abbara.com/litterature/medicale/affixe

[38] *Investopedia - Dictionary of Financial Terms*, printed as Guinan (2009).

s/a.html.

ArtLex. Art Dictionary for artists, collectors, students and educators in art production, criticism, history, aesthetics, and education. Accessed at: http://www.artlex.com/.

BERIS - DOE Human Genome Program - Genome Glossary. Accessed at: http://www.ornl.gov/sci/techresources/Human_Genome/glossary/.

Bio-Top - Lexique de Terminologie Médicale. Accessed at: http://georges.dolisi.free.fr/Terminologie/Menu/terminologie__medicale_menu.htm.

Bojo Novo. Accessed at: http://www.bojonovo.com/index.php?id=48&L=3%22%20onfocus%3D%22blurLink%28this%29%3B.

COPE - Cytokines & Cells Online Pathfinder Encyclopedia. Accessed at: http://www.copewithcytokines.de/cope.cgi.

Crime Victims Services-Criminal Justice System - Glossary of Terms. Accessed at: http://www.azdps.gov/Services/Crime_Victims/cjs/glossary/.

D'Arapri Srl - Piccolo Dizionario dello Spumante e del Vino. Accessed at: http://www.darapri.it/glos2.htm.

Deardorffs' Glossary of International Economics. Accessed at: http://www-personal.umich.edu/~alandear/glossary/.

Dictionary of Botanical Epithets. Accessed at: http://www.winternet.com/~chuckg/dictionary.html.

Dictionary of Corset-related Words and Terms. Accessed at: http://www.staylace.com/resourcelist/diction.htm.

DLG Wein.de – Weinglossar. Accessed at: http://www.wein.de/weinglossar.0.html.

Enchanted Learning - Dinosaur and Paleontology Dictionary. Accessed at: http://www.enchantedlearning.com/subjects/dinosaurs/glossary/.

Epicurus.com - Wine Glossary. Accessed at: http://www.epicurus.com/Glossary/wine/glossary.php.

French Wine a Day - Wine Glossary. Accessed at: http://a-la-recherche-du-vin.typepad.com/french_wine._a_day/2005/04/wine_glossary.html.

Glossario Enologico. Accessed at: http://www.farum.it/glos_enol/list.php.

Glossarist. Accessed at: http://www.glossarist.com/gsearch.asp.

Graham's Port - Glossary. Accessed at: http://malvedos.wordpress.com/glossary/.

Il Faro Finanziario - Glossario. Dizionario dei termini in uso nei mercati anglosassoni. Accessed at: http://www.schorsch.com/de/wissen/glossar/.

Investopedia - Dictionary of Financial Terms. Accessed at: http://www.investopedia.com/dictionary/.

Investor Dictionary. Accessed at: http://www.investordictionary.com/dictionary/.

Jack's wine - Wine Glossary. Accessed at: http://www.jackswine.com.au/wine-glossary.php.

Medical-Glossary.com. Accessed at: http://www.medical-glossary.com/.

Medical-Glossary.com. Accessed at: http://www.medical-glossary.com/.

MedicineNet.com - MedTerms medical dictionary. Accessed at: http://www.medterms.com/script/main/hp.asp.

Metadictionary Dot Com. Accessed at: http://www.metadictionary.com/index.htm.

MetaGlossary. Accessed at: MetaGlossary.com.

Napa Now - Glossary of Wine Terms. Accessed at: http://www.napanow.com/wine.glossary.html.

National Cancer Institute - Dictionary of cancer terms. Accessed at: http://diabetes.niddk.nih.gov/dm/pubs/dictionary/index.aspx.

National Diabetes Information Clearinghouse (NDIC) - Diabetes Dictionary. Accessed at: http://diabetes.niddk.nih.gov/dm/pubs/dictionary/index.aspx.

NHGRI - Talking Glossary of Genetics. Accessed at: http://www.genome.gov/glossary/index.cfm.

OneLook. Accessed at: http://www.onelook.com/.

Sally's Place - A Wine Taster's Glossary. Accessed at: http://www.sallybernstein.com/beverages/wine/wine_glossery.htm.

Sapareta. Accessed at: http://www.sapereonline.it/it/vino/glossario_enologico.html.

Terminologie zur Lichtplanung und Lichtsimulation/Lighting Design and Simulation Terminology. Accessed at: http://www.schorsch.com/de/wissen/glossar/.

The Debenhams fashion dictionary. Accessed at: http://blog.debenhams.com/the-debenhams-fashion-dictionary/trends/.

The Vocabulary of Loss: A Glossary of Suicide-related Terminology. Accessed at: http://lifegard.tripod.com/index-2.html.

Top Side Wine and Spirits - Wine Tasting Terms. Accessed at: http://topsidewineandspirits.com/index.php?page=wine-tasting-terms.

WebFinance - Computing-Dictionary. Accessed at: http://www.computing-dictionary.com/definition/log.html.

WebFinance Inc. - Wine Define. Accessed at: http://www.winedefine.com/.

Welcome Argentina - El Lenguaje del Vino/The language of wine. Accessed at: http://www.welcomeargentina.com/vino/vocabulario.html.

Wikitionary. Accessed at: http://en.wiktionary.org/wiki/Wiktionary:Main_Page.

YOURDICTIONARY. Accessed at: http://www.yourdictionary.com/diction1.html.

Visualizing sloWNet

Darja Fišer, Jernej Novak
University of Ljubljana, Faculty of Arts, Department of Translation
Aškerčeva 2, 1000 Ljubjana, Slovenia
University of Maribor, Faculty of Electrical Engineering and Computer Science, Institute of Informatics
Smetanova 17, 2000 Maribor, Slovenia
E-mail: darja.fiser@ff.uni-lj.si, jernej.novak1@uni-mb.si

Abstract

With the increasing popularity of semantic lexica such as wordnets that are being developed for more and more languages the need for tools which enable displaying and management of their content has risen as well. Dictionary writing systems or tools for ontology management are not suitable for use with wordnets because they are concept-based and relational on the one hand but less formal and more language-oriented on the other. Several specialized wordnet tools have been developed but it is still very difficult to find an all-in-one solution that would freely available and would enable on-line browsing, editing as well as visualization of wordnet content in a mono- as well as a multilingual setting. The goal of this paper is to close this gap with a light-weight and easily portable, browser-independent wordnet tool called sloWTool which supports easy importing of new wordnets or wordnet-like databases from the standard formats such as LMF and DebVisDic XML. The tool also allows adding external, third-party resources, such as wordnet domain hierarchy, coarse-grained sense clusters, and a database of images that are linked to wordnet synsets.

Keywords: wordnet; wordnet browser; wordnet editor; wordnet visualization

1. Introduction

Wordnets are semantic lexicons that have become increasingly popular in the past decade and have been developed, first for English (Fellbaum, 1998) and then also for a number of other languages (see EuroWordnet, BalkaNet, AsianWordNet), including Slovene, which has been developed automatically from heterogeneous resources, such as bilingual dictionaries, bilingual thesauri and parallel corpora (Fišer, 2009).

Wordnets differ from traditional machine-readable dictionaries with their concept-based organization and an explicit encoding of semantic information. To a certain extent, wordnets are similar to ontologies, commonly used for AI tasks in that they too define a set of semantic relations which interlink concepts into a semantic network but wordnets are less formalized and more language-oriented. This is why neither dictionary writing systems such as TshwaneLex (Joffe and de Schryver, 2004) nor tools for ontology development and maintenance such as Protégé (Noy et al., 2003) are not suitable for wordnets.

While several browsers, editors and visualization tools have been developed to deal specifically with wordnets (e.g. Fellbaum, 1998; Louw, 1997; Horak, 2006), they too are not easy use with a new wordnet due to several reasons: many of them are not publicly available, they might be intended as desktop applications for off-line browsing, do not enable the use of several wordnets in parallel, do not allow for simple editing of synsets, and do not include any visualization options. These obstacles consequently encouraged us to develop our own tool for browsing, editing and visualizing wordnet content which we present in this paper.

The paper is structured as follows: in Section 2 we present sloWNet, Section 3 analyses already existing wordnet tools, in Section 4 we present the features of sloWTool which we developed for browsing, editing and visualization of sloWNet, and then conclude the paper with final remarks and ideas for future work.

2. sloWNet

Slovene wordnet was built automatically in three stages, each time using a different approach according to the resources used for extracting the relevant lexico-semantic information. The first and most straightforward approach relied on an existing Serbian wordnet and then translated the literals into Slovene with a traditional digitized bilingual dictionary (Erjavec and Fišer, 2006). This simple approach lacked automatic disambiguation of polysemous dictionary entries and therefore required a lot of manual cleaning. This was improved in the second approach which was able to assign the correct wordnet sense to a Slovene equivalent by disambiguating it with a word-aligned parallel multilingual corpus and already existing wordnets for several languages (Fišer, 2009). The main contribution of the third and final approach was the extraction of a large number of monosemous specialized vocabulary and multi-word expressions from Wikipedia and its related resources (Fišer and Sagot, 2008). The developed wordnet contained about 17,000 literal which belonged to roughly 20,000 synsets.

Since then, sloWNet has undergone two cycles of manual revision; manually validation of all Base Concept Sets (about 5,000), and editing of all nominal synsets included in the semantic annotation of the corpus (about 1,000, see Fišer and Erjavec, 2010).

The final major step in the development of sloWNet 3.0 is the recent large-scale automatic extension in which we combined all the resources from the previous steps in order to exploit them to their full potential and thereby improve coverage of sloWNet without compromising its quality. First, a model was trained on the existing elements in sloWNet, and then a maximum entropy classifier was used to determine appropriate senses of translation candidates extracted from the heterogeneous resources described above (see Sagot and Fišer, forthcoming).

no. of synsets			no. of literals			no. of (synset, literal) pairs		
	PWN3.0	sloWNet3.0		PWN3.0	sloWNet3.0		PWN3.0	sloWNet3.0
Adj	18,156	6,218	Adj	21,538	5,108	Adj	30,004	12,438
Adv	3,621	453	Adv	4,481	514	Adv	5,580	847
N	82,114	30,911	N	119,034	30,319	N	146,345	55,383
V	13,767	5,337	V	11,531	3,840	V	25,047	14,053
total:	117,658	42,919	total:	156,584	39,781	total:	206,976	82,721
BCS1	1,220	1,220	monosemous	130,208	26,339	avg. synset length	1.76	1.92
BCS2	2,213	2,213	mwe	64,383	9,050	avg. polys.-all	1.51	2.07
BCS3	1,238	1,238	proper names	35,002	2,946	avg. polys.-poly	3.39	4.19
total:	4,671	4,671	non-letter lit.	178	32			

Table 1: A comparison of Princeton WordNet 3.0 and sloWNet 3.0

As Table 1 shows, the current version of Slovene wordnet contains 36% of all the synsets in Princeton WordNet. Nouns are still by far the most frequent, representing more than 70% of all synsets. sloWNet contains all synsets from the Base Concept Sets but also a lot of specialized vocabulary; 66% of all the literals in it are monosemous. The extended sloWNet also contains a lot of multi-word expressions and proper names, which are both mostly nominal. A comparison of the average number of literals per synset and average level of polysemy between sloWNet and PWN is interesting because it can indicate how accurate the automatic population of Slovene synsets was. While average synset length is comparable to PWN, the total average polysemy (2.07 vs. 1.51) and the average polysemy excluding monosemous words (4.19 vs. 3.39) show that Slovene wordnet contains noise that will have to be filtered out in the future.

The fact that sloWNet is somewhat noisy due to the automatic construction process is further indicated by the number of literals in the longest synsets which are, at first glance, quite similar to PWN (see Table 2) but a more careful analysis shows that even though these synsets contain several synonyms, not all of them are correct and should therefore be filtered out in the future. This is even more obvious when the most polysemous literals are searched in sloWNet which are clearly very noisy (see Table 2). The most important source of such errors was the inadequate sense assignment for the most frequent words in the language, such as the verb “to be”, the noun “person”, the adjective “big” and the adverb “very”, and will have to be corrected in the future.

While Princeton WordNet contains glosses for all its 117,658 synsets, sloWNet currently contains only 3,178 definitions for nominal synsets that were extracted automatically from Wikipedia articles. 32,881 PWN synsets are also equipped with at least one usage example which is only the case for the 517 sloWNet nominal synsets that were annotated in the corpus. A focused attempt to providing additional definition and example sentences is planned in the near future.

Domains, on the other hand, are much better represented in sloWNet. 46% of all the synsets in PWN that belong to one of the domains exist in sloWNet as well. Of all 161 domains that are present in PWN, only 4 of them are missing entirely, all of them belonging to the Sports domain hierarchy: Rugby, Soccer, Sub and Volleyball, which is a minor issue since there are only 9 synsets in PW that belong to these four domains. Just like in PWN, the most frequent domain is Factotum and the following three most frequent ones are represented in the same order in both wordnets. There are also many similarities among the ten most frequent domains in the two wordnets (see Table 3).

longest synsets		
POS	PWN 3.0	sloWNet 3.0
Adj	23 (02074929-a)	23 (00148078-a)
Adv	10 (00048739-b)	14 (00004722-b)
N	28 (05559256-n)	20 (05921123-n)
V	25 (01426397-v)	24 (00933821-v)
most polysemous literals		
POS	PWN 3.0	sloWNet 3.0
Adj	27 (heavy)	47 (velik~big)
Adv	13 (well)	13 (zelo~very)
N	33 (head)	70 (oseba~person)
V	59 (break)	757 (biti~to be)

Table 2: A comparison of longest synsets and most polysemous literals in PWN 3.0 and sloWNet 3.0

PWN 3.0	Synsets	sloWNet 3.0	Synsets
Factotum	19,454	Factotum	9,701
Zoology	6,270	Zoology	3,345
Botany	5,998	Botany	2,716
Biology	3,004	Biology	1,512
Gastronomy	2,183	Person	793
Chemistry	2,011	Admin.	790
Medicine	1,999	Chemistry	656
Admin.	1,909	Medicine	625
Anatomy	1,768	Building_ind.	575
Person	1,600	Gastronomy	525
Total	77,701	total	33,126

Table 3: A comparison of synsets belonging to domains in PWN 3.0 and sloWNet 3.0

3. Analysis of existing wordnet tools

Several wordnet tools had already been developed, best known among them being the Princeton WordNet Browser (Fellbaum, 1998), Polaris (Louw, 1997) and Periscope (Cuypers and Adriaens, 1997) for the EuroWordNet, DEBVisDic (Horak, 2006) for BalkaNet, WordNet Editor (Derwojedowa et al., 2008) for Polish and WNBrowser (Tufis, 2008) for English and Romanian wordnets.

Because so many tools already existed, it was our goal was to find the one that would best fit our needs and use it. However, our analysis has shown that it is very hard to find a tool which would enable browsing, editing and visualization all in one, and because it is far from trivial to integrate several tools that were developed for different purposes and with different technologies, specialized tools that offer just one of the desired functionalities were discarded (e.g. PWN Browser).

Also, most tools we analysed are not available under an open-source licence (e.g. Polaris, Periscope) and can therefore not be used in the sloWNet project which is based on the open source initiative. We also had to discard the tools that are platform-dependent and are meant for off-line browsing on desktops (e.g. PWN Browser) because they did not meet the requirements of the sloWNet project as such a limitation significantly undermines the usability of the lexico-semantic resource we are developing. Another technical shortcoming we observed is that it is common for wordnet tools to rely on unstandardized, in-house data formats that make it hard to import third-party lexico-semantic resources such as our wordnet (e.g. WordNet Editor, WNBrowser).

Another serious limitation of the available wordnet tools is that a number of them were developed for use in a monolingual setting and are as such unsuitable for bi- or multilingual scenarios (e.g. PWN Browser). Since the development of sloWNet is based a foreign resource, a cross-lingual comparison of concepts is without any doubt a must-have feature. When comparing options for editing wordnet entries it turned out that they are not present in many wordnet browsers at all (e.g. DEBVisDic), and when they are available, they often require installations of client software or do not support creating accounts directly by users, which makes collaborative work on wordnets difficult (e.g. DEBVisDic). This is a very important feature for the sloWNet project because we wish to use crowdsourcing techniques to validate automatically generated synsets.

Finally, when comparing applications for visualizing semantically related words in wordnet, many use Flash or Java technologies that do not perform well in older browsers and with a slower internet connection. A common problem with these applications is also that they produce overcrowded graphs which are not very informative. Similarly, some applications output a static graph for each query that cannot be further explored (e.g. WNBrowser).

Since the beginning of sloWNet development, we have relied on DEBVisDic, which is probably the most widely used wordnet editor and browser in the wordnet development community. The main reason for the change is its inconvenient collaborative on-line wordnet editing that does not support automatic registration of editors, is only possible in certain versions of Mozilla FireFox and requires installation of client packages on each computer the editor wishes to do their job, which is very inconvenient and prevents people to contribute to improving wordnet content. DEBVisDic also does not have a visualization functionality and does not allow integration of third-party resources.

4. Presentation of sloWTool

The all-in-one wordnet tool we developed tries to take all of the above into account. It incorporates browsing, editing and visualization of wordnet content with hyperbolic graphs and images. It is freely available and based on MySQL and PHP technologies, which makes the tool light-weight and portable. It is browser-independent and allows quick queries. Scripts for automatic database transformations from and into several standardized formats, such as DEBVisDic XML and LMF, are provided so that a wordnet for another language can be imported at any time. The on-line browser is simple to use for non-experts but also enables advanced searching and view settings for expert users that can enter complex search queries and decide which fields to display as well as toggle between a mono- and a multilingual option.

4.1 Technical specifications

The sloWTool is a web server application written in PHP scripting language. The tool is using the CodeIgniter open source web application framework for better transparency and maintainability of the source code. The CodeIgniter is based on the model-view-controller (MVC) development pattern. MVC is a software approach that separates an application's logic from its presentation. In practice, it permits your web pages to contain minimal scripting since the presentation is separate from the PHP scripting[1]. The web application data is stored in 12 tables in open source MySQL database which takes approximately 100MB of hard drive. Both technologies, PHP and MySQL, are freely available and can be installed on computers with different operating systems (Linux, Mac OS, and Windows).

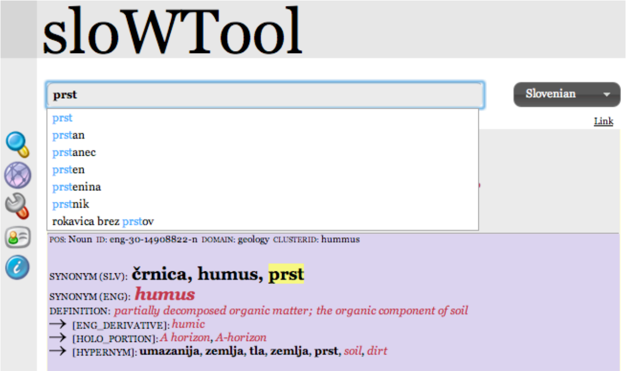
On the client side, in the web browser, a lot of functionality has been written in JavaScript, a scripting language for browsers. Because the client side is quite JavaScript-intensive we are using a quite few add-ons to help us cope with it. For easier HTML traversing, event handling, animation and asynchronous JavaScript and XML (AJAX) performance, we are using jQuery library. The second important add-on is the visualization plugin Springy[2] which we use to draw force-directed graphs on the HTML canvas because it was our initial desiderata not to use Flash for animating the graphs, only the HTML 5 elements. In addition, we use a window plugin mbContainerPlus[3] which helps us draw nice, movable and resizable widows for customizing the page layout.

Because of intensive use of JavaScript in the web browser we created a fluid web application which works fast and without unwanted page refreshes and interrupts. In addition, the client side of the application is using only browser capabilities for displaying the content of the page, enabling the application to work on all modern browsers that includes computers, tablets and even mobile phones with HTML 5-capable browsers.

4.1 Browsing features

The most basic feature of sloWTool is the wordnet browser which is available in simple and in advanced mode. In the simple mode the user can either display the results for a random word or search for a particular word in the desired language. When the search query is entered in the search field, a list of all synsets containing that word is displayed, including multi-word expressions, so that the user can quickly select the word or phrase they wish to see in more detail.

An example of partial search results for the word "prst" (Eng. "soil" or "finger") are shown in Figure 1. All the instances of the searched word are highlighted. Each sense of the searched word is displayed as a separate entry (synset) with language-independent information such as part of speech, synset ID and domain information shown at the top and synset edit stamp at the bottom of the entry. In the main part of the entry all the language-dependent information is provided in all the selected languages, each appearing in different colour for easier reading. In the example below the results are displayed for Slovene (black) and English (red). The most important part of the entry is the Synonyms field which shows all the words that lexicalize the concept in question (literals). In addition, each entry contains a short Definition, currently available only in English for most synsets. Some synsets also have a Usage example where the literals are used in context. Finally, all the semantic relations for that synset are displayed. In order to examine the semantic network for the searched word, it is possible to follow the related synsets and expand them into a tree.



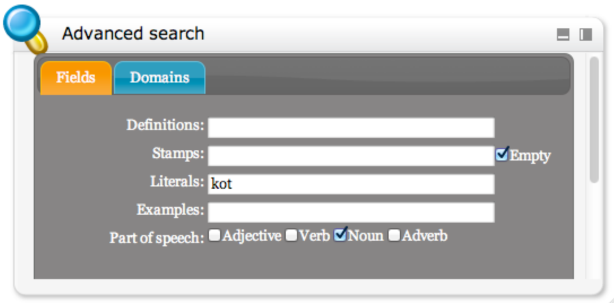
prst
prstan
prstanec
prsten
prstenina
prstnik
rokavica brez prstov

POS: Noun ID: eng-30-14908822-n DOMAIN: geology CLUSTERID: hummus

SYNONYM (SLV): črnica, humus, prst
SYNONYM (ENG): humus
DEFINITION: partially decomposed organic matter; the organic component of soil
→ [ENG_DERIVATIVE]: humic
→ [HOLO_PORTION]: A horizon, A-horizon
→ [HYPERNYM]: umazanija, zemlja, tla, zemlja, prst, soil, dirt

Figure 1: An example of a Slovene synset in sloWTool

More complex search queries can be entered in the advance search window, where the user can use a combination of conditions in several fields. The example of an advanced search query in Figure 2 will find all the nominal synsets in sloWNet that contain the literal "kot" and have not been manually checked. Searches can also be performed over Definitions, Usage examples and Domains. The standard wilcards can be used as well: * for any number of any character and ? for any one character. The results of the search query and dumps of the entire database can be exported in DEBVisDic XML, LMF and tabular formats.



Definitions:
Stamps: ☑Empty
Literals: kot
Examples:
Part of speech: □Adjective □Verb ☑Noun □Adverb

Figure 2: An example of an advanced search query

4.2 Editing features

sloWNet has been developed automatically, which is why synsets need to be manually validated in order to eliminate the noise. We have therefore developed a wordnet editor that is integrated in the browser. We have envisioned two scenarios for editing wordnet content: by

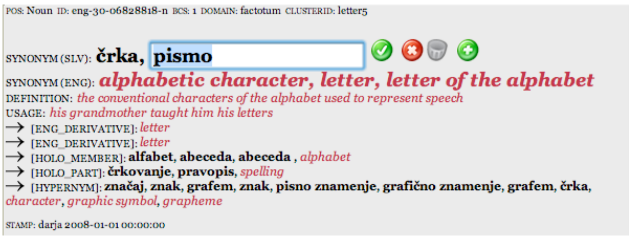
[1] http://codeigniter.com/user_guide/overview/mvc.html
[2] https://github.com/dhotson/springy
[3] http://pupunzi.open-lab.com/mb-jquery-components/mb-containerplus

random visitors of the sloWNet website who spot a mistake in a synset and are willing to correct it immediately, and by a team of lexicographers who perform systematic validation of the developed wordnet.

The first group of users are not willing to invest a lot of effort into the registration process, which is why we enable anonymous editing which does not require a login, making the editing quick and simple. However, anonymous users can only edit literals in synsets, while all the other fields are locked. Also, in order to prevent misuse of the editing option, Captcha tests appear after the maximum number of edits in one session has been exceeded. The changes to synsets that have been suggested by anonymous users are flagged for approval by a database editor, and are only then recorded as such in the database.

The second group of users are lexicographers who log in with a username and password and can edit an unlimited number of synsets, adding changes to all the fields in sloWNet. Because lexicographers are usually carefully selected, approval of the changes they suggest is not required either. Users can edit wordnet content by editing the text in the field (e.g. correcting a mistake in the tedinition), deleting a literal from a synset it does not belong to or by adding a missing literal to an existing synset.

Figure 3 contains an example of a synset which contains an inappropriate literal "pismo" (Eng. letter) for the concept of "alphabetic character" that can be deleted by clicking the Trash button.



POS: Noun ID: eng-30-06828818-n BCS: 1 DOMAIN: factotum CLUSTERID: letter5

SYNONYM (SLV): črka, pismo

SYNONYM (ENG): *alphabetic character, letter, letter of the alphabet*

DEFINITION: *the conventional characters of the alphabet used to represent speech*

USAGE: *his grandmother taught him his letters*

→ [ENG_DERIVATIVE]: *letter*

→ [ENG_DERIVATIVE]: *letter*

→ [HOLO_MEMBER]: alfabet, abeceda, abeceda , *alphabet*

→ [HOLO_PART]: črkovanje, pravopis, *spelling*

→ [HYPERNYM]: značaj, znak, grafem, znak, pisno znamenje, grafično znamenje, grafem, črka, *character*, *graphic symbol*, *grapheme*

STAMP: darja 2008-01-01 00:00:00

Figure 3: An example of synset editing in sloWTool

4.3 Visualization features

The results of a query are visualized in the visualization window that is displayed next to the results of a search query and can be moved and resized, so that the user can directly compare the dictionary view with the graph view. sloWTool visualizer displays all the synsets containing the searched word as well as their first-order relations. Nodes that share any first- or second- order relations are grouped into a cluster. Individual nodes can be dragged closer together or further apart in order to adjust the graph as desired. Figure 4 contains the results of wordnet visualization for the literal "prst". The search query is displayed in the center of the graph and the blue arrows lead to all its senses in wordnet. Additional information about the meaning of the displayed nodes is provided by following the red arrows to the second level of nodes that are semantically related to the original synsets, this displaying a portion of the wordnet's semantic network. Currently, nodes contain Slovene as well as English literals that belong to the same synset but it will be able to limit the display option to a single language in the future when Slovene wordnet gains in size.

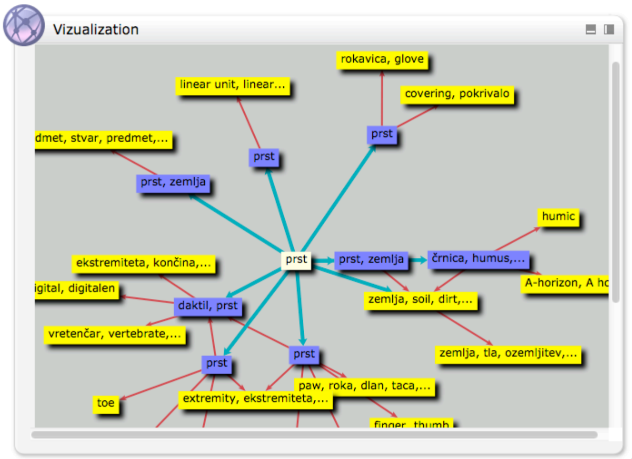


Figure 4: Visualization of wordnet content in sloWTool

4.4 External resources

Apart from developing the browser, editor and visualizer, we have also integrated several external resources into it, which make the tool even more useful. First, in order to enable a comparison between the lexico-semantic inventory in wordnet with actual word usage in context we have integrated the semantically annotated corpus (Fišer and Erjavec, 2010) in the sloWTool that displays the particular senses of the annotated nouns as they are used in context. So far about 5,000 corpus occurrences of 100 most frequent words in the jos100k corpus (Erjavec et al., 2010) have been annotated with approximately 500 different senses. In the future we plan to extend this feature into a platform for annotating all the words in the corpus with wordnet senses.

Second, in addition to sloWNet, we have imported wordnets for English and French in order to be able to compare the lexicalizations of concepts across languages. Plans for incorporating wordnets for other languages are underway.

Next, we have included the WordNet Domains Hierarchy (Bentivogli et al., 2004) which enables the users to look for all the concepts in wordnet that belong to a specific domain, such as Book_Keeping, to its more general parent domain Economy, or to the even more basic domain of Social_Science.

In order to provide a more coarse-grained sense inventory that is sufficient for most users' needs we have grouped the wordnet into meaningful clusters of word senses by mapping wordnet senses to the sense hierarchies of the Oxford Dictionary of English (Navigli, 2006). For example, instead of having to choose between 8 senses of the English word “spirit” in Princeton WordNet, we can use the 3 groups of senses for this word:

- character (2 synsets)
- atmosphere (5 synsets)
- animating force (1 synset)

The clustering is automatic and therefore not without mistakes. Furthermore, it was performed on PWN 2.1, which is why we mapped the clusters to PWN 3.0 that is not perfect either. And, last but not least, clustering was performed on literals, not synsets, which are language-specific and could not be transferred to Slovene as such. This why we conducted an additional sense-oriented grouping of these clusters in order to be able to apply it to Slovene wordnet. Nevertheless, w have already successfully employed the coarse-grained clusters for the extraction of translation equivalents of polysemous words from comparable corpora (see Fišer and Ljubešić, submitted).

Finally, we have enhanced the graph-based visualization module for displaying how words in the wordnet are interlinked by linking the wordnets with an extensive image database called ImageNet (Deng et al., 2009). It contains 12,184,113 images that were carefully selected for their quality and were annotated with 17,624 synsets by humans. Since images are linked to wordnet synsets via word ids, they can be used in other languages as well.

4.5 Availability of sloWTool

The tool is available under the Creative Commons licence of the type Attribution – NonCommercial - ShareAlike. This license lets others remix, tweak, and build upon the tool non-commercially, as long as they credit authors. The new creations of the tool must be available under identical terms. The entire licence can be found on the Creative Commons homepage[4].

The sloWTool full source code is available from Launchpad[5]. Launchpad is a hosting page for free open-source projects. It supports source code hosting using the Bazaar version control system, a bug tracker that allows bugs to be tracked in multiple contexts, a system for tracking specifications and new features, a site for localising applications into different languages, and a community support site.

In order to set up sloWTool, the requirements for the server are a computer that can run web server with the PHP scripting language, support such as Apache and the open source MySQL database. The requirement for running the client part is any modern HTML 5-capable web browser.

5. Conclusions and future work

In this paper we gave an overview of the most important tools for viewing and editing of wordnets and pointed out their shortcomings when trying to use them for sloWNet. We then presented an all-in-one tool we developed ourselves that tries to overcome all the obstacles we ran into with other already existing tools. The first problem with some of the well-known wordnet browsers and editors is that they are not freely available for installation outside the institution where it was developed. Another major issue, especially with the older browsers and editors, is that they have been designed as desktop applications meant for off-line use. Due to our project needs we were also dissatisfied with all the tools that do not support work in a multilingual setting, or tools that enable just one of the desired features.

sloWTool tries to overcome all the identified shortcomings of the available tools and provides a light-weight, easily portable and platform-independent application which is also browser-independent wordnet on the client side. It enables importing of new wordnets or wordnet-like databases from the standard formats such as LMF and DebVisDic XML. sloWTool features include simple browsing and advanced search of wordnet content, anonymous as well as systematic editing of synsets, and a graph-based visualization of the semantic network. The tool also allows adding external, third-party resources, such as wordnet domain hierarchy, coarse-grained sense clusters, and a database of images that are linked to wordnet synsets.

In the future we plan to replace the plain-text wordnet definitions with the Princeton semantically annotated glosses[6] and add links to GeoWordNet[7]. Also, we would like to add wordnets for several other languages for multilingual comparison of wordnet content. And last but not least, we are planning to extend sloWTool to allow assigning wordnet senses to words in the josSENSE corpus.

6. References

Bentivogli, L., Forner, P., Magnini, B. & Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, COLING’04, Geneva, Switzerland, August 28, 2004, pp. 101-108.

Buscaldi, D., Rosso, P. (2008). Geo-wordnet: Automatic georeferencing of wordnet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC’2008*, Morocco, Marrakesh.

Cuypers, I., Adriaens, G. (1997). *Periscope: the EWN Viewer. EuroWordNet Project LE4003, Deliverable D008d012*. University of Amsterdam, Amsterdam.

[4] http://creativecommons.org/licenses/by-nc-sa/3.0/

[5] https://launchpad.net/slowtool

[6] http://wordnet.princeton.edu/glosstag.shtml

[7] http://geowordnet.semanticmatching.org/

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR).*

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawislawska, M. & Broda, B. (2008). Words, concepts and relations in the construction of Polish WordNet. In *Proceedings of the Global WordNet Conference, GWA'2008* Seged, Hungary.

Erjavec, T., Fišer, D. (2006). Building the Slovene Wordnet: first steps, first problems. In *Proceedings of the 3rd International WordNet Conference*, Jeju Island, Korea.

Erjavec, T., Fišer, D., Krek, S. & Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC'10, Malta, May 17-23.

Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fišer, D. (2009). Laveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. In *Human language technology: challenges of the information society*, (LNCS 5603). Berlin; Heidelberg: Springer, pp. 359-368.

Fišer, D., Erjavec, T. (2010). sloWNet: Construction and Corpus Annotation. In *Proceedings of 5th International Conference of the Global WordNet Association*, Mumbay, India.

Fišer, D., Ljubešić, N. (submitted). Addressing polysemy in automatic bilingual lexicon extraction from comparable corpora.

Fišer, D., Sagot, B. (2008). Combining multiple resources to build reliable wordnets. In *Proceedings of TSD'08*, Brno, Czech Republic.

Horak, A., Pala, K., Rambousek, A. & Povolni, M. (2006). DEBVisDic: First Version of New Client-Server Wordnet Browsing and Editing Tool. In *Proceedings of the 3rd International WordNet Conference*, GWA'2006, Jeju Island, South Korea.

Joffe, D., De Schryver, G.M. (2004). TshwaneLex-A State-of-the-Art Dictionary Compilation Program. In *Proceedings of the Eleventh EURALEX International Congress, EURALEX'2004*, Lorient, France.

Louw, M. (1997). *The Polaris User Manual, Internal Report*, Lermout & Hauspie.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of COLING-ACL'06*, Sydney, Australia, July 17-21, 2006.

Noy, N.F., Crubézy, M., Fergerson, R.W., Knublauch, H., Tu, S.W., Vendetti, J. & Musen, M.A. (2003). Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Sagot, B., Fišer, D. (forthcoming). Extending wordnets by learning from multiple resources. In *Proceedings of the LTC'2011 Conference*, Poznan, Poland.

Tufis, D., Ion, R., Bozianu, L., Ceausu, A. & Stefanescu, D. (2008). Romanian WordNet: Current State, New Applications and Prospects. In *Proceedings of the 4th Global WordNet Conference*, GWC'2008, Szeged, Hungary.

The microstructure of Online Linguistics Dictionaries: obligatory and facultative elements

Carolina Flinz
University of Pisa
Via S. Maria 36, 56100 Pisa
E-mail: c.flinz@ec.unipi.it

Abstract

The planning of a dictionary should consider both theoretical and empiric aspects, either for its macro- and microstructure: this is true also for Online Specialized Dictionaries of Linguistics. In particular the microstructure should be standardized and structured so as to fit with the primary and secondary functions of a dictionary. Unfortunately, empirical studies that investigate Online Specialized Dictionaries of Linguistics are rare, making it unclear which microstructural elements are obligatory and which are facultative. This article will present and comment upon the results of an investigation into a corpus of Online Specialized Dictionaries of Linguistics, focusing attention on these aspects and also the most important theoretical issues. An example taken from DIL, a German-Italian Online Dictionary of Linguistics, will end the article.

Keywords: online dictionaries; LSP dictionaries; online dictionaries of linguistics; microstructure bilingual dictionaries of linguistics

1. Introduction

Language Specific Dictionaries are both central to research and core components of basic literature. However, they have only recently become objects of scientific attention: prior to the end of the twentieth century, few lexicographers focused their attention on specialized register[1] and its principal characteristics[2] (cf. Schaeder & Bergenholtz, 1994).

Dictionaries of linguistics are also LSP dictionaries. As such, they have only been investigated since the beginning of the twenty-first century[3] - despite their position as important and relevant tools for the scientific community (in both their printed and online forms). Still today, empirical analyses remain almost non-existent, meaning that no guidelines are present for future LSP lexicographers with an interest in this field.

A particular need also exists for bilingual dictionaries of linguistics in Italy (and particularly German/Italian versions). This is due to the 1999 Italian program of university reforms which separated language studies from literature studies – in turn giving a new role to foreign language linguistics.

This article aims to present the results of an analysis carried out on the microstructure of existing online dictionaries of linguistics. The analysis was originally carried out with the intention of investigating these dictionaries' principal microstructural characteristics[4] - in the hope of providing a set of guidelines for future LSP lexicographers.

Online dictionaries of linguistics are works in their own right, and so should not be seen as simply web-based versions of printed works. Instead, they have their own rights (Barz, 2005), even if some terminology overlaps with printed versions; online dictionaries can be terminology banks, language learning environments (Störrer, 1998), and working and discussion platforms (Abel, 2006). The limits between dictionary, archive, grammar and databank are not strict, and the extensive use of the terms "glossary" and "lexicon" is very common (Flinz, 2010:71).

The microstructure of electronic dictionaries of linguistics – and especially online versions – has only been investigated since the beginning of the twenty-first century. A brief overview of the current state of art will be undertaken in Section 2.

The corpus of the analyses will then be presented in Section 3, and the commented results and some specific examples will end the paper in Section 4. Obligatory and facultative elements will also be focused upon.

2. State of the art: microstructure

The microstructure of dictionaries has long been a point of discussion in lexicography. Many authors have tried

[1] „Was ist eigentlich Fachlexikographie?" (Wiegand, 1988) is considered to be one of the first articles written about this theme. Still in the 80s many lexicographers considered this field "unorganized" and used terms like "vegetating state" (Kucera, 1984), "Wildwuchsgebiet" (Wiegand, 1988) even if there was a "shifting towards the specialized register" (Pilegaard, 1994). The most important publications are from the 90s (Dressler & Schaeder, 1994; Schaeder & Bergenholtz, 1994; Bergenholtz & Tarp, 1995; Hoffmann et al., 1998; Hoffmann et al. 1999 etc.)

[2] Even if the total amount of published LSP dictionaries is very high (Cf. Dressler, 1994), the ones published for a single discipline are very few (excepts some fields like medicine etc.).

[3] Cf. Adamzik, 2001; Lorenzi, 2002; Kreuder, 2003; Flinz, 2010.

[4] The results have been used to plan the microstructure of a German-Italian online dictionary linguistics – a project at the University of Pisa. The results of the analyses concerning the macrostructure were published in elex 2009.

to give the topic a precise definition:
1) Rey-Debove defines it as the "ensemble des informations ordonnées de claque article […] à la suite de l'entrée" (Rey-Debove 1971:13);
2) Hausmann % Wiegand (1989:328) consider it as "the structure of information within the article".

Dictionary microstructures are made up of several different elements that offer information regarding entries' semantics and form (Wiegand, 1996). This composition permits a range of possibilities (cf. Zöfgen, 1994:108):

"Lemma:
a) Aussprache, grammatische Angaben (Artikel, Genus, Pluralbildung, Deklination, Konjugation …), Markierung;
b) (Polysemieangaben: 1°, 2°,…); Bedeutungsparaphrase, Syntagmatik, Valenzangaben, Kollokationen, Beispiele, Paradigmatik, Synonyme, Antonyme, Begriffsfelder, (Homonymie,…);
c) Phaseologische Angaben / Idiomatik." [5]

Of these possibilities, Wiegand distinguishes the three most common variations:
1) „integrierte Mikrostruktur", which gives all of the information in single polisemic entries;
2) „gemisch-integrierte Mikrostruktur", which contains each of the polisemic variations in each entry, with each meaning being followed by its own information;
3) „gemisch-semiintegrierte Mikrostruktur" (Wiegand, 1996), which has both integrated and unintegrated parts, and is generally used for larger articles.

The number and type of information that should follow each entry is determined by the dictionary's type and function. Theoretical studies include both formal and encyclopedic information in LSP dictionaries, as: 1) this information is usually missing in language dictionaries that focus on general language; and 2) this information can help users to better understand lexical lacunae and partial correspondence of meaning.

Standardizing the microstructure is also crucial, as doing so:
1) ensures ease of comprehension;
2) reduced the time required to find information;
3) creates a homogenous style throughout the dictionary;
4) limits individual entries' wording;
5) simplifies the dictionary's readability.[6]

Taken together, these points improve user interaction, causing the "Lesewörterbuch" to be positively judged (Kühn, 1998). Further efforts in this regard include:

1) limited use of links, as too many can cause readers to feel lost;
2) careful use of LSP words, which would not be understandable to the average reader;
3) limited use of abbreviations, saving the reader from regularly having to check up on meanings.

3. Corpus

The corpus is made up of 27 linguistics dictionaries that were found with the help of search engines such as Google, AltaVista and Lycos.

These dictionaries were then categorized in the following manner:
1) monolingual (11 in English, 6 in German, 0 in Italian);
2) bilingual (6 with English as L1 and German, French, Spanish as L2; 1 with German as L1 and English; 1 with Russian as L1 and German as L2; 0 with Italian)
3) plurilingual (2 with Italian as L1).[7]

The dictionaries' microstructure was analyzed according to:
1) article header: typographic relevance of the entry; number and type of information (phonetic, grammar, domain);
2) equivalent or equivalents: languages; direction; presence of grammatical information about the equivalent;
3) definition: use of paraphrasing or citations; presence of links; presence of specific language; presence of abbreviations;
4) syntagmatic elements: syntax information; collocations; examples;
5) paradigmatic elements: synonyms; indication of words belonging to the same semantic field;
6) bibliographical information.

4. Results

The dictionaries were analyzed according to the above-cited categories. The following general considerations were also taken into account:

1) three of the dictionaries were only very simple glossaries – giving equivalents in the foreign language, but no indication of synonymy and/or related terms.
Similarly, they fail to typographically mark and define each entry, don't include bibliographical information, and list every polisemic meaning on its own. They are also organized in a very simple manner, without the typical elements found in online dictionaries (search engine, links etc.);
2) almost all dictionaries had English as L1: only in one case was German used (in a bilingual dictionary) and

[5] In English: „Entry: a) pronunciation, grammatical informations (article, genre, plural, declension, coiniugation…), marks; b) Polysemy (1°, 2°,…); paraphrase of meaning, syntagmatic informations, valency informations, collocations, examples, paradigmatic informations, synonyms, antonyms, conceptual fields (homonymity); c) phraseological informations /idiomatic".
[6] Cf. Wiegand, 1989; Adamzik, 2001.
[7] In this abstract, only the results pertaining to bilingual dictionaries will be presented.

Italian (in a plurilingual dictionary);
3) there was great variation between L2 in the bilingual dictionaries: German, Spain, French, Chinese and Russian were all found. It therefore wasn't possible to compare two dictionaries with the same language couple, meaning that only one dictionary for each language pair has been included;
4) 50% included one entry for each polisemic variation of each term, and 50% had one entry with the various meanings of the lemma (usually marked with the help of letters or numbers).

4.1. Article Header

83% of the analyzed dictionaries had a typographically marked entry in black and bold, although one also included entries in red. In another dictionary, entries are marked with a larger font-size.

There were generally two pieces of information contained in each article's header – most frequently word class and informations about synonymity.

More detailed results are given in the following diagram:

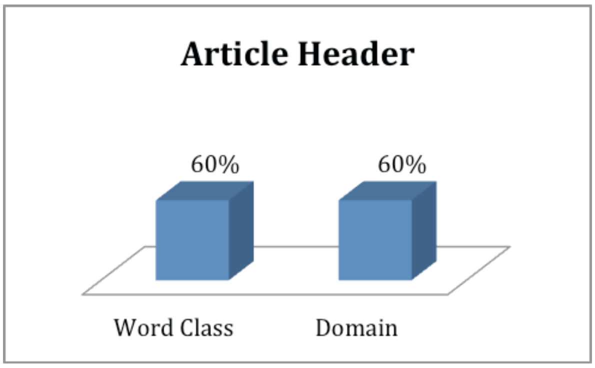


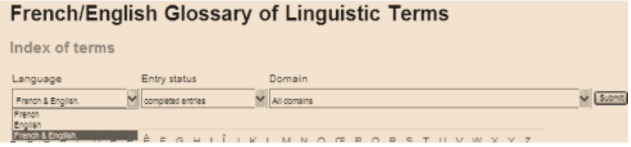
Figure 1: Results from the analyses of the article header

60% of the analyzed dictionaries give information about word class (noun, verb, adjective etc.) and specific domain (lexicography, applied linguistics etc.).

4.2. Equivalents

Equivalents were the most typical elements found in bilingual dictionaries. The most common source language was English, with German and Italian being found as the source language only once.

Bidirectional dictionaries are generally considered rare in theoretical papers, but the results of this analysis do not confirm this fact: 66% of the analyzed dictionaries were bidirectional, even if in some cases this was not total bidirectionality. Both products from the Summer Institute of Linguistics are good references for this type of dictionary: the user of the "French/English Glossary of Linguistic Terms", for example, can chose on the first page between entries that are French, English or bidirectional.


French/English Glossary of Linguistic Terms

Index of terms

Language: French & English.
Entry status: completed entries
Domain: All domains

Figure 2: Screenshot from French/English Glossary of Linguistic Terms (www.sil.org)

Lexicographers also usually consider the inclusion of grammatical information to be common, but this was only found rarely in the dictionaries analyzed (8%).

4.3. Definition

Definition is an important and relevant element in encyclopedic dictionaries, and also in LSP dictionaries (Rossenbeck 1987: 278f; Duvå–Laursen 1994:247f; Schaeder & Bergenholtz, 1994:141). The analyses showed that 83% of the dictionaries were equipped with definitions.

The definition is considered "the epicenter of the microstructure" (Schaeder & Bergenholtz, 1994:225). As such, the primary consideration should be that "the preciseness, the scope and currency of the explanation" should be coupled with "an up-to-date comment on their specific usage" (Schaeder& Bergenholtz, 1994:219). Lexicographers should always have in mind who the user is and what his needs are Kühn (1998), Wiegand (2002) und Zöfgen (1994).

Different possibilities exist for drafting definitions:
1) use of the Aristotle principle of *genus proximum* and *differentia specifica;*
2) information of the intention and the extension of the term;
2) indication about the most important prototypical semantic aspects;
3) evidence about the principle related concepts ("frame concepts").

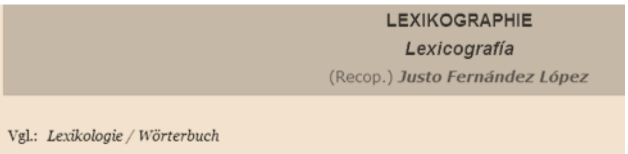
Similarly, various techniques can be used in the writing process:
1) paraphrasing;
2) including synonyms;
3) contrastive analyses;
4) including citations.

The results of the analyses showed that 60% of the dictionaries used paraphrasing in their explanation, and 40% included citations. The Spanish-German dictionary from Hispanoteca and the DLM Project gave citations in both relevant languages or more.

Each definition should present different perspectives on terms, evidenced by the use of paragraphs. (Adamzik, 2001:220). However, the use of paragraphs in the analyzed dictionaries is not common: the only exception is the Spanish-German dictionary from Hispanoteca.
A key concept in lexicography is that words should be

defined using words simpler than themselves (Götz, 1984: 50) – but 100% of the analyzed dictionaries used LSP words in their definitions, even if the proportion was limited to a few words (3-5) per entry. Abbreviations are rare (despite being a typical element of printed linguistics dictionaries).

Links and references to other related terms are very common, being present in all dictionaries with the exception of the Spanish-German dictionary from Hispoanoteca. This exception seems to be a transposition of a written dictionary, having hyperlinks neither in the text nor at its end. Instead, it uses the method commonly found in printed dictionaries:



LEXIKOGRAPHIE
Lexicografía
(Recop.) *Justo Fernández López*

Vgl.: *Lexikologie / Wörterbuch*

Figure 3: Screenshot from Lexikon der Linguistik / Diccionario de Lingüística (www.hispanoteca.eu)

4.4. Sintagmatic Information

Sintagmatic information is very important in some types of dictionary (learner's dictionaries, language dictionaries and LSP dictionaries), because it shows how an entry can be used and what types of words are accepted (Zöfgen, 1994:147). Theories distinguish between:
1) syntactical information, which focuses on the correct use of a term;
2) collocations: the union of two or more words in a sentence. These are also welcomed in bilingual dictionaries;
3) examples, which usually show what type of constructions can be built with the lemma. Examples are also used to show concepts or give particular information.

The results of these points are demonstrated in the following diagram:

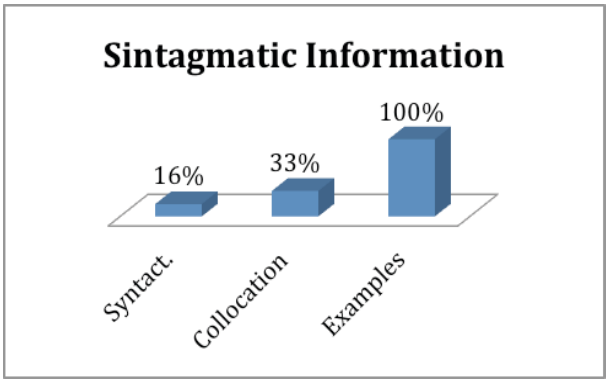
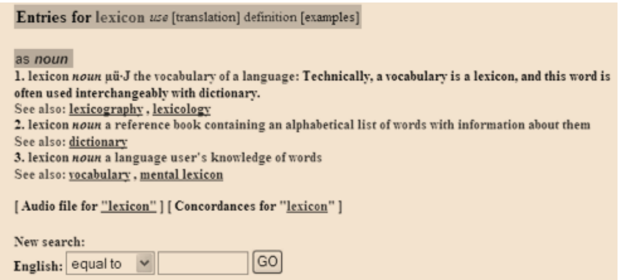


Figure 4: Results from the analyses of the sintagmatic information

Syntactical information is rare, only being used in one dictionary, and even then including little detail. Similarly, information about collocations is not used much in this type of dictionary, with the only exception being the Linguistics Glossary. At the end of each entry here is found the indication: (Concordances for ….):



Entries for lexicon *use* [translation] definition [examples]

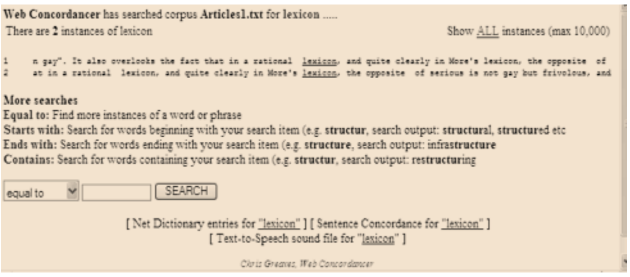
as *noun*
1. lexicon *noun* µü·J the vocabulary of a language: **Technically, a vocabulary is a lexicon, and this word is often used interchangeably with dictionary.**
See also: lexicography , lexicology
2. lexicon *noun* a reference book containing an alphabetical list of words with information about them
See also: dictionary
3. lexicon *noun* a language user's knowledge of words
See also: vocabulary , mental lexicon

[**Audio file for** "lexicon"] [**Concordances for** "lexicon"]

New search:
English: equal to GO

Figure 5: Screenshot from Linguistics Glossary (www.edict.biz/lexiconindex/linguistics)

On opening the link to Concordances for Lexicon, more search options are presented:



Web Concordancer has searched corpus **Articles1.txt** for **lexicon** .....
There are **2** instances of lexicon Show ALL instances (max 10,000)

```
1     n gay". It also overlooks the fact that in a rational  lexicon, and quite clearly in More's lexicon, the opposite  of
2     at in a rational  lexicon, and quite clearly in More's lexicon, the opposite  of serious is not gay but frivolous, and
```

More searches
Equal to: Find more instances of a word or phrase
Starts with: Search for words beginning with your search item (e.g. **structur**, search output: **structural**, **structured** etc
Ends with: Search for words ending with your search item (e.g. **structure**, search output: **infrastructure**
Contains: Search for words containing your search item (e.g. **structur**, search output: **restructuring**

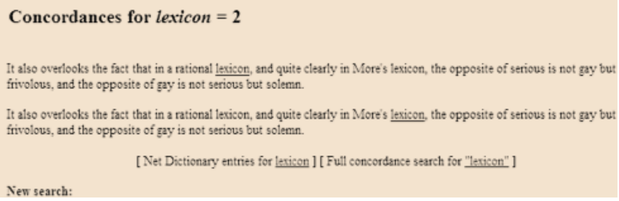
equal to SEARCH

[Net Dictionary entries for "lexicon"] [Sentence Concordance for "lexicon"]
[Text-to-Speech sound file for "lexicon"]

Chris Greaves, Web Concordancer

Figure 6: Screenshot from Linguistics Glossary (www.edict.biz/lexiconindex/linguistics)

An option also exists for searching sentence concordance:



Concordances for *lexicon* = 2

It also overlooks the fact that in a rational lexicon, and quite clearly in More's lexicon, the opposite of serious is not gay but frivolous, and the opposite of gay is not serious but solemn.

It also overlooks the fact that in a rational lexicon, and quite clearly in More's lexicon, the opposite of serious is not gay but frivolous, and the opposite of gay is not serious but solemn.

[Net Dictionary entries for lexicon] [Full concordance search for "lexicon"]

New search:

Figure 7: Screenshot from Linguistics Glossary (www.edict.biz/lexiconindex/linguistics)

Examples are present in 100% of the analyzed dictionaries. They can be:
1) examples of use;
2) examples of concepts (e.g. pronunciation in the definition of particular types of vowels: the "Glossary Spanish-English" has audio files that let the user better understand the audio characteristics of specific vowels).

4.5 Paradigmatic Information

Paradigmatic information about entries is welcome in dictionaries (Kühn, 1998). It is separated into: 1) synonyms; 2) antonyms; 3) terms belonging to the same semantic area.

The analyses concluded the following results:

1) synonyms are present in all dictionaries;
2) antonyms are used only in one dictionary;
3) related terms are typical of these dictionaries: 100% introduce related terms in their entries, which are usually included at either the beginning or end of articles and are signaled with "see also", "cfr" etc.

4.6. Bibliographical Information

Bibliographical information is a typical element of German dictionaries of linguistics, where each entry has at its end a complete indication of the source. These are absent in Italian dictionaries, which instead include them at the dictionaries' ends.

The analyses of online dictionaries showed that only 66% put this type of information at the end of the articles. The dictionaries from the Summer Institute of Linguistics put them in abbreviated form, with a link allowing users to open the full information.

Theoretical studies consider the following as key components of linguistic dictionaries: abbreviations, synonyms, information about grammatical aspects and used contexts, antonyms, etymology and examples. The empirical analyses confirmed these considerations only partially, with bilingual linguistics dictionaries being shown to have:

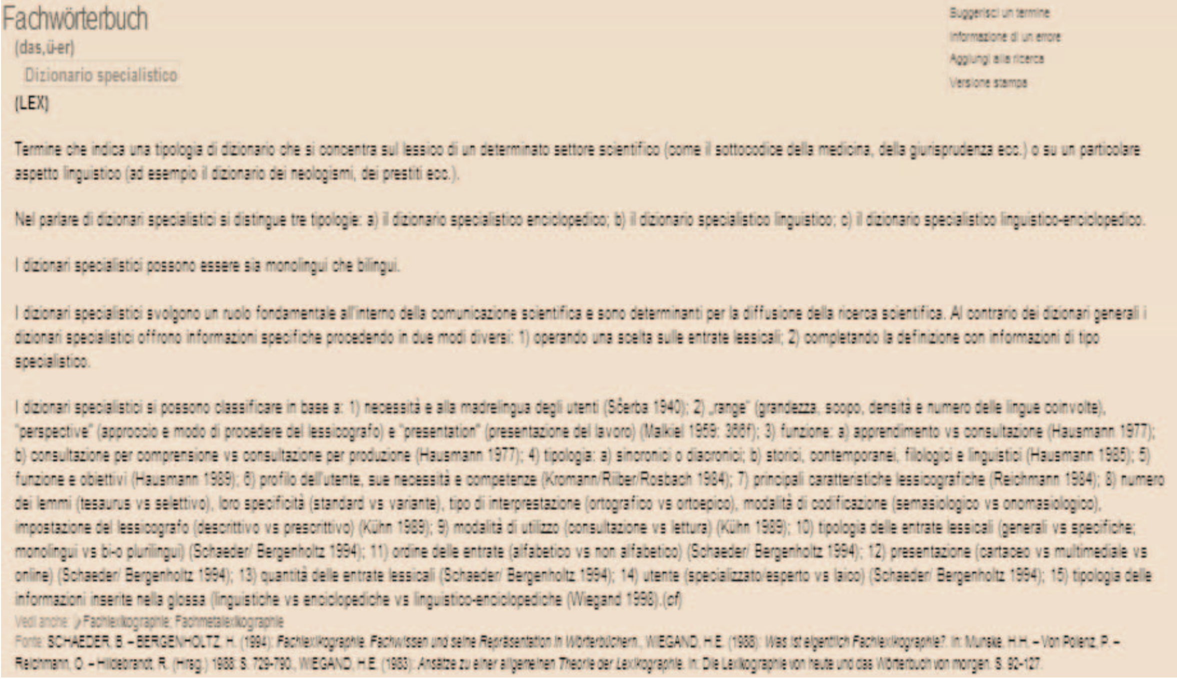
1) lexicographically-marked entries;
2) indication about word class and domain in the article header, but further grammatical information rarely attested even if strongly called for by many researchers;
3) examples of concepts;
4) synonyms and related terms.

The online dictionaries also have a good structured definition: they alternate the use of paraphrasing and citations. Differing perspectives on entries are shown with the help of paragraphs. LSP words, links and abbreviations are also used in the right way, without disturbing the user in his purposes.

These analyses were a great help in the planning of a bilingual German-Italian linguistics dictionary at the University of Pisa (naturally after considering the potential user, his needs, his probable situation of use, and also the type and function of the dictionary).

The microstructure of DIL, a German-Italian online Specialized Dictionary of Linguistics, has tried to follow the above criteria:

Example:



Fachwörterbuch
(das,ü-er)
Dizionario specialistico
(LEX)

Termine che indica una tipologia di dizionario che si concentra sul lessico di un determinato settore scientifico (come il sottocodice della medicina, della giurisprudenza ecc.) o su un particolare aspetto linguistico (ad esempio il dizionario dei neologismi, dei prestiti ecc.).

Nel parlare di dizionari specialistici si distingue tre tipologie: a) il dizionario specialistico enciclopedico; b) il dizionario specialistico linguistico; c) il dizionario specialistico linguistico-enciclopedico.

I dizionari specialistici possono essere sia monolingui che bilingui.

I dizionari specialistici svolgono un ruolo fondamentale all'interno della comunicazione scientifica e sono determinanti per la diffusione della ricerca scientifica. Al contrario dei dizionari generali i dizionari specialistici offrono informazioni specifiche procedendo in due modi diversi: 1) operando una scelta sulle entrate lessicali; 2) completando la definizione con informazioni di tipo specialistico.

I dizionari specialistici si possono classificare in base a: 1) necessità e alla madrelingua degli utenti (Sčerba 1940); 2) „range" (grandezza, scopo, densità e numero delle lingue coinvolte), "perspective" (approccio e modo di procedere del lessicografo) e "presentation" (presentazione del lavoro) (Malkiel 1959: 366f); 3) funzione: a) apprendimento vs consultazione (Hausmann 1977); b) consultazione per comprensione vs consultazione per produzione (Hausmann 1977); 4) tipologia: a) sincronici o diacronici; b) storici, contemporanei, filologici e linguistici (Hausmann 1985); 5) funzione e obiettivi (Hausmann 1989); 6) profilo dell'utente, sue necessità e competenze (Kromann/Riiber/Rosbach 1984); 7) principali caratteristiche lessicografiche (Reichmann 1984); 8) numero dei lemmi (tesaurus vs selettivo), loro specificità (standard vs variante), tipo di interpretazione (ortografico vs ortoepico), modalità di codificazione (semasiologico vs onomasiologico), impostazione del lessicografo (descrittivo vs prescrittivo) (Kühn 1989); 9) modalità di utilizzo (consultazione vs lettura) (Kühn 1989); 10) tipologia delle entrate lessicali (generali vs specifiche; monolingui vs bi-o plurilingui) (Schaeder/ Bergenholtz 1994); 11) ordine delle entrate (alfabetico vs non alfabetico) (Schaeder/ Bergenholtz 1994); 12) presentazione (cartaceo vs multimediale vs online) (Schaeder/ Bergenholtz 1994); 13) quantità delle entrate lessicali (Schaeder/ Bergenholtz 1994); 14) utente (specializzato/esperto vs laico) (Schaeder/ Bergenholtz 1994); 15) tipologia delle informazioni inserite nella glossa (linguistiche vs enciclopediche vs linguistico-enciclopediche (Wiegand 1998).(cf)

Vedi anche: ⟩ Fachlexikographie; Fachmetalexikographie

Fonte: SCHAEDER, B – BERGENHOLTZ, H. (1994): *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern.*, WIEGAND, H.E. (1988): *Was ist eigentlich Fachlexikographie?*. In: Munske, H.H. – Von Polenz, P. – Reichmann, O. – Hildebrandt, R. (Hrsg.) 1988. S. 729-790., WIEGAND, H.E. (1983): *Ansätze zu einer allgemeinen Theorie der Lexikographie*. In: Die Lexikographie von heute und das Wörterbuch von morgen. S. 92-127.

Figure 8: Screenshot from DIL / Dizionario Tedesco-Italiano di terminologia linguistica)
(www.humnet.unipi.it/dott_linggensac/glossword)

5. References

Abel, A. (2006). Elektronische Wörterbücher: Neue Wege und Tendenzen. In F. San Vicente (ed.) *Akten der Tagung „Lessicografia bilingue e Traduzione: metodi, strumenti e approcci attuali (Forlì, 17.-18.11.2005)"*, Open Access Publications: Polimetrica Publisher, pp. 35-56.

Adamzik, K. (2001). *Kontrastive Textologie*, Tübingen: Stauffenburg.

Barz, I., Bergenholtz, H. & Korhonen, J. (2005). *Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*. Frankfurt a.M.: Lang.

Bergenholtz, H., Schaeder, B. (1994). *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern*. Tübingen: Narr.

Bergenholtz, H., Tarp, S. (1995). *Manual of LSP lexikography. Preparation of LSP dictionaries-problems and suggested solutions*, Amsterdam - Philadelphia: John Benjamins.

Dressler, S. (1994). Wörterbücher der Medizin. Eine Bibliographie. In S. Dressler, B. Schaeder (eds.) *Wörterbücher der Medizin. Beiträge zur Fachlexikographie*. Tübingen: Niemeyer, pp. 171-279.

Dressler, S., Schaeder, B. (1994). *Wörterbücher der Medizin. Beiträge zur Fachlexikographie*. Tübingen: Niemeyer. (Lexikographica Series Maior 55).

Duvå, G., Laursen, A.L. (1994). Translation and LSP Lexikography: A User Survey. In H. Bergenholtz,B. Schaeder (eds.) *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern*. Tübingen: Narr, pp. 247-267.

Flinz, C. (2010). DIL- an online bilingual specialized dictionary of linguistics (German-Italian). In S. Granger, M. Paquot (eds.) *eLexicography in the 21st century: New challenges, new applications*. Louvain: UCL, pp. 67-77.

Götz, D., Herbst, T. (1984) (eds.). *Theoretische und praktische Probleme der Lexikographie. Professor Dr. Günther Haensch zum 60. Geburtstag 1. Augsburger Kolloquium,* München: Hueber.

Hausmann, F., Reichmann, O., Wiegand, H.E. & Zgusta, L. (eds.) (1989). *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft(HSK 5.1.) Berlin, New York: De Gruyter.

Hausmann, F., Reichmann, O., Wiegand, H.E. & Zgusta, L. (eds.) (1990). *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK 5.2.) Berlin, New York: De Gruyter.

Kreuder, H.D. (2003). *Metasprachliche Lexikographie. Untersuchungen zur Kodifizierung der linguistischen Terminologie*, Tübingen: Niemeyer.

Kucera, A. (1984). Aus der Werkstatt der praktischen Verlagslexikographie. Übersetzungswörterbücher der Fachsprachen. In: *Mitteilungen für Dolmetscher und Übersetzer* 1/30. pp. 3-6.

Kühn, P. (1998). *Langenscheidts Großwörterbuch Deutsch als Fremdsprache und die deutschen Wörterbücher*. In H.E. Wiegand (ed.) *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von "Langenscheidts Großwörterbuch Deutsch als Fremdsprache"*, Tübingen: Niemeyer, pp. 34-60.

Lorenzi, F. (a cura di) (2002). *DLM - Dizionario generale plurilingue del Lessico Metalinguistico*. Roma: Il Calamo.

Pilegaard, M. (1994). Bilingual LSP Dictionaries. User benefit correlates with elaborateness of „explanation". In H. Bergenholtz, B. Schaeder (eds.) *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern*. Tübingen: Narr, pp. 211-228.

Rey-Debove, J. (1971). *Étude linguistique et sämiotique des dictionnaires français contemporains*. Paris: The Hague.

Rossenbeck, K. (1987). Zur Gestaltung zweisprachiger Fachwörterbücher. In A.M. Cornu, J. Vanparijs, M. Delahaye & L. Baten (eds.) *Beads or Bracelet? How do we approach LSP. Selected Papers from the Fifth European Symposium on LSP* (1985). Oxford: Oxford University Press, pp. 274-283.

Storrer, A., Harriehausen, B. (1998). *Hypermedia für Lexikon und Grammatik*. Tübingen: Narr.

Welker, H.A. (2003). *Zweisprachige Lexikographie: Vorschläge für deutsch-portugiesische Verbwörterbücher*. München: Utz.

Wiegand, H.E. (1988). Was ist eigentlich Fachlexikographie?. In H.H. Munske, P. Von Polenz, O. Reichmann & R. Hildebrandt (eds.) *Deutscher Wortschatz. Lexikologische Studien*. Berlin & New York: De Gruyter, pp. 729-790.

Wiegand, H.E. (1989). Der gegenwärtige Status der Lexikographie und ihr Verhältnis zu anderen Disziplinen. In F.J. Hausmann, O. Reichmann, H.E. Wiegand & L. Zgusta (eds.), *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft *HSK 5.1*, pp. 409-462.

Wiegand, H.E. (1996). *Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographie-Kolloquium*. Tübingen: Niemeyer.

Wiegand, H.E. (ed.) (2002). *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand des „De Gruyter Wörterbuch Deutsch als Fremdsprache"*. Tübingen: Niemeyer.

Zöfgen, E. (1994). *Lernerwörterbuch in Theorie und Praxis. Ein Beitrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen*. Tübingen: Niemeyer.

NLP lexicons: innovative constructions and usages for machines and humans

Nuria Gala[(1)], Mathieu Lafourcade[(2)]
[(1)]LIF-CNRS, 163 av. de Luminy case 901, 13288 Marseille Cedex 9, France
[(2)]LIRMM-CNRS, 161 rue Ada, 34392 Montpellier Cedex 5, France
Email: nuria.gala@lif.univ-mrs.fr, mathieu.lafourcade@lirmm.fr

Abstract

Lexical resources have undergone significant changes with the generalized use of computers and the advent of the Internet. However, while such changes stand for revolutions when it comes to compare machine-readable dictionaries to their paper 'ancestors', machine-readable dictionaries, compiled for human readers, still have serious limitations. Natural language processing lexicons, initially developed for NLP applications, have shed light on some of such shortcomings. In this presentation, we will attempt to bring new elements relatively to NLP approaches aiming to develop present and tomorrow's lexical resources, in particular, using morphological and semantic information to better access lexical items. A special focus will be given on the semantic and on the multilingual side. Our argument is that nowadays lexical resources 1) should be useful both for men and machines, 2) can be constructed in alternative ways from classical lexicographic work, and 3) provide novel accesses and usages that are feasible only in the context of computer and user networks. Such points will be highlighted by means of two resources under development: *LexRom,* as an example of morphological form-based multilingual access, and the lexical network of *JeuxDeMots,* as an illustration of associative and semantic access.

Keywords: NLP lexica; crowd sourcing; semi-supervised learning; morphological and semantic content; multilingual and semantic access

1. Introduction

Since the introduction of computers, lexical resources have undergone significant changes in terms of lexicographical practices and user access to words. For more than thirty years, the growing contribution of computers to lexicography has transformed the way to create and enrich lexical resources[1]. Yet the impact of using machines into the lexicographic field has already been discussed in the literature by leading contributors (Atkins & Zampolli, 1994; Grefenstette, 1998; Rundell, 2002, among others). However, the subject still remains on the table, mainly because the achievements in terms of the resources themselves are far from being as satisfactory as the electronic media could entail.

While lexicographical practices have significantly evolved due to the access to large amounts of data and the use of highly-skilled and linguistically-aware editors (Rundell, 2002), machine-readable dictionaries (MRDs) still stand for electronic versions of their paper 'ancestors'. Doubtless, the use of large amounts of data allowed the incorporation of new information into the dictionaries: statistical – frequencies – (Kilgarriff, 1997), collocational behaviors, and even much complex patterns – syntactic patterns – gathered by corpus query tools (Jakubíček et al., 2010). Furthermore, the electronic media involved the combination of multimedia lexicographic material like sounds (pronunciations), images, videos (sign languages), etc. which might be of help in particular contexts and for specific users: foreign learners (i.e. Merriam Webster[2]), deaf people (i.e. Arasaac[3], Tegnspro[4]), among others. Nevertheless, it should be noted, as Grefenstette (1998) did, that the lexicon represented in the dictionaries is still seen as two-dimensional: "a list of lists" (Heid, 2009), that is to say, a list of words with their associated explanations, be them linguistic, statistical or multimedia.

As far as the access to the lexicographic information is concerned, the major revolution is the fact that the user has a variety of research criteria going from a target key word to more complex search patterns (a specific domain, a grammatical category, etc.). Frequently, s/he may even choose among several possibilities. Yet, in spite of such functionalities, alphabetical lists generally remain, as if the user was unable to get rid of traditional habits.

More recently, online interactive dictionaries appear to be real platforms giving access to several interconnected lexical resources. One example is the *Nuevo Tesoro Lexicográfico de la Lengua Española*[5] (NTLLE), a resource from the Real Academia Española grouping about 70 dictionaries resulting from five centuries of institutional Spanish lexicography. Another example might be *Wordnik*[6], a resource giving access to a variety of lexical resources (dictionaries, corpora, thesauri, etc.) and thus going beyond the user expectations with "as much information as possible" about a word.

[1] The *Trésor de la Langue Française* (1971-1994) innovated French lexicography with the use of computer indexing of a wide corpus of texts (Frantext). The *Collins COBUILD* (1987) was the first dictionary where electronic corpora was used, thus providing primary English data source (7 million word).

[2] http://www.merriam-webster.com

[3] http://www.catedu.es/arasaac

[4] http://tegnsprog.dk

[5] http://buscon.rae.es/ntlle/SrvltGUILoginNtlle

[6] http://www.wordnik.com

Despite such significant progress resulting from computer means (no need to mention increasing storage capabilities, nor reduced response time for a query), MRDs, enriched with hyperlinks but still compiled for human readers, remain two-dimensional repositories and have serious limitations (Fellbaum & Miller 2003), namely on the access to the semantic content otherwise than through words (or their pronunciations / grammatical information) and also on the granularity of the information (they do not include information that they assume the user knows). Additional shortcomings can be raised about the size of the resources in terms of language coverage as well as on the cross-lingual equivalencies.

In this paper, we will attempt to shed light on some of such shortcomings by bringing new elements relatively to NLP approaches aiming to develop present and tomorrow's lexical resources. In particular, a special focus will be given on the semantic and on the multilingual side, using morphological and lexical information to better access lexical objects. The paper is structured as follows. First, MRDs and NLP lexicons are compared. Second, alternatives to human (lexicographic) constructions are dealt with. The following sections are devoted to two lexical resources to illustrate the points already highlighted on the previous sections: accessing to words by their form in a multilingual context (*LexRom*) and through lexical functions (*JeuxDeMots*). The paper concludes by a look at open questions and current developments.

2. Natural Language Processing (NLP) lexicons: from lists to networks

MRDs have been used as a source to collect lexical knowledge for a variety of natural language processing (NLP) applications. Yet considerable research has been done for more than thirty years on automatic extraction of structured knowledge from MRDs: lexical relations, semantic information, taxonomies, etc. However, as Ide & Véronis already pointed out, the results of MRD research for NLP failed to live up to early expectations: "encouraging line of research" (Véronis & Ide, 1990) but "the information they [MRDs] contain is both too inconsistent and incomplete to provide a ready-made source of comprehensive lexical knowledge" (Ide & Véronis, 1994).

At the same time, as they are valuable sources of linguistic information, the NLP community has been actively involved in the creation of a wide range of lexical resources (from computational lexicons –lexical databases– to annotated corpora). While initially created for machine applications, such resources may also be of interest for humans through appropriate interfaces (i.e. language learning, speech therapies, linguistic studies, etc.). NLP researchers have thus been developing computational lexicons leading to significant advances not only on construction methods and techniques (see section 3) but also on the resources themselves: lexical databases are conceived bearing in mind their primary purpose, that is NLP applications, which entails automating –as much as possible– the process of linguistic data analysis with robust technologies. As a result, the information is *structured*, *explicit* and *multi-dimensional*, moving "from lists to networks of lexical objects" (Heid, 2009). To put it in other words, a variety of information is scattered throughout different levels and the user browses to specific contents depending on his/her needs. Lexical resources are thus increasingly *dynamic* as the information is interconnected and available by different means (see sections 4 and 5).

In recent years, a significant number of NLP lexical resources have been developed in a large-scale perspective. Series of projects (EAGLES, MULTEXT, etc.) have converged to standards and models to provide a common framework for their construction, maintenance and extension, i.e. the *Lexical Markup Framework* (LMF) (Francopoulo et al., 2006). These models address linguistic representation and encoding guidelines at different layers (morphology, syntactic behaviors, semantic organization, etc.). Significant projects have come to life, *WordNet* (Fellbaum, 1990) being one of the most outstanding.

3. NLP methods for development, enrichment and evaluation of lexical resources

Due to the nature of language, large-scale lexicon development poses difficult challenges (Calzolari et al., 1999). As manual development is very costly and time consuming, automatic and collaborative building of computational lexicons are real alternatives.

3.1 Automatic acquisition of linguistic knowledge

The cost of manual elaboration and enrichment of resources is generally put forward as a major inconvenience. Within the context of lexical resources and NLP tools development, a response to such uneasiness is the automatic acquisition of linguistic knowledge. Over the last twenty years, a number of unsupervised and (semi-)supervised approaches have become a real alternative yielding to encouraging results at different linguistic layers: morphology (Clément et al., 2004), syntax (Briscoe & Carrol, 1997), semantics (Navigli et al. 2003). The overall idea is to induce linguistic knowledge from available data. Depending on the characteristics of the underlying data (raw or annotated corpora, lexical databases, MRDs, etc.), the target resource would be developed more or less straightforward. In any case, manual evaluation would be necessary, but once again, at different degrees depending on one hand, on the underlying data and, on the other hand, on the aimed granularity of the linguistic description. The more explicit the underlying data, the more explicit the target resource, though more difficult its development.

Due to the availability of large amounts of corpora, statistical models have been playing a major role within the NLP community. Unsupervised techniques do not presuppose explicit linguistic knowledge (annotations): they allow the acquisition of linguistic information from raw corpora. If the results are below other approaches that use annotated data, the major advantage is the availability of unlabeled data (i.e the Web). Very often, such methods are used as a first step for preprocessing raw corpora or to incrementally improve the models .

Semi-supervised approaches exploit some kind of information already encoded or annotated on corpora (i.e. part-of-speech tags). Such methods yield to better results than unsupervised ones because the underlying data allows to induce better linguistic information. In many cases, automatic acquisition of linguistic knowledge for lexical development is based on combining both unsupervised and semi-supervised approaches (see Section 4).

Finally, an alternative to the use of corpora is the use of existing lexical resources. While a number of projects have come to light by using MRDs – with mixed results already mentioned on the previous section – , the use of existing computational lexicons is an interesting option as linguistic knowledge is made explicit. Such line of research is currently widespread, though the resources are not always easily available.

3.2 Collaborative approaches

Collaborative resources, i.e. *Papillon* (Boitet et al., 2002), are based on the principle of sharing contributions, that is, anyone collaborates to enrich the database according to his/her possibilities. The insights of this philosophy are interesting but the results are sometimes disappointing as enriching a resource may become tedious very quickly, and in practice people tend not to participate. Hence, it is hard to get the expected volume of contribution (Cristea et al., 2008).

Over the last decade, the web has led to collaborative projects (*wikis*) based on the participation of volunteers under the supervision of an administrator. Significant projects as regards to lexical semantic resources can be mentionned: *OntoWiki*[7] and *Anawiki*[8] (Poesio et al., 2008) among others. However, if such approaches are appropriate for resources of reasonable size and very good quality (gold standards), they are less suitable for large-scale development (Fort et al., 2010).

One way to avoid such a drawback may be crowd-sourcing through gaming, i.e. games with a purpose (GWAP). In such approaches, volunteers are motivated throughout competition (see Section 5).

Lastly, a new trend has emerged which consist on on-line microworking (a task is cut into small pieces and their execution is paid for). *Mechanical Turk* is one such systems: since its introduction in 2005 it has been increasingly being used for building and validating NLP resources at very low cost (Fort et al., 2010), e.g. transcription, word sense disambiguation, compound relations annotation, categorization, etc. However, a number of drawbacks are being brought to light, namely the small number of trained annotators and thus the annotation quality of the resources produced that way: “if a microworking system is considered desirable by the ACL and ISCA communities, then we also suggest that they explore the creation and use of a linguistically specialized special-purpose microworking alternative to MTurk that both ensures linguistic quality and holds itself to the highest ethical standards of employer/employee relationships” (Fort et al., 2010).

As a first conclusion, lexical resources can be constructed in alternative ways from classical lexicographic work and may be used both for men and machines. Novel accesses and usages may be thus provided, feasible only in the context of computer and user networks. *LexRom* and *JeuxDeMots* appear to be obvious examples.

4. LexRom

LexRom (Gala, 2011) is a project of a multilingual lexicon for Romance languages based on family clusters, providing morphological and semantic information on word families crosslingually. The project aims to be of help in contrastive linguistic research as well as in different NLP and human applications, going from crosslingual information retrieval to interlingual language learning. Spanish and Catalan families have been automatically acquired from corpora and monolingual lexicons, from an initial list of manually encoded words from the French morphological resource *Polymots* (Gala et al., 2010)[9].

To our knowledge, attempts to build multilingual lexical resources have mainly focused on semantic relations between concepts among different languages, i.e. *EuroWordNet* (Vossen, 1998). Other interesting proposals merge lexical and encyclopedic knowledge automatically extracted from WordNet and Wikipedia, i.e. *Babelnet* (Navigli and Ponzetto, 2010). As for morphology, the reference multilingual database is *Celex* (Baayen et al., 1995), yet it has been created as three separated lexicons for English, Dutch and German and thus no interlingual links are available.

4.1 Word-forms and semantic cues

The notion underlying *LexRom* is that of morpho-phonological families. A morpho-phonological family

[7] http://ontowiki.eu/

[8] http://anawiki.essex.ac.uk/

[9] http://polymots.lif.univ-mrs.fr

groups together lexical units sharing phonological, morphological and semantic features. Such a family is usually built around a common stem. For example, in French, the stem 'olive' will induce the family made of lexical units such as 'olivaison' (olive harvesting), 'oliveraie' (olive grove), 'olivier' (olive tree)[10], etc. (see first line on Table 1). For each lexical entry in a family, the following types of information is displayed:

- **morphological structure**: i.e. for 'olivier', base-form *oliv-*, affixes *-i* and *-er* ;
- eventual **phonological alternations**: i.e. 'fleur/flor-' is the stem for words such as 'fleur' (flower), 'fleurir' (bloom) and also 'floraison' (flowering); 'croc/croch-' is the stem for 'croc' (hook) and 'accrocher' (hang);
- **semantic cues**: semantic units associated to the target entry (i.e. for 'olive tree': tree, olive, Jerusalem, etc.). The semantic cues enable to distinguish semantic clusters within a morphological family: words associated to the same idea (see Table 1 for Catalan and Spanish: unlike French, in these languages "oil" and "olive" are two clusters within the same morphological family).

In addition to the linguistic information on lexical entries, for each family, it is possible to see the number of derived items, the number of semantic clusters as well as an indication about how productive the stem might be (low, middle, high).

4.2 Bunches of words cross-lingually

LexRom displays word-families across languages. We thus consider the organization of the lexicon of a language as a set of "bunches of words" sharing a common stem and conceptual fragments. Our hypothesis is that such organization may be found across languages, particularly across closely-related ones. The data obtained will enable to give evidence on (mis)matches in terms of family sizes, lexical holes, equivalent clusters and specific phenomena concerning languages in contrast.

FR	*olive, olivade, olivaire, olivaie, olivaison, olivâtre, oliver, oliveraie, olivette, oliveur, olivier, olivine*
CA	*oli, oliada, oliaire, oliar, oliós, oliva, olivaci, olivaire, olivar, olivarda, olivarer, olivari, olivater, oliveda, olivella, olivellenc, oliver, olivera, oliverar, olivereda, oliverer...*
ES	*aceite, aceitadora, aceitar, aceitera, aceitero, aceitillo, aceitoso, desaceitar, aceituna, aceitunado, aceitunero, aceitunillo, aceituno*

Table 1: "Olive" family for French, Spanish and Catalan

As for lexical productivity, significant differences can come to light by observing the data in *LexRom*[11]. Table 1 shows as example the word 'olive': in Catalan and Spanish such form produces derived words for two different semantic clusters, 'olive' and 'oil' (with two different stems, *oli-* in Catalan and *aceit-* in Spanish). However, the corresponding stem in French (*olive*) only produces derived forms of the 'olive' family (the 'oil' family uses another stem, *huile,* and thus creates another word family). Similarly, *aguja* ('needle') and *agujero* ('hole') are part of the same word family in Spanish although in other similar Romance languages the semantic meaning of the latter is conveyed by means of different word forms: *buco* (Italian), *forat* (Catalan), *trou* (French), etc.). Such differences in terms of structure and productivity in the families will be put forward within the framework of *LexRom*.

In terms of access to words, two primarily functionalities are foreseen (in a similar way to *Polymots*), namely, access by word-form and access by semantic cues.

First, by typing a **key word** ('olive') the user will obtain the equivalent families in all the existing languages. This may include families where a same stem produces several semantic clusters ('olive' and 'oil' in Spanish) or narrower families corresponding to a single semantic cluster ('olive' in French, but not 'oil'). Correspondences between equivalent words will be highlighted. Clusters of one family conveyed by other stems ('oil' in French) will be available by navigation through the lexical graph. The user will thus be able to browse from family to family and from a particular word to its equivalents.

Second, by entering **semantic cues**, the user will obtain a word in the same language as the one used to enter the conceptual items (e.g. 'tree', 'Mediterranian', 'robust' and 'peace' will display the 'olive' family because of the key word 'olive tree' obtained as a result of the query). A list of all the words in the same families in the other languages will also be displayed. Semantic cues, initially extracted from French, will be automatically translated for the other languages. They will provide a dynamic way to access to words.

Finally, other criteria will be proposed for searching, i.e. **productivity** of a family in a particular language or **specific affixes across languages**, to give two examples among others.

At the time of the writing this paper, *LexRom* is under development: 1 741 families for French, 190 for Spanish and 77 for Catalan have been gathered (about 25 000 words overall). Automatic acquisition with very large

[10] This example is a clear evidence that derivational morphology is very frequent in Romance languages, while other languages like English 'prefer' other morphological processes to create new words (compounding, composition, etc.).

[11] Lexical productivity stands for the number of derived forms of a stem. Interlingual contrastive examples may be explored with *LexRom,* e.g. the word *chaise* 'chair' has a single derived word in French *-chaisier-* while its Spanish equivalent *silla* generates up to twenty-three derived items.

corpora and other existing lexicons is needed to scale up the resource (which will be freely available under a Creative Commons licence).

5. JeuxDeMots

JeuxDeMots (Lafourcade, 2007) is a web-based associative game[12]where people are invited to play on various lexical and semantic relations between terms. A particular formatting of the lexical network with a associative dictionary tool is already available for users[13], allowing them to browse the network by jumping from term to term through relations (free association, hyperonym, hyponym, part-of, typical locations, typical subjects and objects for verbs, etc.). Semantic text analysis is the main application for exploiting this resource, however the use as a tool for providing help in the case of the 'tip of the tongue' phenomena is also fruitful.

5.1 A game on popular consensus

Semantic information is collected through non negotiated popular consensus. By consensus, we mean that when several players independently propose during a game the same association, they are memorized in the lexical network. The approach is non negotiated as players are playing without knowing until the result of the game with who they are playing, hence avoiding some interaction than would bias strongly the result. To ensure a system leading to quality and consistency of the database, it was decided that the validation of the relations given by a player should be made by comparison with those of other players. Practically, a relation is considered as correct if it is given by at least one pair of players, making v*alidation by pairs* a form of minimum filtering. This approach is similar to the one used by (von Ahn et al., 2004) for the indexation of images or more recently by (Lieberman et al., 2007) to collect common sense knowledge. As far as we know, this was never done in the field of the lexical networks. In NLP some other web-based systems exist, such as *Open Mind Word Expert* (Mihalcea and Chklovski, 2003) that aims to create large sense tagged corpora with the help of Web users, or *SemKey* (Marchetti et al., 2007) that exploits *WordNet* and *Wikipedia* in order to disambiguate lexical forms to refer to a concept, thus identifying a semantic keyword.

A typical game takes place between two players, in an asynchronous way, based on the concordance of their propositions. When a first player (A) starts a game, an instruction related to a relation type (synonyms, opposite, domains,...) is displayed, as well as a term T pseudo-randomly picked in a base of terms. Player A has then a limited amount of time to answer by giving propositions which, to his mind, correspond to the instruction applied to the term T. The number of propositions is limited inducing players not just type anything as fast as possible, but to choose amongst all answers he can think of. The same term, along the same instruction, is later proposed to another player B; the process being then identical. To increase the playful aspect, for any common answer in the propositions of both players, they receive a given number of points as a reward. The calculation of this number of points is crafted to induce both precision and recall in the feeding of the database. More precisely an answer that is very commonly given would not be much rewarding contrary to an original one. Thus, players have to deal with a double opposite constraints: trying to think like others (to have words in common) while being as original as possible (to get points).

At the time of the writing of this paper, more than 1 100 000 relations linking more than 230 000 terms have been collected. More than one million games (with a mean of 1 minute per game) have been played corresponding to approximately 17 000 hours (about 700 days) of cumulative play. The lexical resources produced with *JeuxDeMots* are freely available under a Creative Commons licence.

5.2 Tool and Evaluation

The question of evaluation the lexical network is difficult. Indeed, there is no comparable resources that could be used as a golden standard. Another way to tackle the problem is to devise a tool that could help people finding a word they have on the tip on the tongue. The success rate of the tool can serve as a rough evaluation of the underlying resource. AKI is such a tool, where people are invited to submit clues one after the other until the system proposes the expected terms, or fails. AKI is used as a tool for lexical access but in fact most people use it as a game where the goal is to challenge the guessing capabilities. Clues can take the form of words or a composition between a relation type and a term. For exemple, the clue *:isa* animal, states that the target word is an animal. Around 20 relations are available. Amongst other, AKI proposes: *:syn* for synonyms, *:antofor* antonyms, *:mat* for mater or substances (like *:mat* wood, the target term is made of wood), *:carac* for typical feature (like *:carac* white), *:part* for typical parts (like *:part* wheel), *:do* for typical action (like *:do* meow), etc.

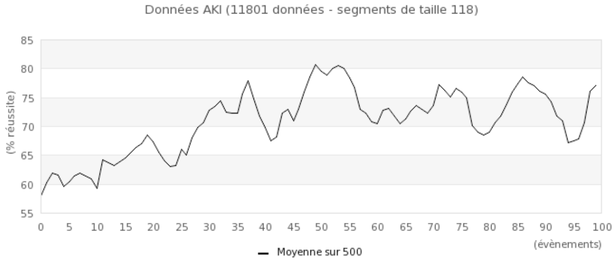
For example, a typical tip of the tongue game would be:

***:isa* boat**	**liner**
shipwrec	**Titanic**
torpedoing	**Lusitania**

Table 2: Typical AKI game (clues are on the left and answers made by AKI on the right)

[12] http://jeuxdemots.org
[13] http://www.lirmm.fr/jeuxdemots/diko.php

We evaluated the AKI performance for about 10 000 games. The overall performance is about 74% of success. We made an assessment of human performances on 200 games randomly taken from those played with AKI, and discover that under the very same conditions people have around 48% success.



6. Conclusion

LexRom and JeuxDeMots illustrate alternative ways from classic lexicographic work to create and enrich lexical resources. They both provide novel accesses and usages that are feasible only in the context of computer and user networks. Obviously, although the major part of the data is acquired automatically or by contributions, evaluation and validation of the resources with human contribution is essential to ensure the linguistic quality of the data. In this sense, NLP and lexicographic approaches converge undoubtedly. Still, the coverage of the resources, i.e. the amount of data obtained with automatic acquisition or collaborative contributions, remains a key issue of such novel approaches.

Last but not least, a number of open issues remain. As for *LexRom*, as the data is gathered mainly from corpora, the constitution of very large high quality corpora in different languages is crucial. At the time of writing this article we are working on such direction to be able to obtain more data. Likewise, theoretical aspects concerning lexical holes and polysemy deserve special attention in terms of language modeling. The *JeuxDeMots project* faces the same issues as previously mentioned concerning the coverage, but also concerning the qualitative evaluation. Some forthcoming research directions will include more common sens knowledge and the introduction of various non standard lexical relations. Collecting lexical information of very particular relations, either specialized or with too few possible answers, doesn't seem to be feasible with games. Thus, some contributive approaches with strong user incentives are still to be invented.

7. Acknowledgments

The authors would like to thank M. Zock as well as the reviewers of the paper for their valuable insights.

8. References

Atkins, B.T.S., Zampolli A. (1994). *Computational Approaches to the Lexicon.* Oxford: Oxford University Press.

Baayen, H.R., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, Philadelphia: University of Pennsylvania.

Briscoe, T., Carrol J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, DC, pp. 356-363.

Boitet, C., Mangeot, M. & Serasset, G. (2002). The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In N. Ide, G. Wilcock (eds.) *Proceedings of on Natural Language Processing and XML, COLING Workshop*. Taipei, Taiwan, pp. 9-15.

Calzolari, N., Choukri, K., Fellbaum, C., Hovy, E. & Fellbaum, C. (1999) *Multilingual resources. Chapter 1. Multilingual Information Management: current levels and future abilities*. Report commissioned by the US National Science Foundation and the European Commission's Language Engineering Office.

Clément, L., Sagot, B. & Lang, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of Fourth international conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, pp. 1841-1844.

Cristea, D., Forăscu, C., Răschip, M. & Zock, M. (2008). How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach. *In Proceedings of Sixth international conference on Language Resources and Evaluation (LREC-2008)*, Marrakech.

Fellbaum, C. (1998). *WordNet: an Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C., Miller, G.A. (2003). Morphosemantic links in WordNet. In M. Zock, J. Carroll (eds.) *Les dictionnaires électroniques*. TAL vol. 44 (2), pp. 69-80.

Fort, K., Adda, G. & Cohen, K.B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2), pp. 413-420.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of Fifth international conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Gala, N., Rey, V. & Zock, M. (2010). A tool for linking stems and conceptual fragments to enhance word access. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC*), La Valetta, Malta.

Gala, N. (2011). Developing a lexicon of word families for closely-related languages. In *Proceedings of ESSLLI International Workshop on Lexical Resources* (WoLeR). Ljubljana, Slovenia.

Gasiglia, N. (2009). Evolutions informatiques en lexicographie: ce qui a changé et ce qui pourrait émerger. *Lexique 19*, pp. 224-298.

Grefenstette, G. (1998), The future of linguistics and lexicographers: will there be lexicographers in the year 3000? In T. Fontenelle, P. Hiligsmann, A. Michiels, A.

Moulin & S. Theissen (eds) *Proceedings of the Eighth EURALEX Congress*. Liège: University of Liège, pp. 25-41.

Heid, U. (2009). Aspects of lexical description for electronic dictionaries. Key note speech at *Electronic lexicography in the 21st century (ELEX-2009)*, Louvain, Belgium. http://www.uclouvain.be/en-271026.html.

Ide, N., Véronis, J. (1994). Machine Readable Dictionaries: what have we learned, where do we go? In *Proceedings of the post-COLING International Workshop on Directions of Lexical Research. Beijing, China.*

Jakubíček, M., Kilgarriff, A., McCarthy, D. & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. In *Proceedings of Workshop on Advanced Corpus Solutions, PACLIC 24.* Tohoku Univeristy, Japan.

Kilgarriff, A. (1997). Putting frequencies in a dictionary. *International Journal of Lexicography*, 10(2), 135-155.

Lafourcade, M. (2007). Making people play for Lexical Acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing*. Pattaya, Thailand.

Lieberman, H., Smith, D.A. & Teeters, A. (2007). *Common Consensus: a web-based game for collecting commonsense goals, IUI'07*, Hawaii, USA.

Mihalcea, R., Chklovski, T. (2003). Open Mind Word Expert: Creating Large Annotated Data Collections with Web Users' Help. In *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora (LINC 2003)*, Budapest.

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M. & Minutoli, S. (2007). SemKey: A Semantic Collaborative Tagging System. In *Proceedings of WWW Conference*, Banff, Canada.

Navigli, R., Ponzetto, S. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010, pp. 216-225.

Navigli, R., Velardi, P. & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent systems*, 18(1), pp. 22-31.

Poesio, M., Kruschwitz, U. & Chamberlain, J. (2008). ANAWIKI: Creating Anaphorically Annotated Resources through Web Cooperation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Véronis, J., Ide, N. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *Proceedings of 13th International Conference on Computational Linguistics (COLING'90)* Helsinki ,vol. 2, pp. 389-394.

Von Ahn, L.L.D. (2004). Labelling images with a computer game. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI) 2004,* Vienna, Austria.

Vossen, P., (1998). *EuroWordNet: A Multi- lingual Database with Lexical Semantic Networks.* Dordrecht, The Netherlands: Kluwer.

Computer-Aided Inflection for Lexicography Controlled Lexica

Maarten Janssen
IULA - UPF
Roc Boronat 138, 08018 Barcelona
E-mail: Maarten.Janssen@upf.edu

Abstract

This article describes the design of a computational system for the development and maintenance of inflected lexica, developed as part of the Open Source Lexical Information Network (OLSIN). The system is built as a tool for lexicographers, and is flexible enough for the lexicographers to deal with any irregularities in the language, and transparent enough for the lexicographers to understand the rules used for the automatically generated inflections. It furthermore allows lexicographers to create and modify paradigm rules by themselves, making it easy to implement the system for any language, including less-resources languages. Apart from the system itself, this article describes some of the challenges and obstacles the design of such a system has to face, and the solutions adopted for them in the OSLIN framework.

Keywords: inflectional morphology; paradigm system; full-form lexicon

1. Introduction

Dictionaries have always included inflectional information, though be it only in limited amounts. Inflectional information is traditionally limited to the key word-forms of irregularly inflected words only, due to restriction on the size of dictionaries. With the rise of electronic dictionaries and the loosening of size restrictions they imply, this situation has changed: more and more dictionaries include the full inflection for all words, or at least for all words of heavily inflecting word-classes: in modern Spanish dictionaries, all verbs are provided with their full inflection, although nouns and adjectives are not. This information is included for two reasons: firstly, it is lexical information that many people are interested in, and therefore information that is useful to include when possible. Secondly, it allows the user to find words in the dictionary without knowing the citation form: if you do not know that *fue* is an inflected form of *ser* (to be) in Spanish, it is difficult to find in a dictionary that does not include inflected forms. With inflection moving into a prominent place in dictionaries, their quality should match up with the rest of the dictionary.

To add inflection to dictionaries, most dictionaries rely on computational tools. Yet with respect to the creation of the inflected forms there is a tension between automation and freedom: it is hard at best to created a large-scale inflected lexicon without the use of computational tools, yet computational tools tend to limit the lexicographer in defining precisely the inflection he deems correct.

This article describes a computational system that aims to overcome this tension by allowing the lexicographer full control over the inflected forms, while at the same time automating the process of inflection as much as possible. This tool, which forms part of the Open Source Lexical Information Network (OSLIN) framework (Janssen, 2005), does this by using a paradigm-based inflection system where the lexicographer himself can create, apply, and modify the paradigms. It is a tool designed for practical usability for lexicographic purposes, without too much emphasis on computational innovation or efficiency.

The set-up of this article is as follows: the next section describes which requirements a computational tool has to meet in order to allow lexicographer sufficient guidance and freedom to develop a high quality inflectional database. Section 3 describes how these requirements are implemented in the OSLIN environment. Section 4 describes how the OSLIN tool can, and has been used in practice to create a large-scale, high-quality full-form lexicon. And section 5 describes some of the more complex issues that play a role in the task of semi-automatic paradigmatic inflection.

2. Design Requirements

It is possible to create a full-form lexicon by manually inserting all inflectional forms. However, to do so is very labour intensive: a full-form lexicon for a morphologically rich language typically contains many more forms than it does lexical entries: for Portuguese for example, every lexical entry has on average 10 forms, meaning that a medium-sized dictionary contains well over a million inflected forms. Not only is it labour intensive, it is also extremely hard to avoid typographic errors in that many inflected forms, especially since the inflected forms of a word typically differ very little amongst each other. Therefore, the use of computational tools in this task is highly desirable, both to save work, and to prevent errors.

However, inflection can be a complex issue: although for the majority of words, it is clear and undisputed how they inflect, there are many cases where inflection is less clear and where information has to be taken into account that is not (easily) computationally available, such as normative rules, etymology and pronunciation. Even relying on usage information is not necessarily sufficient:

it is not unseen for a noun that the plural that is considered normatively correct is not the plural that is most frequently used. Another complication is that sometimes, a form other than the standard form is used as the citation form, for instance for noun that are almost exclusively used in the plural.

In the more complex cases of inflection, the computational tools used should not limit the lexicographer in defining exactly the inflected forms he/she considers correct: for instance, it should be a lexicographic decision whether to include the normative forms, the most frequently used forms, or both in those cases in which these two are not the same. The computational mechanisms should therefore suggest an inflectional paradigm to the lexicographer, but always allow changing or overruling the suggestions made.

The easiest way to implement a computational system that produces suggestions that can be modified at will is by having an external system that only intervenes during the creation of a new lexical entry: when creating new word, the lexicographer can choose to either accept the suggestion made by the tool and insert the computationally created forms into the dictionary, or he can ignore the suggestion, and choose to either insert the suggested forms but modify them afterwards, or completely ignore the suggestion and add the forms manually.

Although this is a workable system, and the method used in the OSLIN framework, it has two major drawbacks. Firstly, such a system often works as a black box, which makes it a lot harder for the lexicographer to spot potential errors. And second, the manually corrected or inserted forms are still subject to typographic errors. This means that over time, it is inevitable that errors creep in. And because it is not longer transparent which forms were automatically created, and which were manually altered, it becomes very hard to spot those errors in a large-scale dictionary. This is not merely a theoretical point, but shown by experience in OSLIN before the introduction of the paradigm system described here.

For these reasons, a more reliable set-up is a system in which the inflectional mechanisms are build into the dictionary system in a more integrated way, so that manual correction can be spotted, and ideally, the manual corrections are even done via the same computational tool. Furthermore, it is important that the computational implementation is done in a transparent fashion, so that the lexicographer can understand why the system suggests the inflected forms it does. This makes it much easier for the lexicographer to work with the system, and avoids errors.

The most intuitive way of computationally treating inflection is by using a paradigm-based system: a word is assigned to an inflectional paradigm, and based on that paradigm, it gets assigned a number of inflected forms. Although there are other, more computationally efficient ways of encoding inflection, it is much more difficult to make the inner workings of such systems clear to a lexicographer without a computational background.

An additional advantage of a paradigm-based system is that it is well-known strategy in the lexicographic tradition, used in grammar books and existing (paper) inflection dictionaries for many inflecting languages such as for instance *Els Verbs Conjugats* (Baptista Xuriguera, 2009) for Catalan. In order to explain to the end user how to inflect the Portuguese verb *bailar* (to dance), it is possible list all the (73) inflected forms, but it will in most cases be sufficient to say that it is completely regular, or that it inflects like *amar* (to love). Indicating the inflectional paradigm not only is a convenient way for the user to understand the inflection, but also allows the user to see other words that inflect in the same way.

If paradigms are presented to the end user, the paradigm system not only has to be able to correctly inflect all words in the lexicon, but also correspond to what are traditionally considered to be the paradigms of the language. This seems a trivial issue, but there are many divergences between a computational and a human perspective on inflectional paradigms (see section 5.1).

There is no single unique way to set-up a paradigm system: there are typically several different sets of paradigms, each of which sets can equally be used to define the inflection of a language. The choice of paradigms itself can in such cases itself become a lexicographic or political matter. Therefore, it is very convenient if the system allows the lexicographer to construe the paradigms himself, and ideally do so without having to rely on a computational linguist. Furthermore, it should be possible to modify the paradigms when the need arises, for instance when the orthography changes, or when it is decided that another choice of paradigms is more appropriate.

Finally, a computational tool for the inflection of a lexicon should ideally be, as far as possible, language independent. That is to say, it should be usable for as many languages that have an inflectional morphology as possible, so that the same computational tool can be used for a wide variety of languages.

2.1 Open Source Lexical Information Network

The inflection system described in this article is part of the Open Source Lexical Information Network (OSLIN). OSLIN is a language independent framework for modeling lexical information, with a focus (for the moment) on inflectional and derivational morphology. The system was originally developed at the ILTEC institute in Lisbon for Portuguese, and has since been extended at the IULA institute in Barcelona to several other language in various degrees of detail, including

Spanish, Catalan, Russian, Dutch, and French. OSLIN is a web-based system, and most of the lexicons can be accessed via the project website (www.oslin.org).

The inflectional part of OSLIN consists of a relational database with two tables. The first table contains the lexical entries with their citation form, word-class and other information, including the inflectional paradigm. The second table contains for each lexical entry all the inflected forms related to it, with their orthography and an indication of which inflected form it is.

The paradigm system of OSLIN, described in this article, is used in several ways within the framework. For the lexicographer, the system is used to create and fill the forms table with all inflectional forms for new words, but also to correct words already inflected. And the system can be used to make selections of similarly inflected words to facilitate correction processes. For the end user, the paradigms assigned by the system are used in the online display of the inflectional information for each word in the lexicon, as well as to display each paradigm with its inflected forms, together with a list of all the words that are inflected via that paradigm.

3. OSLIN Paradigm Manager

In a computational paradigm-based inflection system, there are three different aspects to the design and use of paradigms: firstly, there are the rules or mechanisms that define the paradigm itself. Secondly, there is a need for a system to create those paradigmatic rules, and finally, these rules have to be applied to create the actual inflected forms for the words in the lexicon. This section describes the way these three aspects are implemented in the OSLIN inflectional system. Most of the examples given in this section are in Spanish, but there is nothing specific to Spanish in the design, and examples from for instance Russian would work just as well.

3.1 Paradigm Definitions

A paradigm in the OSLIN system is an entity for creating inflection forms for a selection of words in a specific word class. Each paradigm has a unique identifier, which indicates the word-class it relates to, followed by a sequential number. For example, ADJ01 is the first paradigm for the inflection of adjectives. To make it easier to identify the paradigms, each paradigm furthermore has a prototypical word associated with it, which is typically the most recognizable word that belongs to that paradigm. For Spanish, the example word for ADJ01 is the adjective *gordo* (fat), being the paradigm to inflect *gordo* and all words that inflect like it. In this article, paradigms will be identified always by their prototypical word for ease of reading.

The core of the paradigm is a set of string transformation rules: rules that create the orthography for all the inflected forms, starting from the orthography of the citation form, by transforming the string of characters. The reason why the rules start from the citation form is that they lexical entries in dictionaries are identified by their citation form or headword. These string transformation rules define that the inflected for the word *gordo* with paradigm ADJ01 are: *gordo, gorda, gordos,* and *gordas* respectively, and the likewise relate all similarly inflecting words to their respective inflected forms. There are three types of string transformation rules in the OSLIN system: root-creation rules, inflection rules, and root-alternation rules.

For each paradigm, the root-rule defines how to generate the “root”, or better the invariant part, of the paradigm based on the citation form. The root in the paradigm system is not necessarily the linguistic root, but simply the part of the citation form that remains invariant throughout the forms in the paradigm. For readability, a hyphen is placed at the end of the root in the examples in this article; this hyphen does not correspond to anything in the actual rules. An example of a root-creation rule is given in (1), which is the root-creation rule for the Spanish adjectival paradigm *gordo*. The rule is a regular expression rule, here formulate in Perl for convenience.

(1) ($root = $citation_form) =~ s/o$//;

The root-creation rule in (1) states that the root for *gordo* is formed by removing the *–o* at the end of the citation form. This rule generates the root *gord-* for *gordo*, *blanc-* for *blanco*, etc.

The inflection rules are rules that create individual inflected forms from the root. Each paradigm has as many inflection rules as there are inflected forms for the paradigm. For the paradigm *gordo*, there are therefore four inflection rules, one for each of the four adjectival forms in Spanish (masculine and feminine singular and plural). The inflection rule for the feminine plural of *gordo* is given in (2).

(2) ($inflection[‘fem plur’] = $root) =~ /$/as/;

The inflection rule in (2) states that the feminine plural of *gordo* consists of the root, with the suffix *–as* added to the end. Together with the root-creation rule in (1), this defines that the feminine plural form of *gordo* is *gordas*, for *blanco* it is *blancas,* etc.

The reason why paradigms are defined in terms of simple string transformation rules, and not for instance in the more powerful system of Two Level Morphology (Koskenniemi, 1983), is the aforementioned fact that paradigms are the most intuitive way for lexicographer to deal with inflection.

All inflected forms can be defined in terms of string transformation rules from the citation form, although sometimes only in a trivial way: in a paradigm where no letters are shared between all forms, the “root” form has to be empty, and the word-forms are created by adding the entire word to the empty root.

However, in many cases simple transformation rules lead to paradigms that do not correspond to the traditional paradigms of the language. For example, the Dutch words *jaar* (year) and *boor* (drill) are typically seen as inflecting the same: in their plural form, the double vowel is replaced by a single vowel, leading to the respective plurals *jaren* and *boren*. In a straight-forward string transformation system, these words would, however, end up as having different paradigms: one where the ending *–ar* is replaced by *–ren*, and another where the ending *–or* is replaced by *–ren.*

To implement paradigms in a way that closer resembles traditional paradigms, OSLIN uses root-alternation rules. Root-alternation rules create alternate root forms from the base root form. When using multiple roots, the inflection rules have to indicate which of the root forms is used for each particular form. An example is given in rules (3)-(5). The root-creation rule in (3) defines that the (main) root for *jaar* is identical to the citation form. The root-alternation rule in (4) creates a second form of the root by removing the double vowel in the final syllable of the main root. Finally, the inflection rule in (5) defines that the plural for *jaar* is formed by placing *–en* at the end of the alternate root.

```
(3)  ( $root = $citation_form ) =~ s/$//;
(4)  ( $alt_root[1] = $root ) =~
     s/([aeou])\1([dgklmnprt])$/\1\2/;
(5)  ( $inflection['plural'] = $alt_root[1] ) =~ /$/en/;
```

The rules (3)-(5) for the word *jaren* first create a root *jaar-*, then create an alternative root *jar-*, and then form the plural *jaren* from the alternative root form. In the case of *boor*, the rules create *boor-*, *bor-* and *boren* respectively.

Root alternation rules can be used to group words together under the same paradigm that are traditionally consider to inflect alike. However, they can also be used to make the "root" of the paradigm resemble the linguistic root more closely. For instance, in the Spanish nominal paradigm *actriz/actrices* (actress), rules without root alternation use a root form *actri*, with endings *–z* and *–ces* in the singular and the plural form. With root alternation, it is possible to create the same form in a more linguistically appropriate way, defining *acriz* as the (main) root, with a plural ending *–es* and a root-alternation rule *z/c*.

3.2 Creating Paradigms

Paradigm-based inflection systems using string transformation or transduction rules are far from new, dating back at least to Matthews (1972). It is not difficult to generate this type of rules by hand, even though exceptions in the inflection of a language can make the set of rules quite complex. However, if we want the lexicographer to be able to create and maintain the paradigm system by himself, without having to possess a lot of computational knowledge or a computation linguist, the system should not rely on manually created rules.

Therefore, in OSLIN the rules for the paradigms are created automatically, based on example data provided by the lexicographer. The way this works is very simple: the lexicographer types in all the inflected forms of an example word by hand, and the system attempts to determine which rules have to be defined in order to generate all the forms that the lexicographer entered, starting from the citation form.

In order to allow adding the inflected forms, the only thing the system needs to know is which forms the word has, and a graphical way to organize these forms to make the data easier to read. In OSLIN, these two things are handled by a template: for each (major) word class, there is a template that defines a graphical display of all the inflected forms for that word class. A template is a simple HTML text file, containing this information. An example of a template for Spanish adjectives is given in figure 1.

```
<table>
    <tr><td><th>masulin<th>feminin
    <tr><th>singular<td>{%ms}<td>{%fs}
    <tr><th>plural<td>{%mp}<td>{%fp}
</table>
```

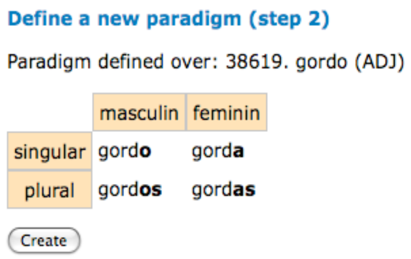
Figure 1: Paradigm Template for Spanish Adjectives

The template on the one hand defines that adjectives in Spanish have four forms, labels by *ms, fs, mp,* and *fp* respectively. And it defines an HTML table to graphically display these four forms in a convenient way. The template is both used to display the inflection of already inflected words, and to create an HTML form for the insertion of a new paradigm.

When the template is created, the lexicographer can select a word for which he want to enter the inflected forms, say the Spanish adjective *gordo*. The system will then use the Spanish adjectival template in figure 1 to display an HTML form with a text box for each of the four individual forms, which the lexicographer is asked to fill in.

After the form has been submitted, the system will establish the longest sequence of characters that remains invariant throughout all the forms that were entered (the root), and define the root-creation rule necessary to create the root from the citation form. Once the root is established, the system will define which inflection rules are need to generate the inflected forms from the root.

To verify no errors were made, the result is then shown to the lexicographer, with the root in normal type, and the "affixes" in bold face. The result for the Spanish adjective *gordo* (fat) is shown in figure 2. Only after the lexicographer confirms that the paradigm is correct will the system store the new paradigm ADJ01 in the system.



Define a new paradigm (step 2)

Paradigm defined over: 38619. gordo (ADJ)

	masculin	feminin
singular	gord**o**	gord**a**
plural	gord**os**	gord**as**

Figure 2: Paradigm definition *gordo* (fat)

While looking for the "root", the system looks for characters, without relying on any type of linguistic knowledge. This makes the process very language independent, and the process works for Spanish just as well as it does languages with another alphabet (like Russian) or languages that have prefixing or circumfixing inflection rules. However, there are paradigms it cannot handle, notably those cases where inflection is not taking place at the beginning and/or the end of the word (see section 5.2).

As mentioned in the previous section, paradigms may contain root alternation rules, and whether or not to use root alternation rules is largely a matter of choice. From only one example, it is impossible for the system to assess what the intended root is, nor what the intended root alternation rules would be. Rather than asking for a large set of examples, the system allows the lexicographer to define the root alternation rules by hand. To facilitate the definition of root-alternation rules, the rules are defined not directly in terms of regular expressions, by in terms of a "from" and a "to" part, separated by a slash, and apply to the end of the root.

To define an alternating pattern for the word *actriz* in Spanish, the lexicographer enters all the inflected forms, plus the (simple) rule *z/c*. With this information, the system automatically compiles all the necessary rules from the example provide, just as in the case without root alternation.

Although simple cases of root alternation rules are easy to define, alternation rules can become rather complex. A somewhat complex rule was given in (4), repeated below, but the rules become even more complex when the root alternation rule adds or removes and accent. Although it would be desirable to facilitate the automatic creation of root alternation rules, root alternation rules currently have to be entered manually. As such, root alternation is the only place in the system where some computational knowledge is required from the part of the lexicographer, but only in cases where the lexicographer wants to use complex root alternation rules.

(4)
```
( $alt_root[1] = $root ) =~
    s/([aeou])\1([dgklmnprt])$/\1\2/;
```

3.3 Recognizing Paradigms

When a complete set of paradigms has been defined, it is possible to use the paradigm system to inflect any word of the language: by assigning a word its appropriate paradigm number, all the inflected forms are implicitly defined. However, assigning the paradigm number by hand is a tedious task, and therefore the system should be able to assign the correct paradigm automatically. That is to say, the system should recognize for any new word what the appropriate paradigm is.

In the OSLIN system, this is computationally implemented by constraints that indicate which paradigms are *not* appropriate, until only the correct one(s) remain(s). There are three types of constraints: hard constraints, blocking constraints, and soft constraints.

Hard constraints define which characteristics the word has to have in order for the paradigm to apply. For instance, in Spanish, there is a paradigm for verbs like *actuar* (to act)*:* these verbs get an accent on the *–u–* in, for instance, the first person present indicative: *(yo) actúo*, to make it clear that the accent is on the *–u–* and not on the *–a–* or the *–o–*. This paradigm applies *only* to verbs on *–uar,* which means that the ending *–uar* is a hard constraint for this paradigm.

Blocking constraints state that one paradigm blocks another: in order to for a Spanish noun to inflect like *casa* (house), it has to be a regular noun, which means it should not follow one of the more restrictive paradigms such as the paradigm for words on a *–z* like *actriz*. Words that fall under the paradigm *actriz* do not inflect like *casa*. So meeting the constraints for the paradigm *actriz* means that the word *cannot* inflect like *casa.*

When properly set-up, the combination of hard constraints and blocking constraints define a complete paradigm system, in which the system will only suggest possible paradigms. There can be more than one paradigm suggested though: for instance in the case of a new word on *–olar* in Spanish, there are no formal clues whether the new word should have *–olo* as its first person singular present indicative, like *molar*, or rather *–uelo*, like *volar*. In such cases, it will be up to the lexicographer to choose between the various possible paradigms. In most cases, however, the majority of words will have only one applicable paradigm, which means that the lexicographer will only have to intervene in a limited amount of cases.

Although a paradigm system with only hard constraints and blocking constraints works, it does not always lead to efficient paradigm recognition. An example is the case of invariable nouns in for example Catalan. There are no hard constraints a word needs to meet in order to be invariable: although most of the invariable words end in *–s*, there are quite a few examples of invariables nouns with other endings, especially when counting loanwords as well. Without constraints, the invariable nominal paradigm will be applicable to any new noun in Catalan, and the lexicographer would have to state for every noun that it is not invariable, which is far from efficient. To solve this, it is possible to define soft constraints. Soft constraints work like hard constraints, except that they can be overruled. In the case of Catalan invariable nouns, we can define a soft constraint that such words should end in an *–s*. With such a soft constraint, the system will by default ignore this paradigm for all other nouns, except when explicitly asked to show all available paradigms.

4. The Paradigm Manager in Action

The OSLIN paradigm manager is a fully implemented inflectional system that has been used to generate and manage the full-form lexicon of a variety of languages. When creating a new full-form lexicon, the manager can be used in two different ways, depending on what is available to start from: the system can either be used to inflect a word-list from scratch, or to create a paradigm system from an already inflected wordlist, and then use the system created from this already inflected word-list to inflect words added to the list afterwards. This is useful, for instance, in cases in which a small inflected lexicon is available, and a (much) larger inflected lexicon is needed. Given that the second option is easier to apply, it will be describe first.

4.1 From an Inflected Lexicon

When starting from an already inflected lexicon (however small), the creation of a paradigm system is relatively straightforward. To start the process, it is necessary to load the inflected data into the OSLIN database system. This means that the lexical entries have to be loaded into the table of lexical entries, and the inflected forms into the table of word-forms. Furthermore, a template has to be defined (see section 3.2) using the same codes for the inflected forms as used in the original lexicon.

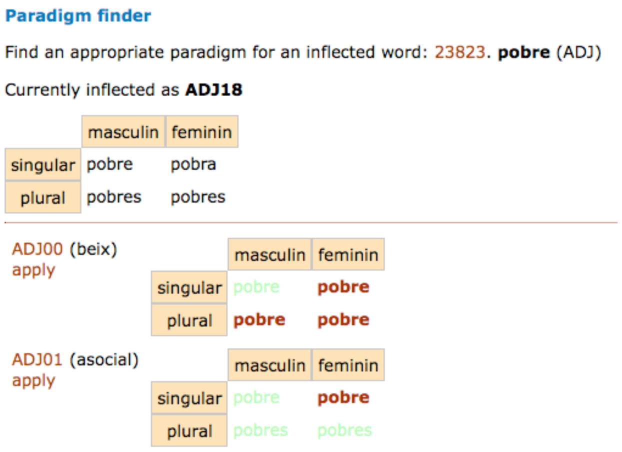
Once the inflected forms have been added to the system, the lexicographer can start setting-up the paradigm system for each of the major word classes in turn, starting for instance with the adjectives. The process is simple: start by selecting the most regular word you can think of, say the Spanish adjective *gordo*, and ask the system to generate a paradigm for it. The system will look-up all the inflected forms of the adjective *gordo* in the database, and determine the longest sequence of letters common to all of them, in this case *–gord–*. It will then suggest a paradigm as was shown in figure 1, where *–o, –a, –os,* and *–as* are the inflectional suffixes, and the root is created from the citation form by removing the last *–o*.

After creating a paradigm, the system will look through all the adjectives in the database that do not yet have a paradigm assigned to it, and check whether they conform to this newly created paradigm. It does this by applying the paradigm to the citation form of the adjectives (one by one), to generate all the inflected forms that the adjective would have if it inflected via this paradigm. The system then verifies if all forms generated by the paradigm are identical to the forms in the database for the adjective in question. If all the forms match, the word does belong to the paradigm, and the system will automatically assign that adjective the paradigm of *gordo*. To see the progress, it is possible to run this process in verbose mode, in which case the system will also indicate for words that do not match the paradigm, what is the first word-form where the generated forms and the stored forms diverge.

After checking all the adjectives, a lot of them will have been assigned the paradigm *gordo*, yet there will still be a large number of adjectives without a paradigm. The next step is to look at the list of adjectives that do not have a paradigm yet, and select one of them, say *grande* (big), and repeat the process: create a paradigm for it, and have the system mark all adjectives that correspond to the newly created paradigm *grande*. And continue this until all adjectives have been assigned a paradigm. Once all adjectives have a paradigm assigned to them, the paradigm system (for adjectives) is complete.

To help in selecting the next paradigm to create, the system can indicate why words that do not yet have a paradigm assigned to them do not belong to any of the existing paradigms. In figure 3, this is shown for the Catalan adjective *pobre* (poor) after a few paradigms have already been created. In this figure, there are two applicable paradigms, and both are compared to the known forms of *pobre*. The forms in green are those for which the form predicted by the paradigm matches the form stored in the database: the paradigm would (correctly) create the female plural form *pobres* if *pobre* would be assigned the paradigm *asocial* (asocial). The forms in bold red are those where the two forms diverge: if *pobre* would inflect like *asocial*, the female singular form would be *pobre*, whereas it is in fact *pobra*. Therefore, *pobre* in Catalan does not inflect like *asocial,* nor does it inflect like *beix* (beige).

In the comparison in figure 3, the system by default only displays applicable paradigms. Therefore, the Catalan paradigm *blanc* is not shown, since *pobre* violates the hard contraint that all adjective of the paradigm *blanc* have to end on a *–c*. If so desired, it is possible to have the system show all paradigms with violating constraints as well.



Paradigm finder

Find an appropriate paradigm for an inflected word: 23823. **pobre** (ADJ)

Currently inflected as **ADJ18**

	masculin	feminin
singular	pobre	pobra
plural	pobres	pobres

ADJ00 (beix)
apply

	masculin	feminin
singular	pobre	**pobre**
plural	**pobre**	**pobre**

ADJ01 (asocial)
apply

	masculin	feminin
singular	pobre	**pobre**
plural	pobres	pobres

Figure 3: Testing an inflected word

When starting from an inflected lexicon, creating a paradigm system in this way is a rapid process: with each new paradigm, the list of words without a paradigm gets shorter, and it is easy to see which paradigms are still missing. Somewhat more complex is defining the set of constraints to avoid the system from suggesting inappropriate paradigms, but given that for each paradigm, a list of examples will be at hand, it is easy to see which characteristics all of the words of a paradigm share.

When all the paradigms of the language have been defined, the list of words without a paradigm should contain only words that do not inflect like anything else in the language, most of which will be loanwords. Yet if the original database contained errors, a large percentage of the words that were incorrectly inflected in the original database will also remain on that list. This means that applying the paradigm manager to an already inflected lexicon is a quick way to detect errors in a full-form lexicon.

4.2 From a Word List

It is not always possible to start from an inflected lexicon, since an inflected lexicon is not always at hand. Therefore, it is also possible to create a full-form lexicon with OSLIN from only a list of words with their word-classes, provided for instance by a dictionary. The word list should be provided in that case as a simple spreadsheet with two columns: the first containing the citation form and the second the word-class the word belongs to. The system will then help to gradually assign a paradigm to each of the words on the list, and fill the OSLIN tables with the lexical entries and word-forms based on these paradigms.

Without inflected examples, the inflection has to be done interactively, working from the most regular paradigms to the most restrictive ones, and then gradually working back to the regular paradigms. How this process works is illustrated here for Spanish nouns.

Looking at Spanish nouns, the most common plural is the word with a *–s* placed at the end, as in the case of *casa* (house). Since the system does not know the plural of *casa*, we have to add it as a new word, and manually add the singular and plural form. After the word with its inflected forms has been added manually, it can be used to create a paradigm as described in the previous section.

Once the paradigm is in place, the system will display all nouns that could potentially belong to this paradigm; since there are no restrictions (yet), that will be the complete list of all nouns. Looking through that list, there are obvious words that do no inflect like *casa*. For instance, in words ending in a consonant, the plural *–s* is not added directly to the singular, but a linking vowel *–e–* is inserted: the plural of *afinidad* (affinity) is *afinidades* and not *afinidads.*

For each such "exceptionally inflecting" class, a paradigm has to be created, with the restriction that apply to that class. In this case by manually inflecting *afinidad* and then creating a paradigm from it for words ending in a consonant, and then verifying if the all the words on the more restrictive list with candidates for the new paradigm inflect indeed with that paradigm, and otherwise repeat the process. In the case of nouns ending in a consonant, nouns ending in a *–z* form an exception, since they get an orthographic root alternation in their plural form: *actriz/actices* (actress) and not *actrizes*.

The paradigm for *actriz* is restrictive enough to apply to (virtually) all nouns on *–z*. Once such a restrictive paradigm is reached, the system can be asked to inflect all words matching the requirements according to that paradigm. That is to say, we can ask the system to inflect all nouns ending on a *–z* in our wordlist via the paradigm *actriz*.

Once the nouns on *–z* are taken care of, it is necessary to return to the more general paradigm (*afinidad*) to verify if there are more exceptions. Once all exceptions to the paradigm *afinidad* have been taken care of, the paradigm *afinidad* can be applied to all remaining nouns ending on a consonant. Once all words ending on a consonant have been inflected, it is time to return to the remaining list of nouns to see if there are other classes of nouns that do not end in a consonant, yet do not inflect like *casa* either, until finally, all remaining nouns can be inflected like *casa*.

Even a restrictive paradigm like *actriz* is not fully without exceptions, although in this case there is only one exception in the *Diccionario de la Lengua Española* (RAE 2001): the word *kibutz* (kibutz) is a foreign loanword and does not change in the plural. When spotting the exception before inflecting all the words on *–z*, it can be inflected by hand, which will mean it will not receive the paradigm *actriz* since it has already been inflected. If it is not spotted before, since there are hundreds of words on *–z* that do follow the paradigm it is easy to overlook, it can always be changed afterwards (see 4.3).

It can happen that a whole paradigm is overlooked, meaning that a class of words got inflected via the wrong paradigm in the process described above. For instance, it is easy to overlook the paradigm for *virgen* (virgin), which receives an accent in the plural: *vírgenes*. When the words of this paradigm have already incorrectly been inflected via the paradigm *afinidad*, this can be corrected by manually correcting the inflected forms for *virgen*, and subsequently create a paradigm out of the corrected inflection, which can then be applied to all words on *–en*.

Using this strategy, we have managed to create reliable, full-forms lexica with around 50.000 to 100.000 lexical entries (over a million inflected forms) for several languages in a relatively small amount of time, with an estimate of around 500 man-hours.

4.3 Post-Verification and Maintenance

Even when created with the utmost care, a large-scale lexicon with over a million word-forms is never fully correct. Therefore, it will be necessary to correct errors after the original creation of the database. The OSLIN administration environment is not built as a tool usable only for the creation of a full-form lexicon, but as a management tool for the creation and continuous maintenance of lexical resources.

The OSLIN tools easily allow to choose a different paradigm for an already inflected word to correct a wrongly inflected word, or to change the inflected forms manually if it does not belong to any paradigm. The problem is to find errors in a database that large. Using external resources such as traditional grammar books and existing dictionary helps in finding words that are known to have an exceptional inflection, and therefore are the most likely to have gone wrong in the semi-automatic inflection process. But the OSLIN databases are built to have an alternative way of finding and correcting errors: improvement by use.

The OSLIN resources are not intended to be passive word-list, but rather lexical resources to be actively used. The database is set-up to be used as the exclusion lexicon for neologism research, and the system comes with integrated tools for use as a part-of-speech tagger and a spelling checker. The part-of-speech tagger can report on words that look like known words that are inflected differently in the corpus than in the lexicon. It does this by automatically lemmatizing unknown words, and then looking for words with a known citation form and word-class, but an unknown inflected form.

Furthermore, the data of the OSLIN lexica are directly available online in a user-oriented web site with rich search capabilities. Each page showing the inflected forms of a word has a “report” button on it, which allows the users to provide feedback on errors in the database (although the report function can be disabled for a language when it is not desired).

Most of the feedback coming from the tagger or the online users is not an indication of an error in the database, but rather mistakes by the users or the authors of texts in the corpus. However, the occasional error in the database is likely to be found by these methods over time.

4.4 Less-Resources Languages

As explained earlier, the OSLIN paradigm system can be used to inflect a lexicon for a large variety of languages, since there is nothing language-specific in its design. And furthermore, the system can be set-up and used by a lexicographic team, with only a minimum amount of external help, and without the need for trained computational linguists.

These characteristics make the OSLIN paradigm inflection system very well suited for use with smaller languages for which fewer resources exist. For less-resourced languages, lexicographic sources and lexicographers often are available, but finding computational linguists to work out the inflectional system of the language is more problematic. With the OSLIN tools, it is possible for lexicographers to create and maintain a reliable, large-scale lexicon for such languages, using a framework that furthermore facilitates the creation of the tools mentioned in the previous section: a part-of-speech tagger, a spelling checker, a neologism detection tool, and an online language consultation site.

5. Issues

The use of paradigms is a powerful and intuitive way to treat inflection. However, there are cases where the use of paradigms for inflection raises problems. This section describe three cases in which issues with the use of a paradigms in a semi-automatic detection tool arise, and sketches how these issues are, or can be dealt with in the OSLIN framework.

5.1 Computer vs. Human Paradigms

As mentioned before, what humans consider to be words that inflect the same does not always correspond to what a computational system would do. The root alternation rules bring the two closer together, but that does not in all cases conflate the two. Below are some cases where mismatches remain, although most of them can be overcome.

For several languages, traditional grammars include paradigms that are computationally speaking fully redundant. For instance, the *Normes Ortográfiques* for Asturian (ALLA 2005) includes a paradigm for *panaderu* (baker), even though it inflect fully regularly like *llobu* (wolf). The paradigm is included for the sake of clarity, and not to indicate an irregularly inflecting group. It is possible to define redundant paradigms in the OSLIN paradigm system, but they have to be forced upon the system, since computationally, there is no reason for their existence. However, the system will not always be able to determine which of the paradigms is the intended one in a given word.

It is common to say that certain words follow more than one paradigm: the Dutch word *leraar* (teacher) is often said to inflect both like *blaar/blaren* (blister) and like *makelaar/makelaars* (house broker). Although it is not impossible to implement this computationally, OSLIN follows the more straightforward method to define a third paradigm for *leraar* in such cases, which is a paradigm that allows both plurals.

Computationally, there is a class of nouns in Portuguese that (at least in some sources) have a *–y* in the singular and *–ies* in their plural: *husky, caddy, body,* etc. These are all (English) loanwords since, until recently, the *y* was not even part of the Portuguese alphabet. Lexicographically, it looks odd to say that there is a paradigm for *husky* in Portuguese. For such cases, it is possible in OSLIN to not assign a word a paradigm at all, but only provide it with manually entered inflected forms.

The most problematic case of mismatch are those cases where for a human, two words clearly inflect the same, whereas for a computer, they do not. An example is the Catalan verb *prevenir* (to prevent). It is a prefixed version of the verb *venir* (to come), and hence inflects like it, as do *sobrevenir* (to overcome) and several other verbs. The third person singular present indicative of *prevenir* is *prevé*, with a accent on the last *é* to indicate the stressed syllable. This stress mark is present in all prefixed verbs from *venir* but it is *not* present in the verb *venir* itself. Since the same form for *venir* (*ve*) is monosyllabic, there is no need for the stress mark. Although it is not fully impossible to define a set of transformation rules that correctly inflect both *prevenir* and *venir*, it is very awkward at best, and definitely not something that can be achieved automatically, or manually by someone without sufficient computational know-how.

5.2 Compounds and Paradigms

In a paradigm-based framework, especially one using string-transformation rules as in the case of OSLIN, inflection is mostly taking place at the beginning and/or the end of the word. For that reason, words where inflection does take place in the middle of the word are problematic.

Infixing inflection is for simple words is not common, but it is much more common to find word-internal inflection in the case of compounds. For instance, hyphenated nominal compounds in Portuguese can pluralize on the left, on the right, or on both part: the plural of *guarda-chuva* (umbrella) is *guarda-chuvas,* whereas the plural of *guarda-nocturno* (night guard) is *guardas-nocturnos*. The same holds for multi-word expressions in English.

In such cases, it would be possible to use a string-transformation rule to place an *–s–* before the hyphen in the paradigm. But the problem with that solution is that if the compound is left-inflecting, it does not necessarily pluralize with an *s*, but can pluralize like any normal noun. Therefore, the OSLIN system can assigns such compounds two paradigms: one for the left part, and one for the right part. For instance, the paradigm SUB01[-]SUB01 can be used for the case of *guarda-nocturno*: it indicates a nominal compound, which inflects on the left and the right, where the two parts are separated by a hyphen, and that the left part inflects via the first nominal paradigm, and the right part via the first nominal paradigm as well.

However, the solution of multiple paradigms relies on the fact that there is a graphical indication what the left and the right part of the compound are. Fortunately, languages have a tendency to avoid left-inflecting compounds where no such indication is present, but they do exist. An example is the Spanish word *hijodalgo* (gentleman) which is morphologically a compound (*hijo de algo* – son of somebody) where the left part is inflecing: *hijosdalgo*. In such cases, it is impossible to automatically determine from the citation form where the plural *s* should be inserted.

In Dutch and German, there is a much larger, well-known class of compounds that are problematic in the same way as left-inflecting non-separated compounds: prefixed separable verbs. The past tense of the Dutch verb *overgeven* (to vomit) is *gaf over*, and the past participle is *overgegeven*. In these two forms, the first component of the verb is separated from the rest, either by displacement, or by the insertion of inflectional material can be inserted between the two parts. Getting the inflected forms correct for separable verbs in a rule-based system is always complicated, but solutions have been implemented in the past (see for instance ten Hacken & Bopp 1998), and these solutions can be implanted in terms over string-transformation rules as well. However, such solutions always rely on a manual indication of the prefix. Although most verbal prefixes are prepositions, there are also verbal compounds with adverbs (*weglopen,* to walk away), or even noun (*brandstichten*, to commit arson), and there is no way to reliably predict which part of the verb will be the prefix.

There are only two solutions in the case of compounds without an explicit indication. The first is to resort to manual inflection for such cases, which is the solution most often used in OSLIN. However, it is possible to manually insert a dummy-separator: by changing the input to the paradigm system from *hijodalgo* to *hijo#dalgo*, and from *weglopen* to *weg#lopen*, it becomes possible to use the multiple paradigm assignment for compounds as described above.

5.3 Defectiveness and Clitics

Defective paradigms, such as impersonal verbs, are verbs that lack certain inflected forms. Such verbs can be straightforwardly dealt with in terms of normal paradigms, where the paradigm itself misses a number of forms. There are, however, two problems with such an approach. Firstly, impersonal verbs *can* typically be used in the defective forms when the verb is used metaphorically. And secondly, the forms that do exist can follow any of the existing paradigms. This means that there is not just the need for one additional defective paradigm, but that theoretically, every paradigm would need a defective counterpart.

To solve both problems at the same time, OSLIN uses meta-paradigms: an impersonal verb like *atardecer* (to get dark) is assigned a normal paradigm, in this case it inflects like *crecer* (to grow). On top of that, it is assigned a defective paradigm, which specifies which forms do and do not exist. There can be various defective paradigms per word class if needed.

The defective paradigms make a distinction between two different types of defectiveness. On the one hand, defectiveness due to semantic restrictions, which can typically be overruled in metaphoric uses of the word. Such forms are shown in the web-interface, but grayed out. On the other hand, thre are cases where the defectivity is due to normative considerations, as in the case of the so-called *euphonic defective verbs,* where the defective forms are never (normatively) acceptable, not even in metaphoric use or otherwise. Such forms are stored, but in principle completely hidden in the web-interface.

Not only defective paradigms can be treated by meta-paradigms, but also clitics in the inflection, as for instance in the case of pronominal verbs in Portuguese or Spanish. In a system based on string-transformation, a pronominal verb like *aburrarse* (to get bored) would need a special paradigm, meaning that as in the case of defective verbs, all paradigms would need to be duplicated. In the OSLIN system, the verb *aburrarse* is inflected like *amar*, and a meta-paradigm is used to add the pronominal clitics in the right forms.

6. Conclusion

As shown in this article, it is possible to have a computational tool for the semi-automatic inflection of the lexicon, where the lexicographer has all the freedom he needs to provide high-grade inflectional data, while at the same time being guided and helped along by the computational tool. With the inflectional tools provided by the OSLIN framework, it is possible to generate large full-scale, lexicographically controlled full-form lexicons within a reasonable amount of time.

Because the system is language independent, and furthermore allows lexicographers to create and apply the paradigm system for a new language, the OSLIN paradigm tool is particularly useful for less-resources languages.

Inflection in dictionaries is an often-underestimated topic: it is often considered a trivial task that can easily be achieved by computational means. This article only mentions problems that have to do with the creation of inflected forms by means of an inflectional paradigm. But there are many other problems that are beyond the scope of this article: how to establish what the correct inflected forms are, how to deal with the inflection of loanwords, when to consider a word to have a defective paradigm, etc. Although the OSLIN tools do not by themselves solve any of these issues, they do provide a platform in which the lexicographer has the option to implement his solutions for these issues.

7. References

Baptista Xuriguera, J. (2009). *Els Verbs Conjugats*. Barcelona: Claret.

Academia de la Llingua Asturiana (2005). *Normes Ortográfiques*, ed. 6. Oviedo: ÁPEL.

ten Hacken, P., Bopp, S. (1998). Separable Verbs in a Reusable Morphological Dictionary for German. In: *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pp. 471-475

Janssen, M. (2005). Open Source Lexical Information Network. In: *Proceedings of the Third International Workshop on Generative Approaches to the Lexicon*. Geneva.

Koskenniemi, K. (1983). *Two Level Morphology: A general computational model for word-form recognition and production*. PhD Thesis, University of Helsinki.

Matthews, P.H. (1972). *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*. Cambridge: Cambridge University Press.

Real Academia Española. (2001) *Diccionario de la Lengua Española,* ed. 22. Madrid: Espasa-Calpe.

Proceedings of eLex 2011, pp. 106-112

Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary

Madis Jürviste, Jelena Kallas, Margit Langemets, Maria Tuulik, Ülle Viks

Institute of the Estonian Language
Tallinn, Estonia
E-mail: madis.jyrviste@eki.ee, jelena.kallas@eki.ee, margit.langemets@eki.ee, maria.tuulik@eki.ee, ylle.viks@eki.ee

Abstract

This paper introduces new functions of the EELex dictionary writing system, developed at the Institute of the Estonian Language. Recently the EELex system has gained many new functions and query possibilities: scheme editor; article preview generator; bulk corrections; testing cross-references; generating Estonian morphological information; displaying the complete data or only chosen types of data; adding image, video and audio files; exporting dictionary data to Word format. By the example of the corpus-based active Basic Estonian Dictionary we describe the implementation of new EELex features for compiling, editing and presenting dictionary data. The dictionary is being compiled for Estonian language learners at the beginner and lower-intermediate levels and will be published in 2013 in both paper and electronic versions. In addition, the Estonian module of the corpus query system Sketch Engine (Kilgarriff et al., 2004) for the extraction and presentation of government and collocation patterns will be illustrated.

Keywords: dictionary writing system; corpus query system; learner lexicography

1. Introduction

EELex (Langemets at al., 2010, henceforth EELex)[1] is a web-based dictionary writing system for compiling, editing and presenting dictionary data, allowing simple and advanced structure-based queries and the sorting of query results; in addition, EELex offers possibilities for group working and has Estonian language support. Altogether nearly 30 dictionaries of various types and different structures have been compiled using the EELex system: monolingual and bilingual dictionaries, terminological and grammatical databases, etc. All dictionaries compiled in the EELex system have a standard XML-markup, which makes EELex a multi-purpose language resource that can be used by lexicographers, terminographers and ordinary end-users.

With the EELex system we intend to create a more efficient bridge between the IT domain and lexicography. The problem of modern lexicography seems to be that only a very small number of the huge variety of possibilities offered by the recent developments in information technology are actually put into practice: the great majority of new e-dictionaries are in substance the same old traditional dictionaries that are simply displayed on an electronical support platform. A veritably modern dictionary should be able to implement in practice the various solutions offered by technology, from corpus-based compilation methods up to real comfort of use (of the final product). Taking into account the possibilites of hypertext functionality one should avoid strictly linear text restrictions, e.g. preferring semantic criteria in entry compilation to use of strict alphabetic order, but also and foremost offering the end-user a user interface that would surpass these restrictions (cf. Atkins, 2002: 11). As electronic dictionaries long ago passed the stage of physical data storage capacity limits, the only real 'capacity' limit to consider is the amount of data shown to the dictionary user on the screen: the user should not be overloaded with too much data (for example a very long and complex article containing various types of data entirely displayed at once in its full form). Rather, the user has to be given the possibility of easily finding and exploring these large amounts of data.

The Basic Estonian Dictionary is taking steps towards a solution of this kind. The conditions within which to achieve such a result have been created: the dictionary is being compiled using an advanced dictionary writing system (EELex), and we use comprehensive Estonian language corpora via the Sketch Engine corpus query system (Kilgarrif et al., 2004)[2]. To the end-user of the e-dictionary we plan to offer a solution thoroughly different from a traditional print dictionary format. In the following chapters we would like to introduce these two tools, EELex and Sketch Engine for Estonian, in more detail, by using the example of the Basic Estonian Dictionary.

2. Specialised Features of the EELex Dictionary Writing System

Even though the main functions of EELex were ready by 2010 (Langemets et al., 2010; Viks et al., 2010), the system is being continuously developed. There are several new functions in the areas of dictionary compilation and customisation, editing and publishing; it is also possible to make global queries from different EELex-based dictionaries.

2.1 Dictionary compilation and customisation

The possibilities for dictionary compilation have been

[1] http://eelex.eki.ee/ (20.09.2011).

[2] http://www.sketchengine.co.uk/ (20.09.2011).

improved. EELex also has new functions such as scheme editor, user management interface and article preview generator.

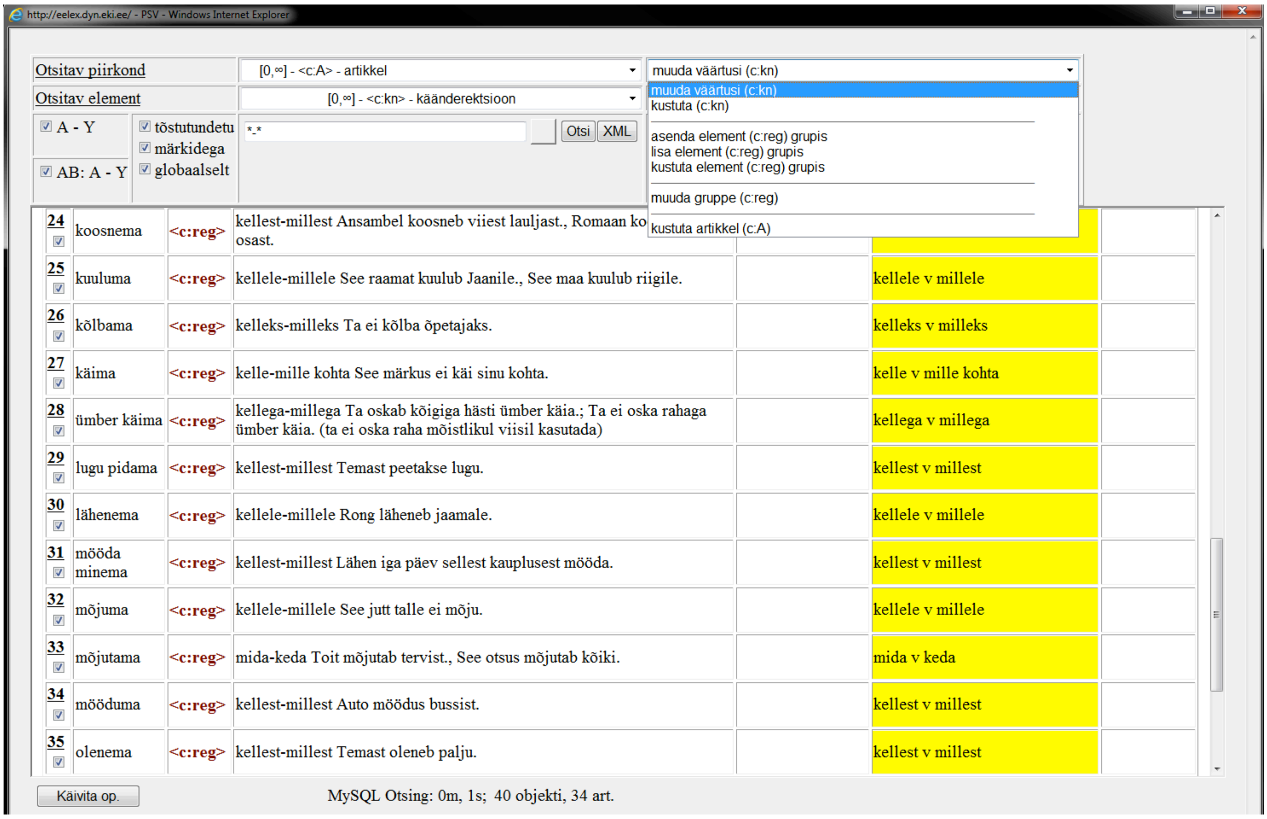
The scheme editor allows the user to customise the XML-structure of the dictionary. It is possible to add and delete elements or attributes, as well as to change the labels and properties of existing elements and attributes.

With the user management interface, the editor can assign different rights to different working group members.

With the article preview generator, the user can modify the dictionary entry preview. The user can set the font, size, colour and background of each element; set a character, text or line break between, in front of or after a specific element or group of elements; show or hide specific elements in the article editing preview and print preview; assign conditions for element display (according to the value of the attribute or neighbouring elements); and assign a hyperlink to an element.

2.2. Dictionary editing

Dictionary editing is made easier by new functions such as bulk corrections, automatic generation of morphological information, use of multimedia data and list data manager. Bulk corrections (see Figure 1) allow simultaneous corrections in all entries of a certain type (entries that do not correspond to the exact criteria can be excluded separately).



http://eelex.dyn.eki.ee/ - PSV - Windows Internet Explorer

Otsitav piirkond [0,∞] - <c:A> - artikkel

Otsitav element [0,∞] - <c:kn> - käänderektsioon

☑ A - Y

☑ AB: A - Y

☑ tõstutundetu

☑ märkidega

☑ globaalselt

_

24 ☑	koosnema	<c:reg>	kellest-millest Ansambel koosneb viiest lauljast., Romaan ko osast.			
25 ☑	kuuluma	<c:reg>	kellele-millele See raamat kuulub Jaanile., See maa kuulub riigile.		kellele v millele	
26 ☑	kõlbama	<c:reg>	kelleks-milleks Ta ei kõlba õpetajaks.		kelleks v milleks	
27 ☑	käima	<c:reg>	kelle-mille kohta See märkus ei käi sinu kohta.		kelle v mille kohta	
28 ☑	ümber käima	<c:reg>	kellega-millega Ta oskab kõigiga hästi ümber käia.; Ta ei oska rahaga ümber käia. (ta ei oska raha mõistlikul viisil kasutada)		kellega v millega	
29 ☑	lugu pidama	<c:reg>	kellest-millest Temast peetakse lugu.		kellest v millest	
30 ☑	lähenema	<c:reg>	kellele-millele Rong läheneb jaamale.		kellele v millele	
31 ☑	mööda minema	<c:reg>	kellest-millest Lähen iga päev sellest kauplusest mööda.		kellest v millest	
32 ☑	mõjuma	<c:reg>	kellele-millele See jutt talle ei mõju.		kellele v millele	
33 ☑	mõjutama	<c:reg>	mida-keda Toit mõjutab tervist., See otsus mõjutab kõiki.		mida v keda	
34 ☑	mööduma	<c:reg>	kellest-millest Auto möödus bussist.		kellest v millest	
35 ☑	olenema	<c:reg>	kellest-millest Temast oleneb palju.		kellest v millest	

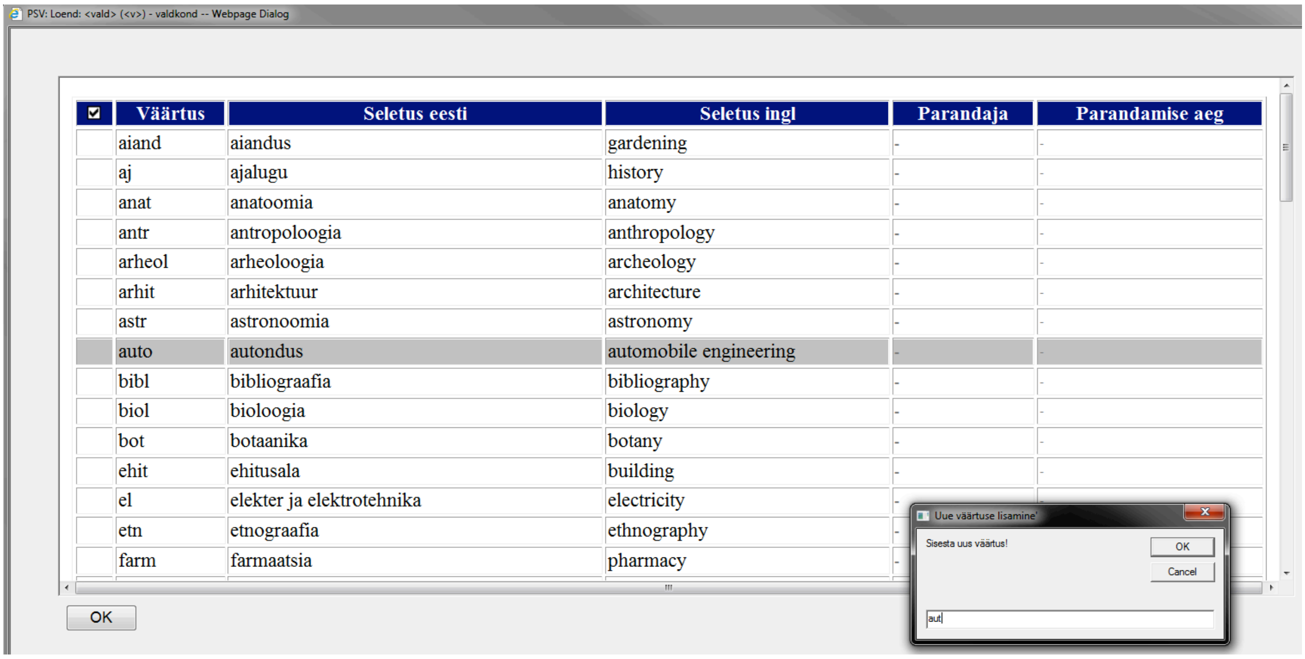
MySQL Otsing: 0m, 1s; 40 objekti, 34 art.

Figure 1: EELex editing: bulk corrections

Automatic generation of morphological information for the entries is possible in two modes: global and local. In global mode, the morphological data is added simultaneously to all entries; homonyms get several inputs, which requires obligatory postediting. In local mode the data is added through a dialogue and the editor can solve problems directly during the working process.

As for multimedia data, audio, video and image files can be inserted into an entry and opened directly.

The list data manager (see Figure 2) allows the user to add and delete list elements like labels, attribute values, semantic types, and so forth.



PSV: Loend: <vald> (<v>) - valdkond -- Webpage Dialog

☑	Väärtus	Seletus eesti	Seletus ingl	Parandaja	Parandamise aeg
	aiand	aiandus	gardening	-	-
	aj	ajalugu	history	-	-
	anat	anatoomia	anatomy	-	-
	antr	antropoloogia	anthropology	-	-
	arheol	arheoloogia	archeology	-	-
	arhit	arhitektuur	architecture	-	-
	astr	astronoomia	astronomy	-	-
	auto	autondus	automobile engineering	-	-
	bibl	bibliograafia	bibliography	-	-
	biol	bioloogia	biology	-	-
	bot	botaanika	botany	-	-
	ehit	ehitusala	building	-	-
	el	elekter ja elektrotehnika	electricity	-	-
	etn	etnograafia	ethnography	-	
	farm	farmaatsia	pharmacy	-	

Uue väärtuse lisamine'

Sisesta uus väärtus!

aut

Figure 2: EELex editing: list data manager

2.3. Publishing

EELex allows dictionaries to be published in print as well as electronic versions.

To prepare a print dictionary, a file in MS Word format with a .doc extension is created, based on a specific query and for a defined set of entries. The entries in the .doc file have the same layout as in the editing window display preview. The query system allows the user to use the database of one particular dictionary to compile different versions of print dictionaries by selecting different structural elements to be shown in the printed version.

To allow electronic publishing of dictionaries, we host a web interface based on the XML format, allowing public access to published dictionaries[3]; lexicographers can also access dictionaries that are being compiled, but are not yet published, via the Institute's intranet. Users can search for information based on entry structure using the web interface; in addition to a simple query or a query from the whole entry, the user can restrict the search to a specific structure element, such as headwords, domain labels, examples, grammatical information, etc.

It is often difficult to easily find the relevant information in long entries. To display the entries in a concise manner, we have conceived a step-by-step display system: initially, only those parts of the entry that contain the queried information are displayed, although the user can also display other parts of the entry (e.g. the Dictionary of Estonian Word Families[4], Viks et al., 2011). We also plan to add the possibility to use structure filters, to allow multiple views, i.e. the display of only predefined structural elements of the entry, such as the headwords and definitions without examples.

2.4. Global queries

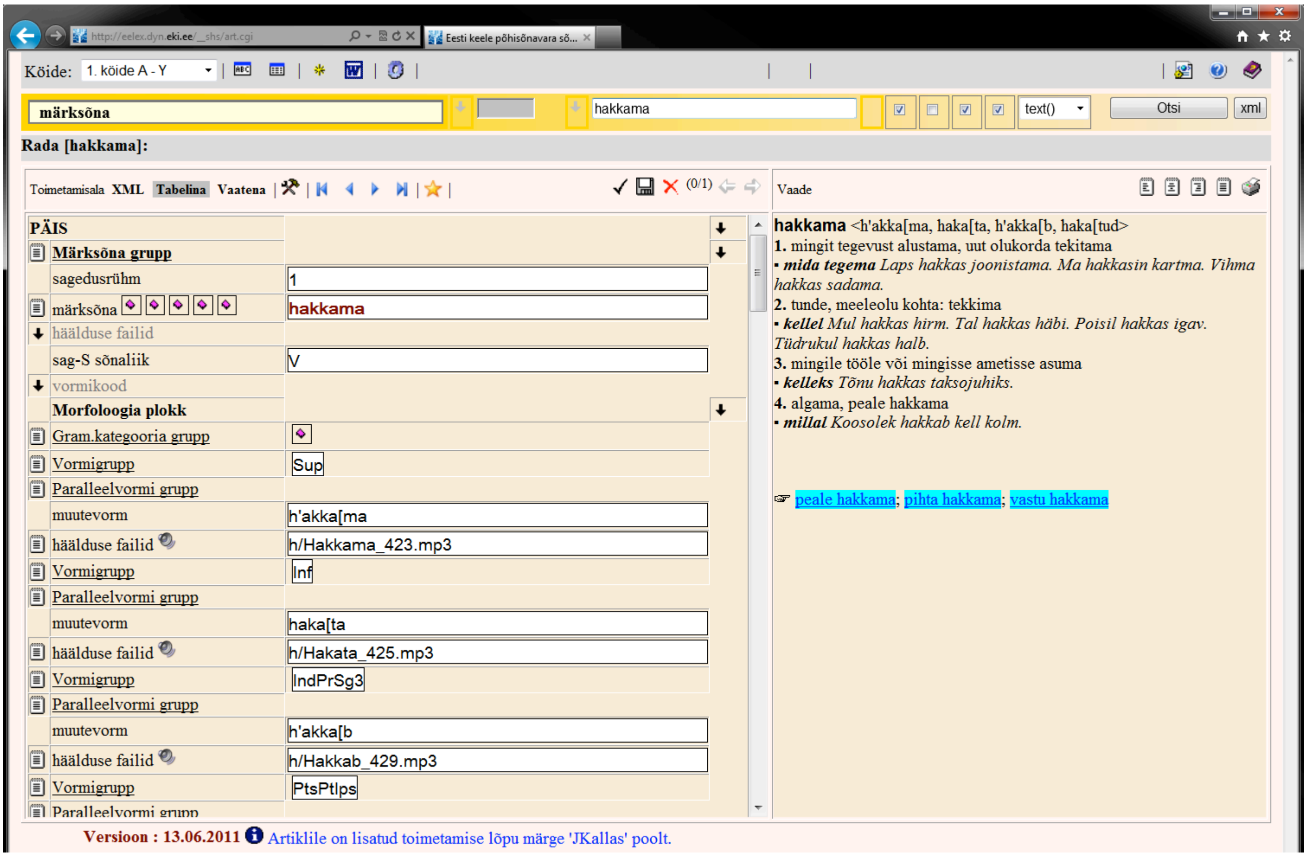
As the number of dictionaries in the EELex system is constantly increasing, a global query interface has been created for lexicographers, allowing a better overview of all dictionaries. The query allows the user to search for a headword in other EELex dictionaries or common keywords in different dictionaries.

3. The Basic Estonian Dictionary Database

The Basic Estonian Dictionary (henceforth BED) is an active dictionary, designed for learners of Estonian (Kallas, 2010; Kallas, Tuulik, 2011). It uses an XML entry scheme containing a detailed entry structure, which the editor-in-chief can complete and restructure if necessary, using the EELex scheme editor. Altogether the scheme has eleven blocks (pronunciation, inflectional formation, definition, word formation, government and collocation patterns, etc.) (see Figure 3).

[3] http://portaal.eki.ee/ (20.09.2011).

[4] http://www.eki.ee/dict/sp/ (20.09.2011).



Rada [hakkama]:

PÄIS	
Märksõna grupp	
sagedusrühm	1
märksõna	hakkama
hääldusе failid	
sag-S sõnaliik	V
vormikood	
Morfoloogia plokk	
Gram.kategooria grupp	
Vormigrupp	Sup
Paralleelvormi grupp	
muutevorm	h'akka[ma
hääldusе failid	h/Hakkama_423.mp3
Vormigrupp	Inf
Paralleelvormi grupp	
muutevorm	haka[ta
hääldusе failid	h/Hakata_425.mp3
Vormigrupp	IndPrSg3
Paralleelvormi grupp	
muutevorm	h'akka[b
hääldusе failid	h/Hakkab_429.mp3
Vormigrupp	PtsPtIps
Paralleelvormi grupp	

hakkama <h'akka[ma, haka[ta, h'akka[b, haka[tud>
1. mingit tegevust alustama, uut olukorda tekitama
▪ ***mida tegema*** *Laps hakkas joonistama. Ma hakkasin kartma. Vihma hakkas sadama.*
2. tunde, meeleolu kohta: tekkima
▪ ***kellel*** *Mul hakkas hirm. Tal hakkas häbi. Poisil hakkas igav. Tüdrukul hakkas halb.*
3. mingile tööle või mingisse ametisse asuma
▪ ***kelleks*** *Tõnu hakkas taksojuhiks.*
4. algama, peale hakkama
▪ ***millal*** *Koosolek hakkab kell kolm.*

☞ peale hakkama; pihta hakkama; vastu hakkama

Versioon : 13.06.2011 Artiklile on lisatud toimetamise lõpu märge 'JKallas' poolt.

Figure 3: EELex editing window: the table view

For the purpose of coherence the metadata of all entries contains information about the semantic type(s) of the word sense(s): the dictionary is being compiled and edited according to semantic types (for the semantic classification of nouns, cf. Langemets, 2010).

In the BED compilation process we use all the new editing functions offered by EELex: bulk corrections, automatic generating of morphological data, use of multimedia data, the list data manager. In the following paragraphs we would like to describe the elements that clearly show the differences between the print and the electronic version.

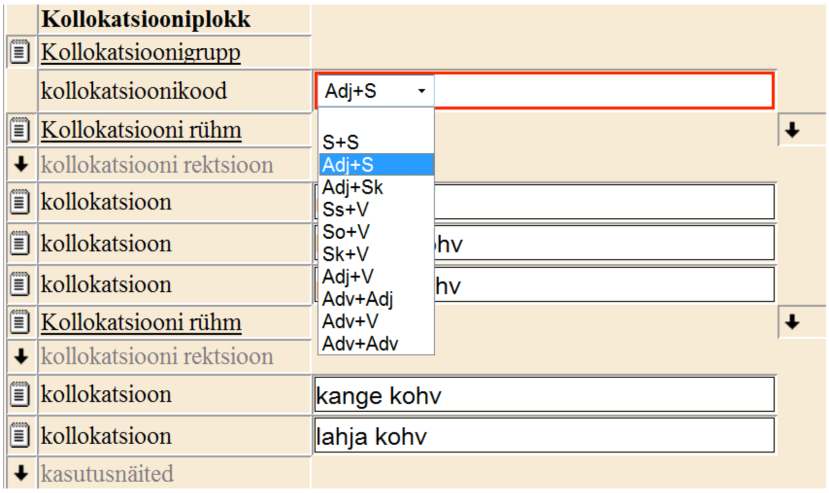
In the pronunciation block we present a sound recording (mp3 audio file) for the most important forms in addition to the graphical presentation of the headword pronunciation. Morphological information is generated automatically. The BED as a learner's dictionary uses a comprehensive form-based presentation of data. In the printed version we give only the minimal inflectional paradigm (i.e. only the main forms: for nominal words the three first cases in singular and plural, for verbs the primary main forms, with secundary main forms if necessary). On the other hand, in the e-version the user will have the option to open the word's full inflectional paradigm: for nominal words all 14 cases in singular and plural, and for verbs all the finite and non-finite forms of the personal voice as well as of the impersonal voice. In the BED, several cross-references are used to show links between words: in the word formation block we use links to show word formation relationships (words that can be formed on the basis of the headword); in the lexico-semantic relationship block links are used to show synonyms, antonyms and paronyms. All the cross-references can be checked with a specific tool in EELex. However, this is not an automatic calculation yet, like in TLex[5] where the updating of sense and homonym numbers is fully automated.

In the sign language block the BED has a video recording of the relevant sign. For every sign the BED contains information about the initial hand form (the handshape with which the sign is articulated; should the handshape change during the formation of the sign, only the initial hand form is shown), the location where the sign is articulated (i.e. face, lips, cheek, chest, neutral space, etc.) and the movement with which the sign is formed. Based on these three parameters it is possible to search for a certain sign choosing the hand form, the location, and the movement of the sign. This enables the deaf dictionary user to find the Estonian equivalent for a sign.

It is possible to add images to the entries. Images are presented both in the paper version as well as in the electronic version of the BED.

[5] http://tshwanedje.com/tshwanelex/ (20.09.2011).

As the BED is an active dictionary, the explicit presentation of syntagmatic relations is of the utmost importance. Government and collocation patterns are presented in separate blocks and grouped according to codes presented, in the form of a drop-down menu (all menus are created with the list data manager). Figure 4 shows the codes for collocational patterns in the BED.



Kollokatsiooniplokk
Kollokatsioonigrupp
kollokatsioonikood: Adj+S
Kollokatsiooni rühm
kollokatsiooni rektsioon
kollokatsioon
kollokatsioon: hv
kollokatsioon: hv
Kollokatsiooni rühm
kollokatsiooni rektsioon
kollokatsioon: kange kohv
kollokatsioon: lahja kohv
kasutusnäited

Figure 4: Codes for collocational patterns in the BED

The most frequent government and collocation patterns are analysed and selected using the Sketch Engine corpus query system.

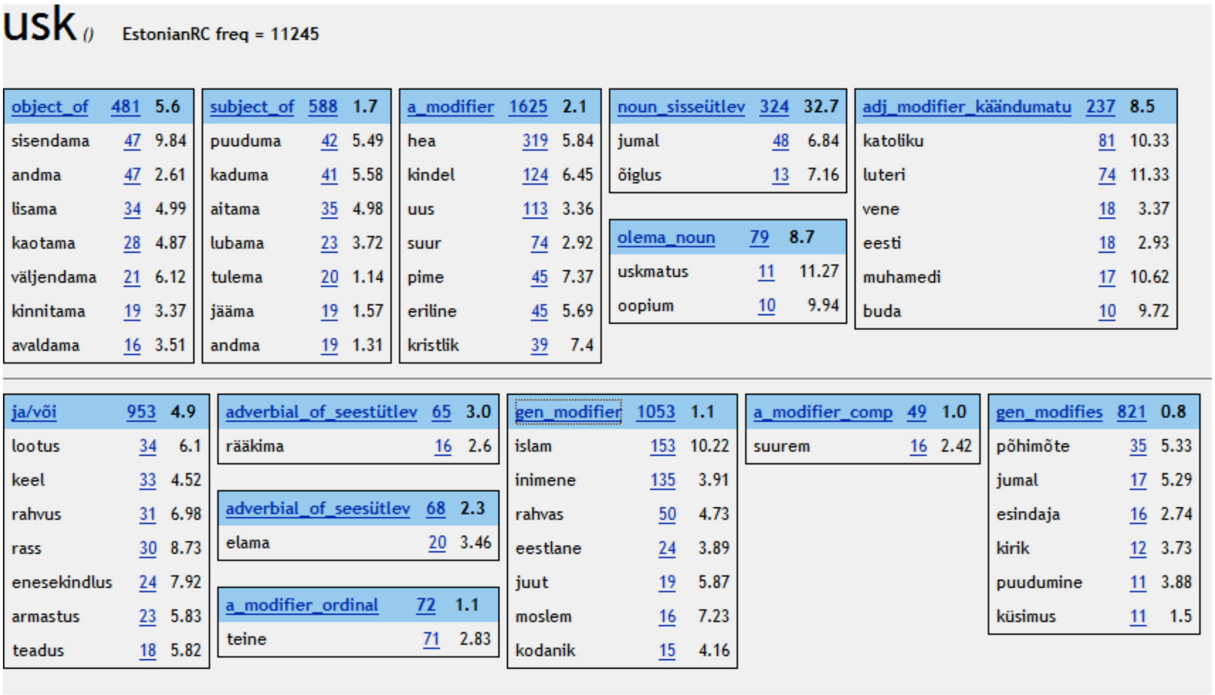
4. Sketch Engine for Estonian: extraction and presentation of government patterns and collocations

Sketch Engine (Kilgarriff et al., 2004) is a web-based corpus query system for several languages (including, among others, French, Spanish, Japanese, etc.). Since autumn 2010 it has been used at the Institute of the Estonian Language to compile two Estonian language dictionaries: the Basic Estonian Dictionary and the explanatory one-volume Dictionary of Estonian (to be published in 2015).

Sketch Engine for Estonian uses the Estonian Reference Corpus[6] of 250 million tokens as input. The corpus had previously been annotated morphologically, lemmatised, partially disambiguated, and annotated by clause by Filosoft LLC[7]. Due to the agglutinative structure of the Estonian language the annotation involves not only lemmata and morphological features (POS-tag), but also inflections, e.g. *majas* /S/maja/s/sg_in 'in the house'.

In order to identify the grammatical relations between words, the syntagmatic (syntactic and collocational) properties of proper nouns, verbs, multi-word verbs (phrasal, prepositional and phrasal-prepositional), adjectives, ordinal numerals, and adverbs were investigated. As a result, 38 grammatical relations for Estonian were defined, using regular expressions and query language IMS Corpus Workbench. The system searches for grammatical relations which correspond to POS-tags and morphological inflections (e.g. such categories as *subject, object, oblique objects, adverbials, modifiers, adverb*); constructions with conjunctions *ja/või* 'and/or', *kui/nagu* 'as'; predicative (complements of the copula-like verb *olema* 'be'); various combinations of finite verbs with non-finite verbs; oblique objects and adverbials of particle verbs, prepositional verbs and noun prepositional phrases. All possible syntactic government patterns are brought forth: the system shows explicitly all possible case, adposition and infinitive government patterns for substantives and adjectives; object, case, adposition, and infinitive government for verbs; case government for particle verbs; adposition government for prepositional verbs and case government for adverbs. Figure 5 shows the word sketch for noun *usk*, 'faith'.

[6] http://www.cl.ut.ee/korpused/segakorpus/ (20.09.2011).
[7] http://www.filosoft.ee/index_en.html (20.09.2011)



usk () EstonianRC freq = 11245

object_of	481	5.6
sisendama	47	9.84
andma	47	2.61
lisama	34	4.99
kaotama	28	4.87
väljendama	21	6.12
kinnitama	19	3.37
avaldama	16	3.51

subject_of	588	1.7
puuduma	42	5.49
kaduma	41	5.58
aitama	35	4.98
lubama	23	3.72
tulema	20	1.14
jääma	19	1.57
andma	19	1.31

a_modifier	1625	2.1
hea	319	5.84
kindel	124	6.45
uus	113	3.36
suur	74	2.92
pime	45	7.37
eriline	45	5.69
kristlik	39	7.4

noun_sisseütlev	324	32.7
jumal	48	6.84
õiglus	13	7.16

olema_noun	79	8.7
uskmatus	11	11.27
oopium	10	9.94

adj_modifier_käändumatu	237	8.5
katoliku	81	10.33
luteri	74	11.33
vene	18	3.37
eesti	18	2.93
muhamedi	17	10.62
buda	10	9.72

ja/või	953	4.9
lootus	34	6.1
keel	33	4.52
rahvus	31	6.98
rass	30	8.73
enesekindlus	24	7.92
armastus	23	5.83
teadus	18	5.82

adverbial_of_seestütlev	65	3.0
rääkima	16	2.6

adverbial_of_seesütlev	68	2.3
elama	20	3.46

a_modifier_ordinal	72	1.1
teine	71	2.83

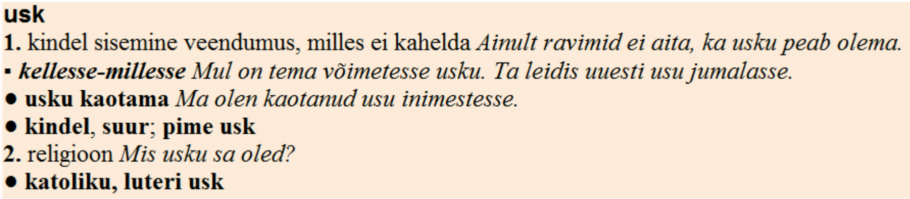
gen_modifier	1053	1.1
islam	153	10.22
inimene	135	3.91
rahvas	50	4.73
eestlane	24	3.89
juut	19	5.87
moslem	16	7.23
kodanik	15	4.16

a_modifier_comp	49	1.0
suurem	16	2.42

gen_modifies	821	0.8
põhimõte	35	5.33
jumal	17	5.29
esindaja	16	2.74
kirik	12	3.73
puudumine	11	3.88
küsimus	11	1.5

Figure 5: Word sketch for *usk*, 'faith'

The word sketch of the noun *usk*, 'faith', reveals the following patterns: noun (as subject) + verb (e.g. *usk puudub* 'to have no faith'), noun (as object) + verb (e.g. *usku sisendama* 'to inspire faith'), adjective + noun (e.g. *hea usk* 'good faith'; *katoliku usk* 'catholic faith'), coordinated nouns (e.g. *usk ja lootus* 'faith and hope'), and case government pattern (*usk* [kellesse-millesse] 'faith [in] someone or something'), etc. The most frequent extracted patterns are included in the entry of this particular noun (see Figure 6).



usk
1. kindel sisemine veendumus, milles ei kahelda *Ainult ravimid ei aita, ka usku peab olema.*
▪ ***kellesse-millesse*** *Mul on tema võimetesse usku. Ta leidis uuesti usu jumalasse.*
● **usku kaotama** *Ma olen kaotanud usu inimestesse.*
● **kindel**, **suur**; **pime usk**
2. religioon *Mis usku sa oled?*
● **katoliku, luteri usk**

Figure 6: BED entry for noun *usk*, 'faith'

The dictionary entry for *usk* presents information about the most frequent government patterns (*usk* [kellesse-millesse]) and collocations (*usku kaotama*; *kindel, suur, pime usk*; *katoliku, luteri usk*).

The quality of the word sketches depends on the quality of the morphological disambiguation. Errors in the output are mainly caused by errors or shortcomings in the morphological annotation. Secondly, the content of the corpus input should be balanced. At the moment (September 2011) 75 per cent of the texts in the corpus represent the media, i.e. journals and newspapers, fiction being significantly under-represented in the corpus. We plan to upload the new version of the Estonian reference corpus into Sketch Engine in January 2012. This new version will contain more texts and additional information, including syntactic tags and data about text types.

When choosing example sentences for the BED we use two Sketch Engine functions: Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) and Tickbox Lexicography Template. In the Estonian module we use the language-independent *vanilla*-version in which the selection of examples depends on the length of the sentence (usually 5–10 words), initial capital letter and punctuation marks. At present, the data has to be manually selected in the Word Sketch and then manually copied into the dictionary. In the future we are planning to create a link between EELex and Sketch Engine to provide the possibility of direct transfer of corpus example sentences into dictionary entries.

5. Conclusion and perspectives

The dictionary writing system EELex has several new functions at different levels: set-up and customisation of a dictionary project (scheme editor, user management interface, article preview generator), editing a dictionary (bulk corrections, automatic generating of morphological data, multimedia data input option, list data manager), electronic publishing (structure-based queries, step-by-step entry data display) and global query interface.

XML-based compilation will allow us to display content for the end-user in the web interface in layers, so that users themselves can choose what information will be displayed on the screen. For example, one might choose to display full entries, entries without the morphological information and/or pronunciation, and whether or not to display government patterns, usage examples, etc.

Moreover, using the EELex global query function, the user will have the option to search for a word in other dictionaries compiled using this system (for example explanatory dictionaries, bilingual dictionaries, etc.).

Being compiled in the EELex system, the BED project is a single database that contains all dictionary related data. This database allows the generation of different outputs: for example a (static) print dictionary or a (dynamic) e-dictionary, as well as specialised dictionaries based on partial database output.

6. Acknowledgements

This project has been supported by research grant SF0050023s09 (‘Modelling Intermodular Phenomena in Estonian’), National Programme ‘Estonian Language and Cultural Memory (2009–2013)’ (‘The Basic Estonian Dictionary’, 2010–2013) as well as the National Programme for Estonian Language Technology (2011–2017) (‘The Modifying of the Lexicographer's Workbench’, 2011–2012).

7. References

Atkins, B.T.S. (2002). Bilingual Dictionaries – Past, Present and Future. In M. Corréard (ed.) *Lexicography and Natural Language Processing: a Festschrift in honour of B. T. S. Atkins.* EURALEX, pp. 1–29.

Kallas, J. (2010). The development of scholary lexicography of the Estonian language as a Second Language in a historical and a theoretical perspective. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress.* Leeuwarden: Fryske Academy, pp. 648–651.

Kallas, J., Tuulik, M. (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. In *Eesti Rakenduslingvistika Ühingu aastaraamat 7.* Tallinn: Eesti Rakenduslingvistika Ühing, pp. 59–75.

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the XI Euralex International Congress.* Lorient: Université de Bretagne Sud, pp. 105–116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress.* Barcelona: Universitat Pompeu Fabra, pp. 425–432.

Langemets, M. (2010). Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras. Tallinn: Eesti Keele Sihtasutus.

Langemets, M., Loopmann, A. & Viks, Ü. (2010). Dictionary management system for bilingual dictionaries. In S. Granger, M. Paquot (eds.) *eLexicography in the 21st century: New challenges, new applications*. *Proceedings of eLex 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, pp. 425–430.

Viks, Ü., Vare, S. & Sahkai, H. (2010). The Database of Estonian Word Families: a Language Technology Resource. In I. Skadina, A. Vasiljevs (eds.) *Human Language Technologies. The Baltic Perspective. Proceedings of the Fourth International Conference, Baltic HLT 2010*. Amsterdam: IOS Press, pp. 169–176.

From Dictionary to Database: Creating a Global Multi-Language Series

Ilan Kernerman
K Dictionaries Ltd
Nahum 8, Tel Aviv 63503, Israel
E-mail: ilan@kdictionaries.com

Abstract

K Dictionaries is developing a global dictionary series for learners and general users, which includes over twenty languages to date. Each language core consists of a database that is used for creating monolingual dictionaries, and for adding translations and making bilingual dictionaries and eventually multilingual ones. A single macrostructure is used for the entire series, and the same microstructure is applied for the entries of all the dictionaries – while being adapted to suit the characteristics of each language. The cores can be modified in relation to any language pair and different components extracted for particular audiences in print editions and digital applications. The data is formatted in XML and Unicode, and the DTD has undergone numerous amendments and improvements over the time to solidify the structure and enrich it. The work is done by lexicographers worldwide using a dedicated editing software, whether offline or online, aided by editorial styleguides, technical manuals and full support from the K Dictionaries team. Further use will include data reversal, linking translations to their own language cores and enhancing multilingual web linkage.

Keywords: dictionary; database; content; technology; application

1. Introduction

In 2005, K Dictionaries (KD) began working on eight bilingual dictionaries for French learners of foreign languages[1]. We started by developing a monolingual dictionary database for French and for each of the eight languages, and then proceeded to translate the French core to the eight languages and these languages to French. The process took four years and the dictionaries were published in France in 2009, with each print edition accompanied by an electronic version for PC.

During that period more language cores and language pairs were compiled, and the French project has become the cornerstone of a global series consisting at present of over twenty languages[2], including forty bilingual titles, twenty in preparation and twenty more currently in planning. Over the last six years the entry structure has evolved and some of the language cores were expanded. The dictionaries are gradually released worldwide in different forms for all types of media.

To enable the creation of such extensive, complex and multilingual data, as well as its efficient processing, maintenance, updating and application, it was necessary to establish and nourish a fine technological infrastructure, which is extremely robust and solid on the one hand and open and flexible on the other hand. From the start we decided to encode the data in XML format and use Unicode to enable work on practically any language[3]. An XML Editor was configured according to the entry structure and provided to the lexicographers for their compilation. The DTD took several months to devise and continued to be modified all along to improve its microstructure and accommodate additions and changes. The data can thus be used with adaptable XSL documents to prepare for print and various digital versions. A desktop application was developed to produce an electronic version of each dictionary, and eventually online versions were conceived as well.

To our knowledge, this is a first such systematic attempt to develop a worldwide dictionary series for language learners and general users on such a relatively broad scale. Our editorial and technical concepts have evolved in the making, many errors were made and changes introduced. Some of our lessons may prove valuable to others as well.

2. Content

On the outset, the main target users of the initial French bilingual dictionary series were identified as French native speakers learning a foreign language at lower to intermediate levels. The publisher, Assimil, who is a leader in foreign language teaching materials in France, was keen on offering this added value to users of its *Méthod Assimil* coursebooks. At the same time, these dictionaries had to stand on their own and satisfy other users too.

It was decided that the dictionaries would be

[1] Arabic, Chinese, Greek, Japanese, Polish, Portuguese, Russian, Turkish.

[2] The cores currently available include Arabic, Chinese (Simplified, Traditional), Czech, Dutch, English, French, German, Greek, Italian, Japanese, Korean, Latin, Norwegian, Polish, Portuguese (Brazil, Portugal), Russian, Spanish, Thai and Turkish; in addition, Hebrew, Hindi and Swedish are due.

[3] XML (Extensible Markup Language) is a set of rules for encoding documents in machine-readable format. The DTD (Document Type Definition) is a set of markup declarations that define a document type for XML, and the XSL (Extensible Stylesheet Language) is used to transform and render XML documents. Unicode is a standard for consistent text encoding, representation and handling in most writing systems. (Definitions from Wikipedia, http://wikipedia.org/)

medium-size and bi-directional, catering for both encoding and decoding purposes. To satisfy Assimil’s needs, they had to cover the vocabulary in their *Méthod* for each language. They were going to have up to 680 pages, but when their data proved to be growing size this limit was first extended to 800 pages and finally the average page number is about 1,000. The print edition was going to be accompanied by an electronic version on CD-ROM.

We began by preparing a sort of monolingual learner’s dictionary core for each language, which could function as a base for adding translations and developing bilingual dictionaries. The entries include a brief definition for each sense, and examples often consist of short phrases rather than full sentences. The defining vocabulary consists of all the headwords in the dictionary and their inflections.

Generally-speaking, developing the L1 database puts into play a *deconstruction* of the language, mapping its ‘atoms and quarks’, and enabling its *reconstruction* in a variety of lexicographic terms. Then, in the translation process, a brand new equilibrium is reset concerning each specific language pair in play.

An important editorial change occurred half-way through the compilation, in the form of replacing the definition in French entries by sense indicators. The reasoning was to facilitate the use for French users and focus solely on their need for disambiguation of the different meanings of polysemous French words. Figure 1 demonstrates the French definitions in the first occurrence substituted by sense indicators in the second (both underlined):

accueil [akœj] *nm*
1 manière de recevoir qqn ou qqch
◊ *faire bon / mauvais accueil à qqn*
2 lieu où l’on reçoit des visiteurs
◊ *Adressez-vous à l’accueil !*
accueil [akœj] *nm*
1 réception ◊ *faire bon / mauvais accueil à qqn*
2 lieu ◊ *Adressez-vous à l’accueil !*

Figure 1: Replacing definitions by sense indicators

To save on space in the print editions, the definitions were later removed altogether from the entries of the other languages, but were kept in the electronic versions. The translation consists chiefly of providing L2 equivalents for each sense, example and phrase, often including pronunciation and grammatical details. The translators may modify the L1 entry to suit it to the L2, particularly by changing examples of usage, and sometimes they also suggest substantial changes in the L1 entry structure, such as concerning the classification of senses, that can help to improve it. Figure 2 demonstrates an alteration of the original example of usage from the Arabic version to the Russian one, to better suit the French-Russian pair:

auxiliaire *adj*
1 qui aide [mu'saːʔ’id] **مساعد**
◊ *personnel auxiliaire* شخص مساعد
auxiliaire *adj*
1 qui aide **вспомогáтельный** [fspəma’gatʲilʲnɨj], **подсóбный** [pat’sobnɨj]
◊ *économie auxiliaire* подсó бное хозя́йство

Figure 2: Modifying the example of usage

Basically, each language core consists of 12,000 main headwords[4], but some languages were doubled or quadrupled[5]. The list of headwords was devised primarily according to frequency and importance. The principle components of the entry are outlined in Table 1:

○ *the headword*:
L1 headword, pronunciation and alternative script[6], part of speech, grammatical gender and number, irregular forms
○ *the attributes*:
L1 geographical usage, subject field, register, sense qualifier, range of application, synonyms, antonyms, notes
○ *the sense indicator*:
L1 (when no attribute was indicated for a sense of a polysemous entry) a hyperonym or ‘preposition- filler’ (e.g. *of* something) – for disambiguation
○ *the definition*:
L1 a succinct definition for each sense
L2 translation for each sense of the headword (NOT of its definition), pronunciation, grammatical gender and number as appropriate
○ *the examples of usage*:
L1 example(s) of usage for each sense of polysemous entries, usually consisting of short phrases rather than full sentences
L2 idiomatic translation of the example(s)
○ *the compositional phrases*:
L1 collocational phrases (idioms, compounds, etc), with/without definition and/or example – may be part of a given sense or form a sense of its own
L2 idiomatic translation of the phrase (and its examples of usage)
○ *the sub-headwords*:
L1 run-ons usually consist of part of speech change and derivates, and may include part or all of main headword components
L2 translation of each L1 component, as above

Table 1: The entry’s main components

[4] Chinese has 3,000 main headwords consisting of single characters, and their derivations appear in the form of sub-entries

[5] Dutch, French, German, Italian and Portuguese have 25,000 main entries, Spanish has 50,000.

[6] The IPA (International Phonetic Alphabet) is used for most languages. Chinese includes Pinyin, Hebrew includes alternate script with ‘vowels’ (*nikud*), and Japanese includes Kanji, Katakana, Hiragana and Romaji.

The French titles were published by Assimil in 2009 (DAK), each accompanied by a desktop application that is downloaded from the publisher's server by using an individual code that is inserted in the inside back-cover, and installed on the user's PC.

In the meantime, work on more languages was launched, and the first dictionary to actually appear in the global series was Norwegian/Spanish in Norway, accompanied by the electronic version on CD-ROM (*Spansk ordbok*). The publisher, Vega, published the next title, Italian, in 2010 (*Italiensk ordbok*). German is published this autumn, to be followed by French and Polish in spring 2012.

The next countries in which titles are due to appear in print are Japan (2011) and Brazil (2012). Meanwhile, Dutch bilinguals have been available online since 2010 on http://mijnwoordenboek.nl/, and iPhone applications were released by Abbyy as part of their Lingvo series.

3. Technology

K Dictionaries operates worldwide on the development of lexicographic content, and its final dictionary products and services are released by others. In this spirit, since the turn of the century our cooperation has gradually shifted from traditional publishers to a wide range of technology firms, while our focus has increasingly centred on creating content that can be used and re-used, again and again, whether fully or partially, for any purpose and in any type of media. To attain the goal of offering high-quality content we must rely on high-technology. Moreover, we started to develop our own electronic applications, both to experiment with how the dictionaries might look like and to offer them to 'low-tech' partners from the publishing industry. As a result our company has become to some extent a *technology-based* (and *technology-oriented*) content provider.

As an aside, for a long time we have considered that all contemporary *lexicography* actually is (or should be) *e-lexicography*, so this distinction today is redundant (and not quite justifiable for the last quarter century, since the advent of corpus-based dictionaries[7]).

From the outset of this project we devoted close attention to the technology-related aspects of creating the data and putting it to play. Nevertheless, fresh insights and new issues kept appearing all along, leading to constant amendment of the editing software and the other tools used for our work.

Overall, it has become increasingly difficult and artificial to separate the content from the technology features of our work, and the division made in this paper is not fully precise and is mainly intended to serve its description purposes. The terms used herein are thus not always applied in their conventional sense. The following is a general overview of some of the main technology-derived dimensions of this project.

3.1 Editing Software

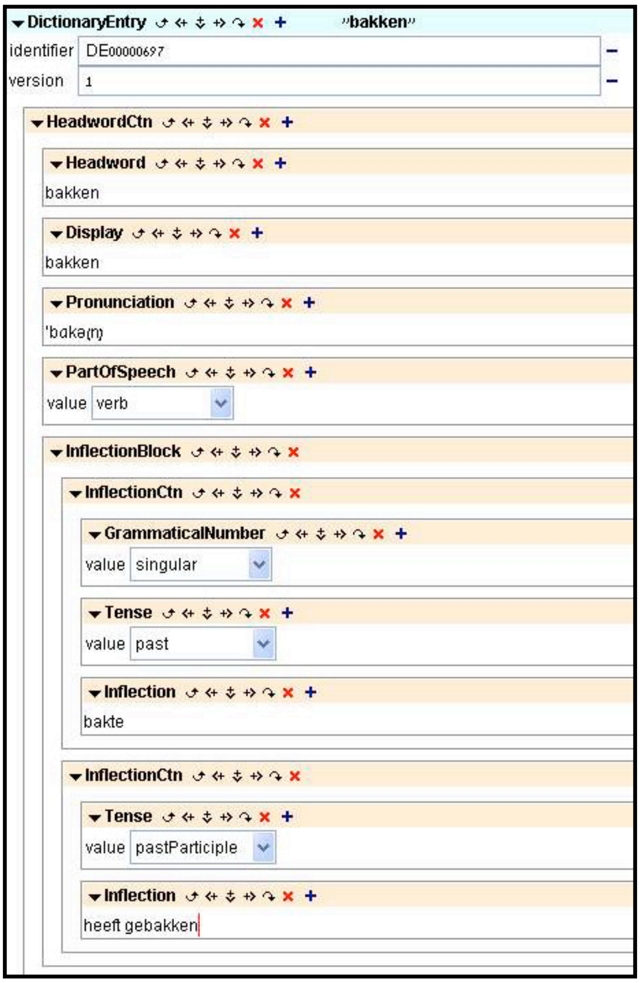
To maintain our independence and freedom of creation, as well as to help to reduce costs, we opted to use an XML editor to compile the data and to configure it for our purposes on our own, rather than utilize an existing dictionary writing system. However, it might come as no surprise to those familiar with this topic if we admit that after six years of experience it is still not absolutely clear which of these options is actually preferable and whether there indeed is any clear-cut answer to this question.

The editing software was developed for over half a year, but has continued to undergo endless updates since then. The main reasons for revising this software were:

- correction and improvement;
- adjustment to characteristics of new languages;
- adaptation to modifications in entry components;
- compatibility with changing operating systems.

The editing is done at a distance, usually from the editor's home, and the software can be used for either compiling L1 entries or adding L2 translation equivalents. The use is enabled either online or offline, and to work offline the editors must first install the software on their machine, which often requires personal adjustment to each individual computer and operating system. The software is accompanied by detailed documentation concerning its installation and manipulation, including a meticulous account of the entry microstructure. Whenever necessary, our staff provides full technical support for any need the lexicographer might have. Figure 3 demonstrates an extract of a Dutch entry in the XML Editor:

[7] COBUILD (1987) was the first dictionary to be entirely based on a computerized corpus.



```
DictionaryEntry    "bakken"
identifier DE00000697
version 1
  HeadwordCtn
    Headword
    bakken
    Display
    bakken
    Pronunciation
    'bakə(n)
    PartOfSpeech
    value verb
    InflectionBlock
      InflectionCtn
        GrammaticalNumber
        value singular
        Tense
        value past
        Inflection
        bakte
      InflectionCtn
        Tense
        value pastParticiple
        Inflection
        heeft gebakken
```

Figure 3: An extract of an entry in the XML Editor

3.2 Headword List

The chief editor of each language begins by preparing an editorial styleguide for compiling the entries and by drafting the list of main and sub-headwords. The headword list includes also the part of speech and cross-references, and for certain languages other elements are also indicated, such as alternative scripts. The headwords are selected according to frequency and importance, and will form the essence of the entries. Table 2 demonstrates an extract from a typical headword list.

The technology team then processes this headword information into 100 XML files, including 120 headwords each (i.e. 12,000 main entries in total), which will form the backbone for compiling the actual entries. In the compilation process it is possible to remove, add and change headwords from the initial list.

MAIN	RUNON	POS	XREF
antever		vt	
antiaéreo		adj	
antialérgico		n	
antialérgico	antialérgico	adj	
antibiótico		n	
anticaspa		adj	
anticoncepcional		adj	
anticoncepcional	anticoncepcional	n	
anticonceptivo			anticoncepcional
antidepressivo		n	

Table 2: An extract from the Portuguese (Portugal) headword list

3.3 Entry Microstructure

The editing software provides in advance for all possible components of the entry, many of which are going to be selected from a pre-defined dropdown menu. Elements that were not pre-defined in the software can be attributed as free values. For example, the word class for many languages will indicate the grammatical gender and grammatical number, as demonstrated in the DTD extracts in Figure 4:

```
<!ELEMENT GrammaticalGender EMPTY>
<!ATTLIST GrammaticalGender
  value notPredefined
  |masculine|feminine|masculine-feminine|
  neuter|masculine-neuter|masculine-feminine-n
  euter|feminine-masculine) #REQUIRED
  freeValue CDATA #IMPLIED
>
<!ELEMENT GrammaticalNumber EMPTY>
<!ATTLIST GrammaticalNumber
  modifier (only|usually) #IMPLIED
  value (notPredefined|singular|dual|plural)
  #REQUIRED
  freeValue CDATA #IMPLIED
>
```

Figure 4: Extracts from the DTD

3.4 Metalanguage & Localization

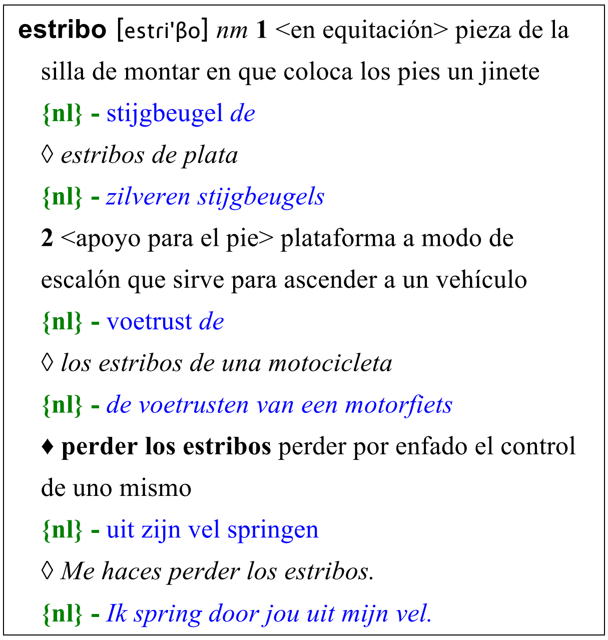
To facilitate the work on all different languages within a single macrostructure, all the labels used in the editing software appear in English, abbreviated. Their L1 equivalents are recorded in a separate list and will eventually replace the English labels in the final product. Other localization documents include in particular the transcription key for each language. Table 3 presents a brief extract from the key for Arabic:

IPA	(Sampa)	Unicode	EditorSet	Transcription	Translation	Arabic
Consonants						
Plosives						
b	b		Keyboard	baːb	door	باب
t	t		Keyboard	tisʔ’	nine	تسع
D	d		Keyboard	daːr	home	دار
t’	t`	t + 02bc 700	Keyboard + IPA Ext	t’aːbiʔ’	stamp	طابع
d’	d`	d + 02bc 700	keyboard + IPA Ext	d’arab	he hit	ضرب

Table 3: Extract from the Transcription Key for Arabic

3.5 Editor-Friendly

The editing software features a Preview button that enables the lexicographer to instantly view the XML data in a clear reading style of an HTML document, to easily review the entries and introduce changes on the fly. Figure 5 demonstrates an HTML preview of the Spanish-Dutch entry that appeared in Figure 3 above:



estribo [estri'βo] *nm* **1** <en equitación> pieza de la silla de montar en que coloca los pies un jinete

{nl} - stijgbeugel *de*

◊ *estribos de plata*

{nl} - *zilveren stijgbeugels*

2 <apoyo para el pie> plataforma a modo de escalón que sirve para ascender a un vehículo

{nl} - voetrust *de*

◊ *los estribos de una motocicleta*

{nl} - *de voetrusten van een motorfiets*

♦ **perder los estribos** perder por enfado el control de uno mismo

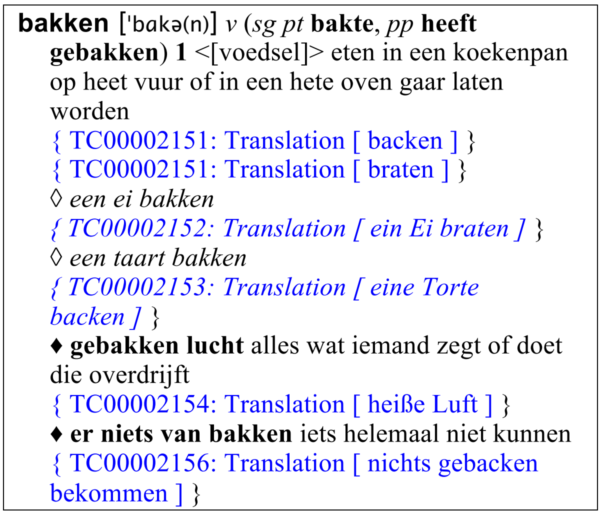
{nl} - uit zijn vel springen

◊ *Me haces perder los estribos.*

{nl} - *Ik spring door jou uit mijn vel.*

Figure 5: An extract from a Spanish-Dutch entry in the HTML preview

Recently we devised an alternative for using the XML Editor to enable translators to work directly on an MS Word DOC, where they just need to insert the translation components in ready-made fields that were converted from the XML data and will be re-converted back to XML with the help of the ID tags. The number of possible translation equivalents is limited to three per sense and one for each example and expression. Figure 6 shows a sample Dutch-German translation in DOC:



bakken ['bɑkə(n)] *v* (*sg pt* **bakte**, *pp* **heeft gebakken) 1** <[voedsel]> eten in een koekenpan op heet vuur of in een hete oven gaar laten worden
{ TC00002151: Translation [backen] }
{ TC00002151: Translation [braten] }
◊ *een ei bakken*
{ TC00002152: Translation [ein Ei braten] }
◊ *een taart bakken*
{ TC00002153: Translation [eine Torte backen] }
♦ **gebakken lucht** alles wat iemand zegt of doet die overdrijft
{ TC00002154: Translation [heiße Luft] }
♦ **er niets van bakken** iets helemaal niet kunnen
{ TC00002156: Translation [nichts gebacken bekommen] }

Figure 6: Translation in an MS Word document

3.6 Corpus

When we first set on the French project we discovered there were no publicly accessible corpora for French and the other languages. It was therefore decided to rely on whatever private corpora held by some of the editors and to refer cautiously to evidence found on the Internet. In addition, each editor-in-chief signalled out existing dictionaries that the lexicographers may turn to for general reference only, with copying strictly forbidden. Recently we began using the Sketch Engine corpus of Lexical Computing for Dutch[8], with satisfactory results.

3.7 Defining Vocabulary

This paper generally forgoes most editorial aspects that are not strongly concerned with the technological features of the series, but one that is most noteworthy and also somewhat underlines the *e* spirit of our venture is concerned with the defining vocabulary. It was decided that the vocabulary for each language will consist of all the headwords and their various inflections.

[8] http://sketchengine.co.uk/.

The reason is our expectation that most of the use of these dictionaries will be electronically, so that hyperlinking any word in the text to its appropriate entry should be possible. To enable this, all that is necessary is to have morphological connections among all the words.

Thus, once the full language core is compiled, we process a list of all the words used therein, then proceed to associate each word that is not a headword to its entry.

Unfortunately, although such an extensive defining vocabulary (based on the 12,000 main headwords) would seem to grant tremendous room for the lexicographers to maneuver, we found that often they did not fully abide by this regulation and did include other words in their definitions. Our solution to this problem will be to incorporate a new feature in the software that alerts the editor to any word that is not part of the list of headwords.

Meantime, those words that cannot be associated to entries become prime candidates for inclusion in any expansion of the dictionary.

3.8 QA & Processing Tools

Once data is received from editors or translators, it undergoes checking by the project manager as well as initial automated analysis of the contents, to basically confirm that all relevant components are in place. For example, that each sense has at least one example of usage, that an attribute or an indicator is included in addition to the definition, that the translation is accompanied by its phonetic transcription, etc. Figure 7 shows the home screen of our Utilities tool (currently being revised):

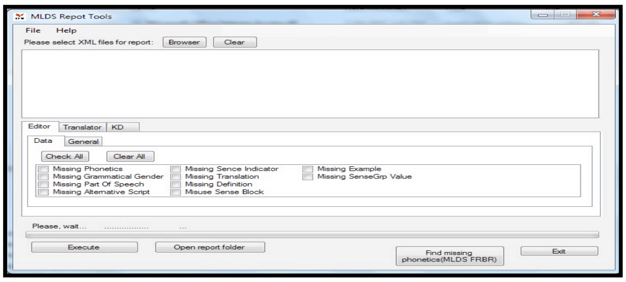
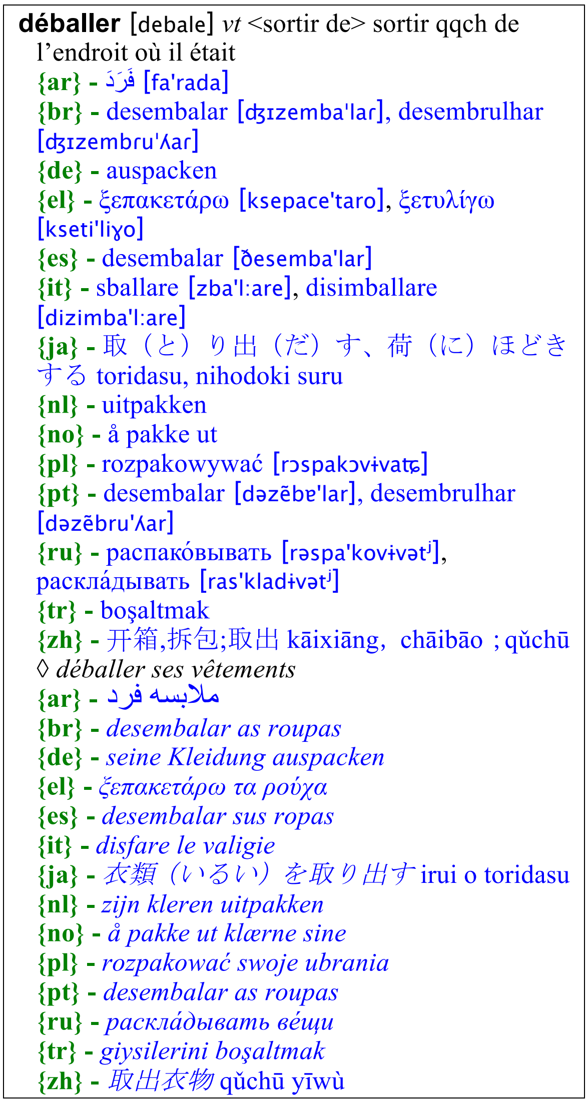


Figure 7: A snapshot of (former) automatic QA tool

Further processing of the data might concern changes made in the L1 entry during the translation, listing words that are not on the headword list, running statistics on the number of senses or examples, etc.

Each language core is organized in its own database, with all the translations available for the L1, as demonstrated in Figure 8. The data construction makes it possible to process specific language combinations or extract any components.



déballer [debale] *vt* <sortir de> sortir qqch de l’endroit où il était

{ar} - فَرَدَ [fa'rada]

{br} - desembalar [ʤɪzemba'laɾ], desembrulhar [ʤɪzembɾu'ʎaɾ]

{de} - auspacken

{el} - ξεπακετάρω [ksepace'taro], ξετυλίγω [kseti'liɣo]

{es} - desembalar [ðesemba'lar]

{it} - sballare [zba'l:are], disimballare [dizimba'l:are]

{ja} - 取（と）り出（だ）す、荷（に）ほどきする toridasu, nihodoki suru

{nl} - uitpakken

{no} - å pakke ut

{pl} - rozpakowywać [rɔspakɔvɨvatɕ]

{pt} - desembalar [dəzẽbɐ'lar], desembrulhar [dəzẽbru'ʎar]

{ru} - распакóвывать [rəspa'kovɨvətʲ], раскла́дывать [ras'kladɨvətʲ]

{tr} - boşaltmak

{zh} - 开箱,拆包;取出 kāixiāng，chāibāo ；qǔchū

◊ *déballer ses vêtements*

{ar} - ملابسه فرد

{br} - *desembalar as roupas*

{de} - *seine Kleidung auspacken*

{el} - *ξεπακετάρω τα ρούχα*

{es} - *desembalar sus ropas*

{it} - *disfare le valigie*

{ja} - *衣類（いるい）を取り出す* irui o toridasu

{nl} - *zijn kleren uitpakken*

{no} - *å pakke ut klærne sine*

{pl} - *rozpakować swoje ubrania*

{pt} - *desembalar as roupas*

{ru} - *раскла́дывать вéщи*

{tr} - *giysilerini boşaltmak*

{zh} - *取出衣物* qǔchū yīwù

Figure 8: An HTML preview of an extract from the French multilingual database

4. Application

Having the raw dictionary data in XML format enables putting it into use in numerous ways, whether for print or digital media, or for applying different components to suit different user groups.

4.1 Same Translation in Polysemous Entries

Usually each sense of the entry has all its relevant information appearing together, including the translation equivalent of the meaning and the example of usage. However, in DAK it was decided that when all senses of a polysemous entry happen to have the same translation equivalent it will be placed at the headword level before the first sense, to facilitate the user’s comprehension of the entire entry and to economize on the typographical representation. Figure 9 demonstrates two ways of displaying an entry, when each sense has a different translation and when all the senses have the same translation.

croyance *nf*
1 fait de croire **wiara** [vjara] *f* ◊ *la croyance à la liberté wiara w wolność* ◊ *la croyance en Dieu wiara w Boga*
2 *conviction* **wierzenie** [vjɛʒɛɲɛ] *nt* ◊ *les croyances religieuses wierzenia religijne*
croyance *nf* **crença** [ˈkrẽsɐ] *f*
1 fait de croire ◊ *la croyance à la liberté* crença na liberdade ◊ *la croyance en Dieu crença em Deus*
2 conviction ◊ *les croyances religieuses* as crenças religiosas

Figure 9: Different presentation of an entry when each sense has a different translation (Polish, above) or the same translation (Portuguese, below)

4.2 Same Entry for Different Users

Since the data contains all the relevant entry components, it is possible to make use of different elements to cater specifically for each target group. The next two figures demonstrate different uses of the French/Japanese database in dictionaries targeted for French learners of Japanese and for Japanese learners of French.

Figure 10 shows a sample entry in the French-Japanese section. For French users it provides sense indicators and Romaji, whereas for Japanese users it provides phonetic transcription of the headword, definitions instead of sense indicators, and no Romaji.

destin *nm*
1 fatalité 運命（うんめい）、宿命（しゅくめい） **unmee, shukumee** ◊ *accepter son destin 運命を受（う）け入（い）れる* unmee o ukeireru
2 vie 人生（じんせい）、生涯（しょうがい） **jinsee, shoogai** ◊ *un destin cruel 残酷（ざんこく）な人生* zankoku na jinsee
destin [dɛstɛ̃] *nm*
1 avenir décidé à l’avance 運命、宿命 ◊ *accepter son destin 運命を受*
2 existence, vie 人生、生涯 ◊ *un destin cruel 残酷*

Figure 10: French-Japanese entry for French speakers (above) and Japanese speakers (below)

Figure 11 shows a sample entry in the Japanese-French section. For French users the entries are arranged in Roman alphabetical order and the Romaji script of the headword appears first, and the entry includes definitions. For Japanese users the entries are arranged according to Kanji and do not include Romaji, the part of speech of the headword is in Japanese, there are sense indicators for the purpose of disambiguation, rather than definitions, and the French translation of each sense is accompanied by its phonetic transcription.

ho¯o1 ほう 方 *n*
1; 方向（ほうこう）, 方面（ほうめん）hookoo, hoomen **direction** ◊*山の方へ行く yama no hoo e iku aller en direction de la montagne*
2 いくつかの中のひとつ ikutsuka no naka no hitotsu **ceci** ◊*こちらの方を選ぶ kochira no hoo o erabu choisir celui-là*
ほう 方 *名詞*
1 =方向 **direction** [diʀɛksjɔ̃] *f* ◊ *山の方へ行く aller en direction de la montagne*
2 選択 **ceci** [səsi] ◊*こちらの方を選ぶ choisir celui-là*

Figure 11: Japanese-French entry for French speakers (above) and Japanese speakers (below)

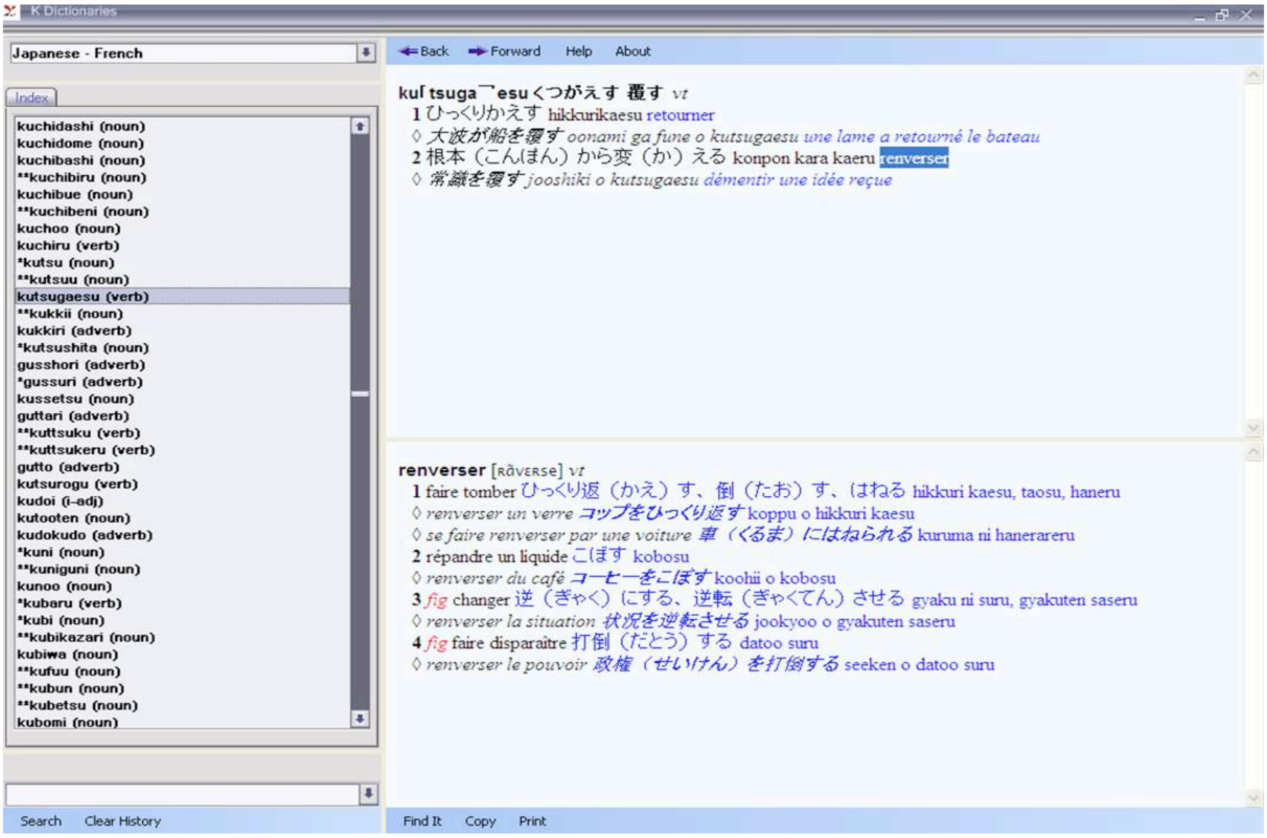
4.3 Desktop (Offline) Application

The print edition of each dictionary is accompanied by an electronic version that the user can install on his/her computer. It can either be downloaded from the publisher’s server, by inserting an individual code that appears in the book, or is offered on a CD-ROM that is added to the book (and can, of course, be released also on its own, regardless of the print edition).

In order to produce this application for each title in a semi-automated mode, we developed a generic XML dictionary ‘shell’ that can absorb any lexicographically structured data in XML format and process it into an electronic dictionary application for PC according to its own configuration. Its main features are:

- dual display of both dictionary sections on the same screen, and full linkage between the two with easy transfer from one to the other;
- various search options in either language, including advanced, wildcard and soundex searches, on different entry components;
- full hyperlink of the words used in the dictionary to their appropriate entries in either section, including headword inflections;
- hyperlinking items in illustrations and words in supplements to their relevant dictionary entries;
- compatibility with other desktop applications, including a hotkey for direct connection;
- back/forward paging that keeps track of all the entries that were previously loaded, which may be erased and restarted at any point;
- audio pronunciation, including self-recording;
- adapting the part of speech and all other labels, as well as the help guides and installation instructions, to the user’s native language;
- a skin engine enabling users to transform the visual aspects of the interface to their liking;
- loading an unlimited number of dictionaries concurrently, and selecting which languages to work with – whether for L1 or L2.

Figure 12 reproduces a screenshot from the dictionary application for French learners of Japanese, where the French translation of the second sense of the Japanese entry was doubleclicked and opened in the French-Japanese dictionary section appearing below it.



kuchidashi (noun)
kuchidome (noun)
kuchibashi (noun)
**kuchibiru (noun)
kuchibue (noun)
**kuchibeni (noun)
kuchoo (noun)
kuchiru (verb)
*kutsu (noun)
**kutsuu (noun)
kutsugaesu (verb)
**kukkii (noun)
kukkiri (adverb)
*kutsushita (noun)
gusshori (adverb)
*gussuri (adverb)
kussetsu (noun)
guttari (adverb)
**kuttsuku (verb)
**kuttsukeru (verb)
gutto (adverb)
kutsurogu (verb)
kudoi (i-adj)
kutooten (noun)
kudokudo (adverb)
*kuni (noun)
**kuniguni (noun)
kunoo (noun)
*kubaru (verb)
*kubi (noun)
**kubikazari (noun)
kubiwa (noun)
**kufuu (noun)
**kubun (noun)
**kubetsu (noun)
kubomi (noun)

ku⌈tsuga⌉esu くつがえす 覆す *vt*
1 ひっくりかえす hikkurikaesu retourner
◊ *大波が船を覆す oonami ga fune o kutsugaesu une lame a retourné le bateau*
2 根本（こんぽん）から変（か）える konpon kara kaeru renverser
◊ *常識を覆す jooshiki o kutsugaesu démentir une idée reçue*

renverser [ʀɑ̃vɛʀse] *vt*
1 faire tomber ひっくり返（かえ）す、倒（たお）す、はねる hikkuri kaesu, taosu, haneru
◊ *renverser un verre コップをひっくり返す* koppu o hikkuri kaesu
◊ *se faire renverser par une voiture 車（くるま）にはねられる* kuruma ni hanerareru
2 répandre un liquide こぼす kobosu
◊ *renverser du café コーヒーをこぼす* koohii o kobosu
3 *fig* changer 逆（ぎゃく）にする、逆転（ぎゃくてん）させる gyaku ni suru, gyakuten saseru
◊ *renverser la situation 状況を逆転させる* jookyoo o gyakuten saseru
4 *fig* faire disparaître 打倒（だとう）する datoo suru
◊ *renverser le pouvoir 政権（せいけん）を打倒する* seeken o datoo suru

Figure 12: A screenshot from the desktop application of the Japanese/French dictionary

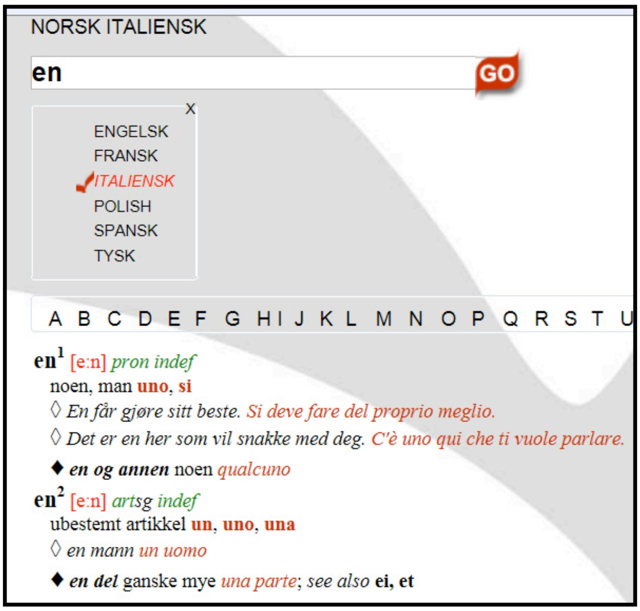
4.4 Online Application

Over the last couple of years we developed a test-site for online dictionary applications. The initial purpose is to investigate how to offer dictionaries online and how features of online dictionaries might affect the actual compilation and use of lexicographic content. The site also serves to demonstrate online apps to our partners. Since KD is basically a B2B company, and does not present its products and services directly to the actual end-users, this site is not targeting the general public.

As with the desktop application, the XML structure of the data enables its fairly straightforward application for online use as well. Figure 13 shows a screenshot from a subset of the main site that is dedicated to a series of Norwegian bilingual dictionaries, where it is possible to easily switch among languages, hyperlink words to their entries, etc.

4.5 Other Applications

KD does not create other electronic applications on its own, but cooperates with a wide range of technology partners for this purpose. The same XML data of our dictionaries is then used just as easily to develop versions for their own desktop and online usages, as well as for smartphones, tablets, handheld devices and any other form of digital media.



en[1] [e:n] *pron indef*
noen, man **uno, si**
◊ *En får gjøre sitt beste. Si deve fare del proprio meglio.*
◊ *Det er en her som vil snakke med deg. C'è uno qui che ti vuole parlare.*
♦ ***en og annen*** noen *qualcuno*
en[2] [e:n] *artsg indef*
ubestemt artikkel **un**, **uno**, **una**
◊ *en mann un uomo*
♦ ***en del*** ganske mye *una parte*; *see also* **ei, et**

Figure 13: A screenshot of a site for Norwegian bilingual dictionaries

5. Conclusion

Developing our dictionaries as an extensive database rather than as specific products demands a considerably higher initial investment over a considerably longer period of time on the one hand, but the consequences include many more potential by-products over a much longer term on the other hand.

In addition to continuing to solidify and enrich the content, we plan to further extend and exploit its database applications particularly as a base for semi-automatic development of new content – such as by data reversal, linking translations to their own language entries or combining several languages together – and to enhance its integration with various corpora and related applications.

6. References

COBUILD. *Collins COBUILD English Language Dictionary*. Glasgow: Collins. 1987.

DAK. *Dictionnaires Assimil Kernerman. arabe, chinois, grec, japonais, polonais, portugais, russe, turc*. Paris: Assimil. 2009.

Italiensk ordbok. Italiensk-norsk / Norsk-Italiensk. Oslo: Vega Forlag. 2010.

Spansk ordbok. Spansk-norsk / Norsk-spansk. Oslo: Vega Forlag. 2008.

Comparable Corpora BootCaT

Adam Kilgarriff, Avinesh P.V.S, Jan Pomikálek
Lexical Computing Limited
Brighton, UK
E-mail: adam@lexmasterclass.com, avinesh.pvs@gmail.com, xpomikal@fi.muni.cz

Abstract

The BootCaT method (Baroni and Bernardini, 2004) has proved a fast, effective and versatile approach to corpus building. The method has been applied to small specialist corpora for finding terminology and translations (as originally envisaged by Baroni and Bernardini), and to large, general corpora, for large numbers of languages. First we review BootCaT, and present some figures for the sizes of corpora that can be built in a few minutes, on various parameter-settings. To date BootCaT has not been applied multilingually. We explore this by building matching corpora for different languages from matching seeds. We consider three ways of obtaining matching seeds: manual translation, automatic translation, and by finding keywords from corresponding Wikipedia articles. In one experiment, we present a bilingual word sketch based on seed-translation by Google Translate. In another, seeds are from Wikipedia, and we evaluate the corpora by seeing, firstly, how many domain terms they deliver, and secondly, by seeing how often the terms in the one language are translation equivalents of the terms in the other.

Keywords: Comparable Corpora, Bootcat, Terminology

1. Introduction

The BootCaT method (Baroni and Bernardini, 2004) has proved a fast, effective and versatile approach to corpus building. Starting from a set of seed words, tuples (typically triples) of the seeds are randomly generated and sent as a query to a search engine. The pages which the search engine puts at the top of its search hits pages are retrieved, and, after a certain amount of filtering, de-duplicating, and cleaning, you have a corpus. For a bigger corpus, all that is required is a large-enough seed set and more queries to the search engine. The method benefits from all the work that the search engines do to identify relevant, non-spam, text-rich pages. The method has been applied to create small specialist corpora for finding terminology and translations (as originally envisaged by Baroni and Bernardini), and also large, general ones.

To date it has not been applied multilingually. In this paper we describe Comparable Corpora BootCat, a program that takes a set of seed terms for a domain in Language 1 (L1), bootcats an L1 domain corpus, finds corresponding seed terms for Language 2 (L2) and bootcats a matching L2 corpus. We describe challenges and procedures, and present a first pass at a 'bilingual word sketch' and a pilot evaluation.

2. BootCaT

2.1 Implementations

The implementation of BootCaT that we use throughout is WebBootCaT[1](Baroni et al., 2006) which provides a web interface, with the BootCaT process running on a remote server. This is in contrast to the original toolset, which was for installation on the user's computer and for running from the Unix command line or Dos prompt.

The WebBootCaT suite uses Onion[2]for deduplication and Justext[3] for filtering (Pomikalek, 2011).

As the original toolset is a set of open-source perl scripts, they are readily open to customisation by anyone wishing to use them, and there are numerous BootCaT variants in use at various places.

2.2 Uses

Uses of the tools for creating small, specialist corpora, including translation and teaching translation, terminology and teaching terminology and domain lexicography, are widespread, though used rather than reported on, so the evidence is anecdotal.

The BootCaT method has also been used to create large, general-language corpora for lexicography and general linguistic research (Sharoff, 2006; Baroni, 2009; Kilgarriff et al., 2010).

2.3 Parameters

There are numerous parameters to select when running a BootCaT procedure. The ones which can be set in the WebBootCat interface (advanced options) are:

- File Types: HTML, RSS, MS-Word, pdf, plain text, any.
- **Creative Commons licence only:** (to address possible copyright concerns).
- **Tuple size:** how many items in the search to be sent to the search engine.

[1] Access to WebBootCaT is available to all registered users of the Sketch Engine service, see http://www.sketchengine.co.uk.

[2] http://code.google.com/p/onion/

[3] http://code.google.com/p/justext/

- **Sites list:** it is possible to restrict the search to sites, or to domains, for example .it for pages with URLs ending in .it only.
- **Max tuples:** The number of queries to be sent to the search engine.
- **Max URLs per query:** For each query result, how many URLs do we attempt to retrieve (from the top of the search hits list).
- **Sites list:** it is possible to restrict the search to sites, or to domains, for example .it for pages with URLs ending in .it only.

There are further options determining which retrieved pages are filtered out. They are left at their default settings in the experiments.

There is also the choice of search engine and, of course, the choice of seeds (or, more generally, the methodology for selecting seeds).

2.4 Corpus Sizes

One question of interest for potential BootCaT users is: how large a corpus do I get? We ran some experiments using seed terms drawn from wikipedia (see discussion below). We used three domains - Stradivarius, volcanoes and pancreatic cancer; two search engines - Yahoo and Bing; three 'sizes' - 10, 50 and 250 search engine queries; and four languages - Czech, English, French and German. Results are given in Table 1.[4]

The corpora took between 30 seconds and 15 minutes to create, depending on corpus size.

The URLs figure has a maximum of ten times the 'Queries sent' figure, since we take up to ten URLs from each query. In fact it was a lower number in each case, as, for some of the queries, the search engine offered less than ten hits, or there were duplicates among the hits for different queries. The URLs figures for Czech are lowest, as there were often not ten hits for a query, and different queries in the same domain often pointed to the same URL.

The Docs figure has a maximum of the URLs figure, and they would be the same if all URLs sought were found, and provided text which passed through Web BootCaT's de-duplication procedures and filters for pages which do not appear to contain running text of the language in question. Typically around one third of URLs are not found, or the page that is retrieved is rejected.

Web pages are of sizes that vary by orders of magnitude, and a page with 100,000 words in it can turn a small corpus into a large one, so there is no very dependable relation between number of documents and size of corpus. If we divide the number of words in each corpus by the number of documents contributing to it, to give an average document length for each corpus, the figures vary from 1,400 to 26,000. 26,000 was an outlier: in most cases the average document length was between 2,000 and 8,000 words, with there being more long documents in 'Stradivarius' than 'volcanoes'.

One would expect there to be interactions between search engine, language and domain, as different search engines will have prioritised different languages and types of page, and different domains will tend to have different kinds of page. We note some differences, but the experiment is too small-scale to draw inferences. The two search engines provided comparable sizes of corpora.

2.5 Earlier evaluations

In the context of large, general language corpora, the papers mentioned above (Sharoff, 2006; Baroni, 2009; Kilgarriff et al., 2010) all makes efforts to evaluate the resulting corpora. Here, our focus is on small specialised corpora where the one evaluation we are aware of is (Bernardini et al., 2011). The authors present a group of trainee translators with a translation task: to translate Patient Information Leaflets (as found in drug packets) from English into Italian. They are offered a variety of resources including several bootcatted corpora, with seeds emphasising 'domain' or 'genre'; the two bootcat corpus types were found to be the most useful inputs for the task.

3. Going Multilingual

The core method for producing a multilingual BootCat corpus is:

- take a set of seed terms for a domain in L1
- bootcat an L1 domain corpus
- take corresponding seed terms for L2
- bootcat an L2 domain corpus.

We call this Comparable Corpora BootCaT as the resulting corpora will be comparable in the sense of the Building and Using Comparable Corpora workshop series:[5] different languages but similar content.

One question that the outline begs is, how do we find corresponding seed terms across languages?

3.1 Finding Corresponding Seeds

We would like the L2 seeds to demarcate the same domain as the L1 seeds. The obvious thing to do is to translate them. This might be done by the user, or by automatic lookup in a bilingual dictionary.

[4] Results are only shown for Stradivarius and volcanoes. For pancreatic cancer our wikipedia-based method for finding equivalent articles could not be applied because there was no corresponding article.

[5] http://comparable.limsi.fr/bucc-workshop.html

Language	Search Engine	Queries sent	Volcanoes			Stradivarius		
			Urls	**Docs**	**Kwds**	**Urls**	**Docs**	**Kwds**
English	Bing	10	84	51	244	70	46	230
		50	318	180	679	230	150	1230
		250	941	515	1580	808	483	5326
	Yahoo	10	67	39	152	60	47	148
		50	281	176	445	267	196	1071
		250	867	527	1232	937	649	3700
French	Bing	10	79	45	150	74	52	264
		50	246	152	461	225	145	1020
		250	755	506	1445	612	379	3815
	Yahoo	10	79	36	118	82	60	720
		50	285	154	695	257	156	1155
		250	994	527	1737	843	510	2317
German	Bing	10	49	39	112	32	19	126
		50	174	139	236	183	136	1071
		250	460	339	1288	407	279	2511
	Yahoo	10	59	44	88	40	21	36
		50	246	161	609	147	82	142
		250	775	449	2135	446	250	792
Czech	Bing	10	38	24	79	26	16	154
		50	78	47	168	44	24	624
		250	239	73	463	194	102	1210
	Yahoo	10	47	27	96	44	31	44
		50	120	69	153	54	29	75
		250	453	158	535	315	182	1108

Table 1: BootCat corpus sizes (in URLs sought, documents contributing text, and thousands of words) for two domains, two search engines, three numbers of queries sent to the search engine, and four languages. 'Urls' is number of URLs sought. 'Docs' is the number of web pages that passed through the filters to contribute a document to the corpus. 'Kwds' is the final corpus size, in thousands of words.

The main problem with the first method is that the user may not know the domain, or the language pair, well enough to do the translation. Also it can be time-consuming, and, from the perspective of system development and evaluation, it introduces a large extra unknown into the process: some people will do it better than others, or from a different perspective. We conclude that it is good to offer users the option of translating seeds themselves, or editing automatically-translated seeds, or simply writing the L2 seeds from scratch, but we also need to offer an automatic method.

The main problem with the dictionary-lookup method is the availability and coverage of dictionaries. We would like to cover a large number of language pairs, but each language pair requires its own dictionary (or, ideally, two, one for each direction). We need to cover technicalvocabulary, as that is of most use for building domain corpora, so the dictionaries will need to be big, with good coverage of very many domains. Accessing any one such dictionary typically involves extensive negotiation and we ideally want hundreds.

Two resources covering very many language pairs and directions in a convenient online format are Google Translate and Google Dictionary.

These two resources differ in several ways. For Google Translate, the expected form of input is a text, and the engine aims to disambiguate each term in the input according to context so we get just one translation. This may be convenient for term-translation, if the set of L1 terms is presented as a text (with suitable delimiters between them) as the terms may mutually disambiguate. Google Dictionary often presents multiple translation candidates, like a standard bilingual dictionary.

Second, many more language pairs are offered for Google Translate than Google Dictionary. As at April 2011 Google Translate was available for 1171 directed language pairs and Google Dictionary, for 50.

Third, they operate on different Terms of Use. Google recently stated:

> Due to the substantial economic burden caused by extensive abuse, the number of requests one may make per day will be limited and the API will be shut off completely on December 1, 2011.[6]

We currently have a variant of CCBC that uses Google Translate, but it seems we shall not be able to use it for long. Our conclusion on translation-via-dictionary-lookup is that it is good where we have access to a good dictionary, but getting access is a problem, language pair by language pair, and leaves us at the mercy of dictionary providers.

A third route does not translate at all, but uses wikipedia, viewed as a comparable corpus, as input.[7] It exists for 265 languages and is freely available, and it is often possible to find corresponding articles in different languages. In some cases they are translations but more frequently they are not. Where we have a corresponding pair we can find keywords and key terms from the L1 wikipedia article and the L2 wikipedia article and use them as seed words for the BootCaT processes. Note that we use wikipedia for seeding the process, but go outside wikipedia to build the corpus: we do not depend on the wikipedia text for the main phase.

The corpora described for Stradivarius, volcanes and pancreatic cancer were created in this way. They were also the corpora usd in the evaluation. The keywords to use as seeds for BootCat were the words which with the highest ratio of frequency in the wikipedia article, to frequency in a reference corpus: the algorithm is given below.

```
-----------------------------------------------------------------
Algorithm Pesudo Code
For focus corpus F and for reference corpus R
   - Make a frequency list
   - Normalise to per-million
   - add 100 to each normalised figure[8]
   - for each word w
        Score(w) = Value for corpus F/ Value for corpus R.
-----------------------------------------------------------------
```

Once all the scores were calculated the words were sorted by scores, and the top 100 from the list were then used as seeds for WebBootCat.

4. Automatic Term Recognition and Term Translation Spotting

A principle use for domain-specific corpora is term-finding, as a manual, semi-automatic, or fully automatic procedure. The fully automatic approach, ATR, is a topic with a substantial literature: see (Zhang et. al., 2010) for a recent review and an evaluation of alternative approaches.

We expect our corpora to be used for term-finding, most likely in a procedure where an automatic process proposes candidates which are then accepted or rejected by a person. So it is reasonable to evaluate BootCaT according to how good its corpora are as sources for ATR.

Three relevant observations from ATR are:

- Two distinct dimensions for assessing candidate terms are ‘unithood’ and ‘termhood’. Unithood (only applicable to multi-word candidates) concerns the extent to which the distinct words in a candidate expression should be treated as a single unit. Termhood concerns the extent to which a candidate belongs to the domain, as distinct from the language in general
- Different domains are very different (so a good procedure in one domain may not be good in another)
- Evaluation is very hard. There is little overlap between different resources. Experts differ. Most evaluation efforts only support limited and local conclusions.

An area neighbouring ATR is ‘Term Translation Spotting’, which makes use of comparable corpora, a field inaugurated in (Fung, 1995). To evaluate CCBC corpora, this area is highly salient. It is also relevant for statistical machine translation, as it is closely related to the SMT challenge of finding sentence-pairs that correspond across languages.

While we use ATR to evaluate BootCaT and CCBC, it has not been the focus of our research. Our focus has been thecorpus-building itself. At time of writing our ATR machinery is underdeveloped, so ATR results are not yet as good as the corpora may justify.

[6] http://code.google.com/apis/language/translate/overview.html

[7] The approach has been suggested by Silvia Bernardini, Federico Zanettin and Federico Gaspari, personal communication.

[8] We add N=100 to solve two problems (a) If a word doesn't occur in a corpus, its freq-per-million is zero, and divinding by zero isn't possible (b) by varying N, we can get a keyword list that tends to have lower-frequency words in (with low N) or higher-frequency words in (with high N).

5. Bilingual word sketches

CCBC corpora can be used for term-finding, as two matching but independent datasets, and it is likely that this will be their most common use. But perhaps we can do more, offering candidate translation pairs. We have observed that L1 and L2 key term lists often contain translation pairs. Could we find them automatically, offering the user a list of likely translations for each L1 term?

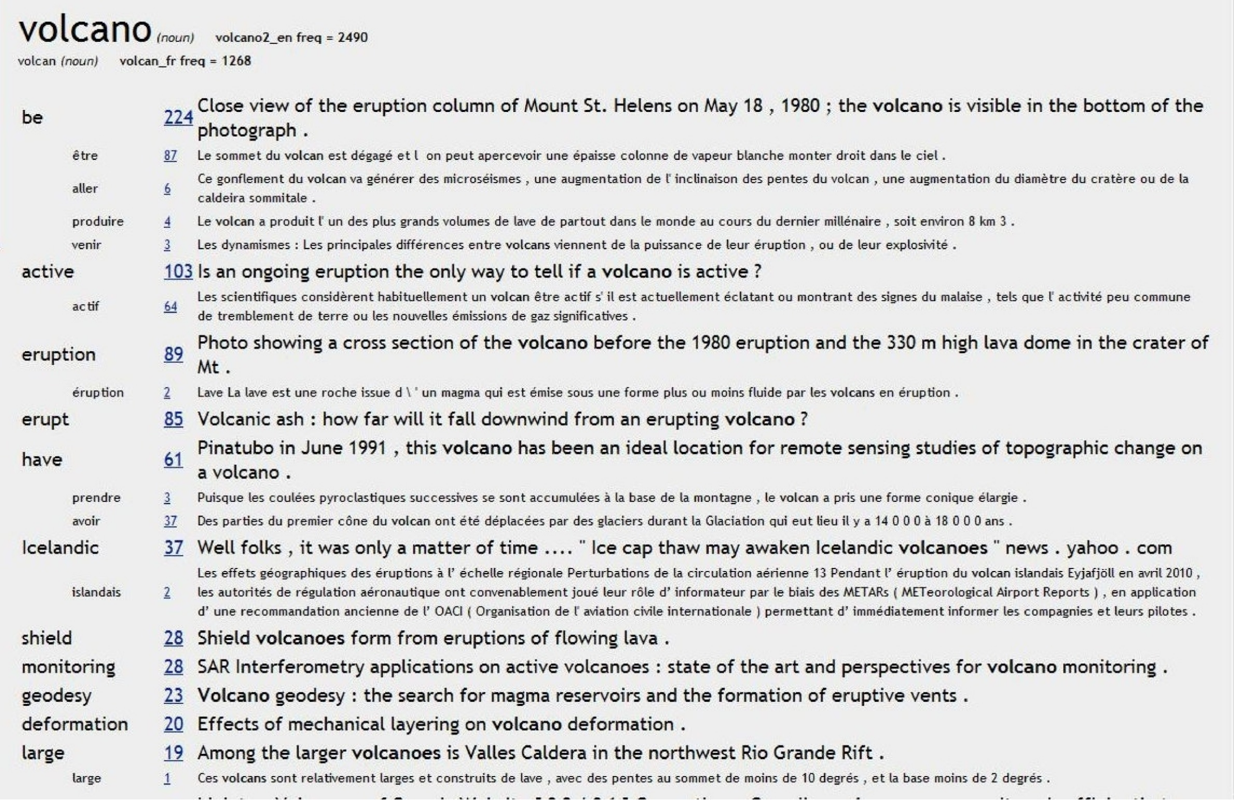
We tried this as follows:
The Sketch Engine already has a collocation-discovery method, based on a grammar to find candidates, and statistics to find the most salient candidates (Kilgarriff et. al., 2004). We view terms as a subset of collocations, and use the existing machinery, with a reduced grammar, as a term grammar. (We excluded some grammatical relations that give rise to collocations but are not considered as giving rise to candidate terms, for example, adverb-modifying-verb.)

We then use a bilingual dictionary to translate all the component words appearing in the L1 term list into L2. We lemmatise the corpus and base the analysis on lemmas (eg dictionary headwords) rather than wordforms throughout. We then apply the 'cross-product' method first proposed by Gregory Grefenstette to find, for each multiword unit, combinations of their translations which are present in the L2 corpus:

```
-----------------------------------------------------------------
for each L1 collocation < a,b >
  - for each translation of a : ta
      - for each translation of b : tb
        - see if < ta , tb > is in the L2 collocation list
-----------------------------------------------------------------
```

If it is, we have a candidate translation pair. For any L1 collocation, there may be 0, 1 or multiple L2 collocations (and vice versa).

We would like to produce 'bilingual word sketches'. Monolingual word sketches are one-page automatic corpus-based summaries of a word's grammatical and collocational behaviour and have been in use for lexicography since 1998. It is far from clear how the definition should be extended to cover the two-language case, but a first pass at the bilingual word sketch was prepared using the method above (but only with the single translations for each word that Google Translate provided) and is shown in Figure 1.



volcano *(noun)* volcano2_en freq = 2490
volcan *(noun)* volcan_fr freq = 1268

be	224	Close view of the eruption column of Mount St. Helens on May 18 , 1980 ; the **volcano** is visible in the bottom of the photograph .
être	87	Le sommet du **volcan** est dégagé et l on peut apercevoir une épaisse colonne de vapeur blanche monter droit dans le ciel .
aller	6	Ce gonflement du **volcan** va générer des microséismes , une augmentation de l' inclinaison des pentes du volcan , une augmentation du diamètre du cratère ou de la caldeira sommitale .
produire	4	Le **volcan** a produit l' un des plus grands volumes de lave de partout dans le monde au cours du dernier millénaire , soit environ 8 km 3 .
venir	3	Les dynamismes : Les principales différences entre **volcans** viennent de la puissance de leur éruption , ou de leur explosivité .
active	103	Is an ongoing eruption the only way to tell if a **volcano** is active ?
actif	64	Les scientifiques considèrent habituellement un **volcan** être actif s' il est actuellement éclatant ou montrant des signes du malaise , tels que l' activité peu commune de tremblement de terre ou les nouvelles émissions de gaz significatives .
eruption	89	Photo showing a cross section of the **volcano** before the 1980 eruption and the 330 m high lava dome in the crater of Mt .
éruption	2	Lave La lave est une roche issue d \ ' un magma qui est émise sous une forme plus ou moins fluide par les **volcans** en éruption .
erupt	85	Volcanic ash : how far will it fall downwind from an erupting **volcano** ?
have	61	Pinatubo in June 1991 , this **volcano** has been an ideal location for remote sensing studies of topographic change on a volcano .
prendre	3	Puisque les coulées pyroclastiques successives se sont accumulées à la base de la montagne , le **volcan** a pris une forme conique élargie .
avoir	37	Des parties du premier cône du **volcan** ont été déplacées par des glaciers durant la Glaciation qui eut lieu il y a 14 0 0 0 à 18 0 0 0 ans .
Icelandic	37	Well folks , it was only a matter of time .... " Ice cap thaw may awaken Icelandic **volcanoes** " news . yahoo . com
islandais	2	Les effets géographiques des éruptions à l' échelle régionale Perturbations de la circulation aérienne 13 Pendant l' éruption du **volcan** islandais Eyjafjöll en avril 2010 , les autorités de régulation aéronautique ont convenablement joué leur rôle d' informateur par le biais des METARs (METeorological Airport Reports) , en application d' une recommandation ancienne de l' OACI (Organisation de l' aviation civile internationale) permettant d' immédiatement informer les compagnies et leurs pilotes .
shield	28	Shield **volcanoes** form from eruptions of flowing lava .
monitoring	28	SAR Interferometry applications on active volcanoes : state of the art and perspectives for **volcano** monitoring .
geodesy	23	**Volcano** geodesy : the search for magma reservoirs and the formation of eruptive vents .
deformation	20	Effects of mechanical layering on **volcano** deformation .
large	19	Among the larger **volcanoes** is Valles Caldera in the northwest Rio Grande Rift .
large	1	Ces **volcans** sont relativement larges et construits de lave , avec des pentes au sommet de moins de 10 degrés , et la base moins de 2 degrés .

Figure1: English-French bilingual word sketch for *volcano*.

6. CCBC : Pilot evaluation

We conducted a first evaluation of CCBC by asking bilingual experts, for a small set of corpora and term candidates, “should this term be in a specialised dictionary for the domain”, and, for the two-language case, for each L1 item, is its translation in the L2 list.

We used eight of the corpora described in Table 1: we selected only the ones that used Yahoo, and only the largest, based on 250 search-engine queries, from each set of three sizes. We used two corpora for each language, one on volcanoes and one on Stradivarius. One of the evaluators also assessed the English and German pancreatic-cancer corpora.

For each corpus we identified 30 keywords and 100 top collocations. The keywords, all single words as opposed to multiwords, were the words that had the highest ratio between normalised frequency in the domain corpus and in a large web-crawled reference corpus for the language. In addition keywords had to occur in at least ten different documents.

Who	Wds	Trans	Mwds
	Volcanoes, En		
E-Cz	29/30		10/85
E-De	29/30	10/29	16/85
E-Fr	29/30	19/29	24/85
	Stradivarius, En		
E-Cz	19/29		13/85
E-De	26/30	3/26	9/85
	Stradivarius, De		
E-De	16/30	2/16	6/84
	Cancer, En		
E-De	27/30	9/27	
	Cancer, De		
E-De	22/30	10/22	8/90
	Volcanoes, Fr		
E-De	27/30	19/27	5/83

Table 2: Manual evaluation results, by corpus and evaluator.

Wds: One-word term candidates assessed as good.

Trans: the good single-word terms for which translations were found in the corresponding list for the evaluator's other language.

Mwds: multi-word term candidates assessed as good.

The 100 top collocations were identified as the items with the highest scores in the domain corpus (with salience as defined on the Sketch Engine website). This used the technology for collocation-finding, so the collocations were in fact 3-tupes of <word1, word2, grammatical relation> (or in some cases, 4-tuples with the fourth item being a preposition). A consequence was that the same word-pair sometimes occurred twice in the collocation list, once as, eg, <ice, glacial, modifier> and once as <modified, glacial, ice, modified>. There were 15 such duplications in each of the English files, so once we had de-duplicated, there were just 85 items assessed rather than 100.

Whereas the single words were selected purely on the basis of their termhood, the multi-word candidates were selected purely on the basis of their unithood.

The evaluator was presented with the two-part list (single words, and multiwords) and given four possible answers to the question “should this term be in a specialised dictionary for the domain?” - yes, probably, possibly, no. In the event evaluators almost always used ‘yes’ or ‘no’ and the few ‘probably’ values were treated as ‘yes’ and the ‘possibly’ ones as ‘no’. Then, for the multilingual part, the evaluators were asked to judge, whether each of the ‘good’ terms in the L1 list had a translation amongst the ‘good’ terms on the L2 list. There was one evaluator each for Czech and English, German and English and French and English, called E-Cz, E-De and E-Fr in Table 2. All were language professionals, native speakers in one of their languages and of near-native competence in the other. Not all translators completed all parts of the exercise, hence the blank cells in the table.

6.1 Discussion

It is immediately apparent that the system performed well on the single-word terms and poorly on the multiword ones. The majority of the single-word candidates were good in all cases, with only one bad item in thirty in ’volcanoes-en’. (The same bad item was picked out by all three evaluators.) By contrast, the best result for multiword candidates was under one in three.

Likewise for translations: matched corpora often furnished translation pairs from among the single-word lists, with over half of the lists falling into translation-pairs in a couple of cases. For multiword translations, we do not provide a column in Table 2 for the simple reason that our evaluators, when they looked, did not find any. The column would have contained only zeroes and blanks.

For the English lists that all three evaluators assessed, there was very high agreement on what was good for the single-word items, but low agreement for the multiwords. This is, we believe, because it is a hard judgement for a non-expert in the domain to make (specially outside one’s mother tongue, as E-De said when explaining the blanks in her results).

We believe the reasons for the poor performance on multiwords are, firstly, insufficient care in adopting our collocation grammar to a term grammar, and second, the fact that our multiword canddiate selection was based only on unithood, and not at all on termhood.

This was a small pilot evaluation, and over the coming months we shall be undertaking a more careful evaluation. The pilot has shown us that, as measured by results for single-word terms, our corpora look satisfactory, but we need to adopt lessons from ATR and translation-term evaluation in order to improve performance on multiword candidates.

7. Summary

We have presented CCBC, a suite of methods for 'bootcatting' comparable corpora. We first reviewed BootCaT, and presented some data on the size of corpora that one might expect to generate with a range of search engines, languages, domains, and query-set sizes. We then discussed various ways for turning BootCaT bilingual. We presented a first draft 'bilingual word sketch'.

We made an initial evaluation of BootCaT and CCBC by considering the corpora that were produced as sources for automatic term recognition, and asking experts to evaluate the candidate term lists. This gave some evidence that the corpora were useful, and many pointers for what we need to do next.

8. Acknowledgements

This work was supported by EU FP7 grant 248307 'PRE-SEMT' and EU LLP grant 505630 'Kelly'. We would also like to thank our three evaluators: Marie-Aude Lefer, James Thomas and Marion Winters.

9. References

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004,* Lisbon: ELDA. pp. 1313-1316.

Baroni, M., Kilgarriff, A., Pomikalek, J. & Rychly, P. (2006). WebBootCaT: a web tool for instant corpora. In *Proceedings of Euralex 2006*, Alessandria: Edizioni dell'Orso, pp. 123-132.

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

Bernardini, S., Ferraresi, A. & Zanchetta, E. (forthcoming). Old needs, new solutions: comparable corpora for language professionals. In S. Sharoff, R. Rapp & P. Zweigenbaum (eds.) *Building and Using Comparable Corpora. Springer*.

Fung, P. (1995). Compiling bilingual lexical entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Annual workshop on Very Large Corpora,* Boston, Massachusetts, pp. 173-183.

Kilgarriff, A., Reddy, S., Pomikalek, J. & Avinesh PVS. (2011). A Corpus Factory for Many Languages. In *Proceedings of LREC'10,* Valletta, Malta, May 19-21.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX 2004,* Lorient, France, pp. 105-116.

Pomikálek, J. (2011). Removing Boilerplate and Duplicate Content from Web Corpora. *PhD Thesis, Masaryk University, Brno, Czech Republic.*

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

Zhang, Z., Iria, J., Brewster, C. & Ciravegna, F. (2010). A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, May 26-30, Marakech, Morocco.

Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project KELLY – issues on reliability, validity and coverage

Sofie Johansson Kokkinakis, Elena Volodina
University of Gothenburg, Sweden
Språkbanken, Department of Swedish, University of Gothenburg, Box 200, 405 30 Göteborg
E-mails: sofie.johansson.kokkinakis@svenska.gu.se, elena.volodina@svenska.gu.se

Abstract

At present there are relatively few vocabulary lists for Swedish describing modern vocabulary as well as being adapted to language learners' needs. In Europe including Sweden there exist approaches to unify ways of working consistently with language learning, one example worth naming in this respect is the Common European Framework of Reference (CEFR) which provides guidelines for systematic approach to language teaching and assessment of language proficiency. This article describes EU project Kelly (KEywords for Language Learning for Young and adults alike, 2009-2012), the main objective of which was to create vocabulary lists for nine languages (Swedish, English, Norwegian, Greek, Italian, Polish, Arabic, Chinese and Russian) and adapt them to CEFR levels. We describe the process of compiling and validating the Swedish Kelly-list, dwell on benefits and limitations of using a corpus based approach in this project; as well as mention the impact of the methodological approach for compiling vocabulary lists for specific purposes.

Keywords: language learning; corpus-based approach; frequency-based vocabulary list

1. Introduction

The EU project Kelly (KEywords for Language Learning for Young and adults alike), was granted to ten partner organizations[1] for the period of 2009-2012. The main objective was to develop a bilingual language learning tool for nine languages; Swedish, English, Norwegian, Greek, Italian, Polish, Arabic, Chinese and Russian and to adapt it to the above-mentioned CEFR levels[2] (Council of Europe, 2001). Monolingual vocabulary lists for the nine project languages were translated into the eight partner languages, generating 72 language pairs.
CEFR covers six proficiency levels, starting with the beginner level (A1, A2), covering the intermediate level (B1, B2) and up to the mastery level (C1, C2). Proficiency levels are partly defined in terms of what a learner should know as far as grammar and communication skills are concerned in the form of "can-do"-statements, and partly in terms of topical (domain) knowledge (e.g. education, sports, etc.). In the light of the above mentioned systematic learning- and assessment strategies which are nowadays practiced in Europe and Sweden, the project has been aiming to adapt the selected vocabulary to the CEFR-levels and to evaluate to which extent the CEFR-specific domain vocabulary should be a part of the Kelly lists.

A corpus based methodological approach was used to ensure that the vocabulary list coverage corresponds to empirically based evidence and authentic language and language use. The Corpus Factory tool (Kilgarriff et.al. 2010) was used to aid the creation of a new Swedish corpus (SweWAC); SketchEngine (Kilgarriff et.al. 2004) was used as a workbench for statistically based selection of potential headwords. Various Swedish electronic lexical resources such as SALDO (Borin & Forsberg, 2009) and SMDB (Berg & Cederholm, 2001) were used for proofreading the Swedish list of selected headwords. A database was built to facilitate comparison of the Kelly vocabulary lists and to ensure the validity of the vocabulary item selection across all languages. The final product – a web based language learning tool – is planned to be evaluated quantitatively and qualitatively by a web based vocabulary levels test and a questionnaire.

The Swedish monolingual vocabulary list is at present a freely available electronic resource that reflects a selection of 8 425 most frequent words in modern Swedish as described in Johansson Kokkinakis & Volodina (forthcoming) and Volodina & Johansson Kokkinakis (forthcoming-a, b). In this paper we discuss the technologies we have used mentioning their strengths and limitations and their overall impact on the quality of the Swedish list. Validation and coverage are also described in detail to demonstrate the linguistic appropriateness of this approach.

2. From corpus to wordlist in a nutshell

The main principle of the KELLY project was that the final vocabulary lists should reflect modern language, constitute the most frequent core vocabulary, plus be based on objective selection.

2.1 Corpus Factory for corpus collection

To start with, the corpora for vocabulary selection had to reflect present-day language. Moreover, to ensure comparability between vocabulary lists for the nine

[1] Adam Mickiewicz University, Poland; Cambridge Lexicography and Language Services, UK; Consiglio Nazionale delle Ricerche, Italy; Institute for Language and Speech Processing/R.C. "Athena", Greece; Keewords, Sweden; Lexical Computing Ltd, UK; University of Gothenburg, Sweden; University of Leeds, UK; University of Oslo, Norway; University of Stockholm, Sweden (coordinating partner)
[2] http://www.coe.int/t/dg4/linguistic/cadre_en.asp

languages and to guarantee objectivity of word selection, the corpora had to contain at least 100 mln words and preferably to be collected from the web.

Given the above-mentioned prerequisites for the project we faced the problem of an appropriate web corpus of the defined size. There were at the time only two annotated general-language corpora available for Swedish – Parole Corpus and Stockholm-Umeå-Corpus (SUC) (Källgren et al., 2006). Neither of the two could qualify as a candidate core corpus for the KELLY-list. Parole dates from 1976-1997 and does not meet the requirement of being a collection of modern language samples. SUC is a balanced corpus dating from 1990's, but comprises only 1,2 mln. words and does not meet the requirement of size.

Therefore a big modern corpus of Swedish, a web-corpus SweWAC (Swedish Web-Acquired Corpus) was compiled by the KELLY partner "Lexical Computing Ltd" using Corpus Factory tool (Kilgarriff et. al., 2010). SweWAC is at present available via commercial concordance tool SketchEngine in its original form (http://www.sketchengine.co.uk/) as well as via the concordance tool Korp freely available through the Swedish Language Bank as a "citation" corpus, in which sentences are mixed in random order so that the full texts cannot be retrieved (http://spraakbanken.gu.se/korp/).
Compiling a web-based corpus for Swedish was a process consisting of several steps:

1. Collect "seed word" list, approximately 500 mid-frequency words whose frequency range is between 1000 and 6000. This was done using texts on Wikipedia – first a "Wiki-corpus" was compiled as a primary corpus for seed-word selection, word form frequency was calculated (as opposed to base forms/lemmas), and then 500 mid-frequency word forms were selected for further web-search. Length restriction was set on the seed words: they should be at least 5 characters long to sort out coinciding word forms in other languages (e.g. Swedish versus English *fast*). Words containing digits or other non-characteristic for the language characters were discarded.
2. Repeatedly select three random seed words to create a query, and send the query to a search engine.
3. Retrieve hit pages and clean the text, e.g. remove navigation bars, ads, duplicates. The web-corpus finally consisted of 114 million words.

Among the advantages of web-collected corpora we can name the following:

- Since its construction is a highly automated process, short collection time at low costs is ensured.
- Texts collected from the web tend to contain more spoken-like interactional language since there are a lot of forums and blogs; therefore, compared to classical corpora, they have a benefit of complementing strictly written mode of language with everyday colloquial language.

Among the disadvantages or rather limitations of a web corpus we can name the following:

- First of all the absence of control over the kinds of texts that constitute the corpus. Such corpora are therefore unpredictable as to their structure and contents, presenting an unclear mixture of domains and most probably devoid of balance between domains and genres.
- As our experience of SweWAC has shown, besides texts in Swedish there is a minor percentage of texts written in other languages, among them Norwegian, Danish and English. Presumably the reason for that is presence of ambiguous seed words, for example international proper names, e.g. *Albert, Alexander, Berlin, Chris, Chicago, Daniel*; non-Swedish spelling of words, e.g. *America* (as opposed to the Swedish *Amerika*), British (as opposed to *brittisk*), *company* (Swedish *företag*), *college, corporation* etc. A number of seed words coincided in form with English words, even though their length was longer than or equal to five characters, e.g. *album, attack, civil.* One way out of this may be POS-tagging of the wiki-corpus and filtering seed words of unwanted word classes (e.g. proper names and foreign words) prior to sending queries to the search engines. Another even better alternative is to have a language team prepare a list of seed words (or even better several lists for different genres) and thus ensure the more or less balanced and predictable structure of the corpus.
- Another problem with a web corpus is that automatically collected web-texts may appear with different character encoding.

However, these limitations have proven to be minor problems. The method of working on the KELLY-lists was formed in such a way that most of problems mentioned above were corrected during the validation phase through word list comparisons between languages. This and some other selection strategies are described later in this article.

2.2 Lemmatization and POS tagging of SweWaC

Tokenization, lemmatization and POS-tagging was performed by the Swedish team using the tools developed by Kokkinakis & Johansson Kokkinakis (1997). SweWAC was tagged for part-of-speech, lemma and morphosyntactic information (case, gender, number), thus facilitating frequency analysis of word forms, lemmas, and grammatical features. The way lemmatization was performed has naturally influenced the headwords in the Kelly-list. That is why it is important to comment on what we understand by lemma in this context.

The frequency count in the Swedish KELLY-list was calculated upon lemmas (or lem-pos as they are otherwise called), i.e. base form of the word plus its part-of-speech. More closely, in SweWAC context lemma (lem-pos) stands for a set of word forms having the same stem or base form and belonging to the same word class, e.g. all occurrences of the word forms *flicka, flickas, flickan, flickans, flickor, flickors, flickorna, flickornas* are counted together since they have the same base form *flicka* (Eng. *girl*), the same word class *noun* and the same gender *uter*. However, such definition of a lemma allows grouping together words that share the same base form and word class, but not grammatical features (inflectional morphological aspects), e.g. *fil* (noun, -en, -er; the uter gender, 3rd declension; Eng. *traffic lane*) and *fil* (noun, -en, -ar; the uter gender, 2nd declension; Eng. *file* as in nail-*file*) are counted together in frequency statistics. The missing information about the declension of a noun or conjugation group of a verb results in a partially misleading frequency information. The verb *vara* irrespective of which one of the two verbs is meant – *to be* or *to last* – has always the same frequency value, in spite of the fact that the two verbs are conjugated differently, one being a strong verb (conjugation group 4), the other being a weak verb (conjugation group 1); they also have unrelated meanings, the meaning "*to last*" being much more rarely used. Different lexemes of the same lemma have similarly been summarized, e.g. *rom* (Eng. *caviar, roe deer, rum, Rome*). Thus, neither polysemy nor homography within the same word class have been taken into account during the lemmatization process and consequently during the frequency analysis.

Another aspect which would need further improvement in annotation of the SweWAC corpus is derivational morphology, i.e. mark-up of root morphemes and word-building affixes of each lexical item. The suggested markup could have allowed collecting frequency statistics according to the word family principle, i.e. words that share the same root being grouped together (e.g. *lära, verb* and *lärare, noun* would make the same entry). The frequency statistics summarized from SweWAC at present does not allow to group words on this principle, which means a learner that knows the verb *läsa* (Eng. *read*) cannot be assumed to know the noun *läsare* (Eng. *reader*). On the other hand, we don't believe that the word family concept is appropriate for language learners at beginner level.

Errors in frequency calculations of the homographs within the same word class of the type "*vara, verb* (Eng. *to be*) – *vara, verb* (Eng. *to last*)", though being a systematic drawback, influence only a few rare cases in Swedish and thus have to be neglected in want of a better analysis software. Multiword items that are most frequent in Swedish are marked up as units and do not add misleading information to the statistics used for L2 learners. Finally, taking derivational morphology into account is an arguable demand. Some researchers build their word frequencies upon the notion of word families but they aren't many (Gardner 2007). Thus the two features - having less frequent multiword units, phrasal verbs and idioms marked up as units and having roots and affixes marked up for each lemma - refer rather to desirable than to absolutely necessary features. Therefore, we consider word frequency statistics based on lem-pos as described here both reliable and appropriate for language learning purposes.

2.3 SketchEngine as a workbench for frequency analysis

To generate frequency-based wordlists over SweWAC, the lemmatized and POS-tagged corpus was uploaded into SketchEngine. SketchEngine offers a number of options for working with statistics. We have used the options of collecting lemma-pos lists with raw frequency alternatively with dispersed frequency.

There are three frequency measures that have been used in the Swedish Kelly-list: raw frequency (RF), relative frequency (word per million or WPM) and average reduced frequency (ARF). Raw frequency gives an absolute count of the words in the corpus. WPM is the relative count where raw frequency is divided by the total number of running words (tokens) in the corpus and then multiplied by one million. WPM is a measure which makes word frequencies from different sources/corpora comparable. ARF takes into account dispersion of the words in different subcorpora and throughout the whole corpus. If the word/lem-pos is used in only one of the subcorpora, or if the distance between the word occurrences in the whole corpus is not regular, it is not considered to be representative of the basic vocabulary, and its rank is reduced according to the formula explained in Savický and Hlavácová (2002). The measure is used to ensure that only domain-independent general-purpose vocabulary is selected, i.e. words that are frequent in a few texts of a certain domain (e.g. law or medicine) but otherwise not regularly used in all types of texts are disqualified from the general vocabulary status.

We generated two wordlists: one with lemma-tags in combination with RF; and one with lemma-tags in combination with ARF. The RF-based list from SweWAC contained 402 446 items; whereas ARF-based list contained only 232 900 items. This means that only half of the lemmas in SweWAC have qualified themselves into the general-purpose vocabulary list. We collected raw frequencies for the items on the ARF-ordered list and calculated WPM (word per million) ratio based on raw frequencies. If WPM was less than 1, the item was not included into the list. As a result we gained a list of 153 061 items.

2.4 Working on headwords

2.4.1. Word classes

The main guideline in selecting headwords for the Swedish KELLY-list was defined as "Proposal for inclusion of word types in Kelly". According to those guidelines each language team should include lem-pos with normalized spelling, avoid "language-family" principle, i.e. include derivational forms as legitimate independent items; avoid including idioms or other phraseological units; avoid proper names with a few exceptions. Homonymy, polysemy, multiword expressions (mwe) and abbreviations were left for each language team to decide upon.

The following word classes were suggested for inclusion: noun, verb, adjective, adverb, pronoun, determiner, conjunction (and subjunction), exclamation and some numerals, namely; 1-20, 30, 40, 50, 60, 70, 80, 90, 100, 1000, 1000000, 1st, 2nd, 3rd (but not 4th, 5th, ...), half, quarter, third.

The following word classes were suggested for exclusion: participle, proper nouns, foreign words, punctuation.

2.4.2. Prescriptive versus descriptive character

During our work we came to a point where we had to decide whether our list should be of a prescriptive or descriptive character. On the one hand, the aim of the project was to produce word lists for L2 learners, and in this respect the entries in the list should be of prescriptive character, e.g. incorrect spelling excluded, appropriate words selected. On the other hand, we set as our priority aim to use a modern corpus of Swedish to identify lexical items that are frequent in present-day Swedish and which therefore are necessary for the language learner to study in the first hand. Thus, if we had started applying "selection" rules based on our judgment rather than statistics, it would have been a step back and we would risk ending up with a regular list.

On the basis of this, we made a decision to keep our list descriptive in character. That entailed among other things inclusion of several alternative spellings of certain items and refusal from our part to delete certain vocabulary that didn't look "appropriate" for language learners at the pre-translation stage, e.g. words like *stalinistisk, adj*, *marxistisk, adj*, *sovjetisk, adj* (Eng. *Stalinist, adj; Marxist, adj*; *Soviet, adj*). It was planned to check our Kelly-items during the "post-translation" (validation) stage and evaluate every item in the list against translations into Swedish, and if the above-mentioned words could "prove" their basic-vocabulary status by being present in other languages, they would be kept in the final list. If, on the other hand, no other list contained these words, they would be considered for deletion from the final list. Such an approach ensured objectivity and consequence in handling all items, and not only the ones that seemed out-of-place during the initial stage.

2.4.3. Filtering of unwanted words

30% of 153 061-long list was constituted of "unwanted words and characters" that we removed automatically. By noise we understood the following groups:

a. All entries (lemmas) containing digits or other characters than letters, e.g. > < = etc. We preserved items containing underscore (_) since underscores are used in multiword items (e.g. *d_v_s, i_alla_fall*).

b. Some word classes:

- Proper names – we have assumed that these were not as important for L2 learners as lexical words. The only proper names that have been added manually to the list are the ones standing for the countries involved in the project (China, Greece, Great Britain, Italy, Norway, Poland, Russia, Sweden), and large Swedish cities (Stockholm and Gothenburg). Automatic sorting was performed after using a name tagger to differ between nouns and proper names.
- Numerals have been removed from the list on the assumption that the number of numerals in the list was too high to handle them manually whereas the most necessary numerals (43 of them) were added manually.
- Punctuation marks were removed.
- Participles were removed on the assumption that students will learn verbs and eventually learn to apply grammar rules to create participles. Another motivation was that most dictionaries, e.g. SAOL (Swedish Academy Word List), do not provide participles as separate entries; they are, instead, listed together with the verb.
- Foreign words that have been recognized by the tagger, were also removed.

Altogether 51 522 lemmas have been removed as ”unwanted words” reducing the original 153 061-long list to approximately 100 000-long list.

Final reduction in lemma-number was done automatically by collecting all morphological variants of the same lemma under one unique entry. To illustrate this, the original list contained all forms of the adjective *livlig* (Eng. *lively*):

lemma:-:POStag	Word form
livlig:-:AQPUSNIS	livligt (neutrum)
livlig:-:AQP0PN0S	livliga (plural)
livlig:-:AQPNSNIS	livlig (utrum)
livlig:-:AQC00N0S	livligare (comparative)
livlig:-:AQS00NDS	livligaste (superlative)

All the five forms referring to *livlig, adjective* (i.e. livlig:-:AQ) have been reduced to one unique entry for *livlig, adj*; all respective frequencies have been summed up resulting in one entry as follows:

ARF	RF	WPM	lemma	POS
572	907.0	7.955	livlig	AQ

The last reduction provided us with a list of 54 338 unique lemmas.

To go through a list of 54 000 lemmas isn't an easy task, therefore we cut the list at 9000-point and started working with it.

2.4.4. Manual analysis of the lemma list

During this stage we made a number of decisions about headwords and the way we want to present them, among other things abbreviations, spelling and form variants, homonymy, polysemy, stylistically marked vocabulary, multiword units, and some marginal cases as described in Johansson Kokkinakis & Volodina (forthcoming) and in Volodina & Johansson Kokkinakis (forthcoming-b). Lemmatization and tagging errors were identified and fixed, often with the help of concordance searches in SweWAC, for example the noun *fånge* (Eng. *prisoner*) was erroneously lemmatized as a non-existent noun *fångare* from its definite plural form *fångarna*. In some other cases we consulted SAOL online (http://www.svenskaakademien.se/svenska_spraket/svenska_akademiens_ordlista/saol_pa_natet/ordlista) before we made decisions on, for example, which variant should be made headword and which one provided in brackets as an alternative variant.

2.4.5. Automatic proofreading against other Swedish lexical resources

It is easy, to make omissions during a manual control. Therefore, to double-check that the resulting list contained only existing words, an automatic matching against an associative lexicon, SALDO (Borin & Forsberg, 2009), was performed. About 500 warnings were issued which were double-checked manually – certain passive verbs that didn't contain suffix "s" were corrected, e.g. *envisa→ envisas* (Eng. *to persist*); some reflexive verbs have been corrected for the reflexive pronoun *sig*, e.g. *befinna → befinna sig* (Eng. *to be present*), some missing word forms in SALDO have proven to be existing in SAOL online; other legitimate items seemed to be too modern to be present in either SAOL or SALDO, e.g. *blogginlägg* (Eng. *blog entry*).

Another automatic control was performed matching the Swedish Morphological Database, SMDB (Berg & Cederholm, 2001), which resulted in a shorter list of warnings which were taken care of manually in the same way as described above.

2.4.6. Finalizing entries for translation

Before sending the list for translation two last steps were performed:
- 85 relevant items were added; 43 numerals, 11 geographic names for partner countries, some missing names for family members, words for meals, measures, one missing weekday and some other domain-specific vocabulary items after comparison with the Swedish Lexicon for Immigrants, LEXIN.
- one last manual proofreading was performed where articles were assigned to nouns and infinitive markers to verbs; as well as consistency of headword presentation was checked.

3. Validation through translation

3.1 Homonymous and polysemous items in translation

Some teams within the project decided to disambiguate homonymous (and in certain cases polysemous) items manually prior to the translation phase to avoid multiple translations. The Swedish team decided to go after the lem-pos principle to make the process more automatic and fast. It was a part of the decision to run an experiment that will help identify number of one-to-one mappings there are between different language pairs; number of homonymous and polysemous items which can be identified through translation; and to which extent the list could expand depending on different target languages.

Yet, in certain cases we chose to add an "example" of a typical word context for the translator and eventually for the language learner, though we didn't intend to limit the translations by the provided context. We therefore left disambiguation decisions to the subjective judgment of translators.

Translators needed to provide only one translation using the most frequent alternative and to keep in mind that the list was intended for language learners. Where impossible, several translations were provided. The motivation behind the "single translation variant" approach was that items having only one meaning could be used as bidirectional translations of each other and eventually even multidirectionally between several languages, if translated accordingly. This experiment, demonstrated that this was impossible. If translators had been asked to provide several translation equivalents, it could have secured better mini-lexica. Translation of the polysemous word *rom* provides an illustrative example;
In different contexts headword *rom, noun-en* can mean a drink (Eng. *rum*), food (Eng. *caviar*), an animal (Eng. *roe-deer*), a collective name for gypsy people, or a city (*Rome*). In all the cases the noun is of a non-neuter gender, i.e. takes definite ending "-en". Some of the translators showed a "good" sense of humor choosing the meaning of "alcoholic drink" as the most appropriate translation equivalent for L2 learners. Table 1 shows the translation equivalents for the Swedish headword *rom, noun-en* in six languages:

Language	Translation of the Swedish "rom, n-en"	Meaning in English
English	rum;roe	(1) rum (drink); (2) caviar/roe deer
Greek	αβγοτάραχο	roe deer
Itlian	uova di pesce, rum	(1) caviar; (2) rum (drink)
Norwegian	rom	(as polysemous as in Swedish)
Polish	ikra	caviar
Russian	ром	rum (drink)

Table 1. Translations of the Swedish noun *rom*

According to the provided translations, the equivalents for the Swedish *rom* in the other languages are mostly used as a drink, caviar or roe deer; none of the translators has offered the alternative for the name of the city (probably because of the word class *noun* instead of *proper noun*), nor the collective name for gypsies. The translation also shows that the translated items cannot be used as translations of each other. Generalizing further, we can admit that with the exception of 5 symmetrically translated items which are mentioned later, none of the translations from the same source word in Swedish can be used as translations between the other 8 partner languages.

Totally there are 2100 unique Swedish words that have been provided with multiple translations, of those 383 items had multiple translations into more than one language. They were distributed as follows between the CEFR levels: A1 – 658; A2 – 167; B1 – 584; B2 – 627; C1 – 497; C2 – 0 items with multiple translations.
In table 2 we have collected some information on multiple translations from Swedish per target language.

Language	Multiple translations (homonyms)
English	319
Greek	1021
Italian	857
Norwegian	1
Polish	325
Russian	7

Table 2. Multiple translations from Swedish

It is quite unexpected to see only 7 multiple translations in Russian that is a more distant relative of Swedish compared to 319 multiple translations into English, a closer language family member. It points to the fact that translation process is highly subjective and the translator personality and experience influences the resulting work.

3.2 The Kelly database

To make it possible to store, analyze and compare the nine original lists and their translations a special database Kelly DB was created by Lexical Computing Ltd. Users can search for a word in a web based user interface and find out whether the word is present in the database and how it is translated into other languages.

3.2.1. Universal, common and unique vocabulary

The main reason for the database was to match original lists for each language with the eight translations into these languages to see how many words are present in all 9 languages (symmetric translations, i.e. items that can be safely used as translations of each other), how many are common to 8 languages, 7 languages, etc. and to generate the following lists:

- Words universal to all 9 languages
- Words specific for each individual language pair
- Words unique for each individual language

A symmetric pair means that the translator of one language, e.g. from English to Swedish has translated let's say *library* as *bibliotek* while the translator from Swedish to English has translated *bibliotek* as *library*. The two translations can therefore be used bidirectionally as translations of each other. A non-symmetric translation can be demonstrated by the following example:
- *angå* (Swe source item) – *regard* (Eng translation)
- *regard* (Eng source) – *betrakta* (Swe translation)

Symmetric set of translations means that (randomly or not) translators between all language pairs chose the same variants for the pairs "source word" – "target word".

It has turned out that only 5 words belong to the universal vocabulary, i.e. they are translated in symmetrical sets. These words are *music, library, sun, hospital, theory*. The constellation of the "universal" vocabulary appears to be rather random depending on translators' preferences and seems to rely on chance rather than on some linguistic reasons.

Surprisingly enough some expected words like weekdays, months, numbers, names for relatives and basic foods haven't gained the status of universal vocabulary. For example the word *bread* is (almost) symmetrically translated, with the exception of one translation where an extra variant (synonym) – *corn* – is provided. The same refers to the word *mother*: all translators into Swedish chose the variant *mor* except the one who translated it with *moder*. As far as *father* is concerned, there were different translation variants to Swedish, including *pappa*, *far* and *fader* which made translation sets asymmetrical.

The symmetrical sets for 8 and 7 languages do not seem to reveal much of a language either apart from the fact that certain languages have more variants for the same notion and therefore they do not add to the symmetry. Certain asymmetrical sets are the result of incorrect translations or different interpretation of the source words. A very interesting example is weekdays. In Chinese at least three different names for each weekday are used (depending on the translation equivalent for *week*). In Arabic there are at least two names for each weekday; which of course has made it impossible for weekdays to enter a symmetric set for 9 or 8 languages.

Absence of ordinal numerals (one, two, three, etc.) among symmetric sets for 9 or 8 languages is also rather surprising at first glance. It takes to know the other languages to see the reason why it happens that way.

The numbers for common vocabulary between different language pairs comprise symmetric pairs for each language combination. Table 3 shows the numbers for languages paired with Swedish:

Language combination	Nr of symmetric pairs
Swedish – Norwegian	3109
Swedish – English	3002
Swedish – Italian	2641
Swedish - Polish	2495
Swedish - Russian	2271
Swedish – Greek	1966
Swedish – Chinese	1123
Swedish – Arabic	618

Table 3. Common vocabulary for Swedish-X language combinations.

The numbers indicate how many entries in the two languages can be used bidirectionally.

Numbers of the common vocabulary between different language pairs seem to confirm the fact of "closeness" between the languages depending on which language family they belong to – the closer relatives the languages are, the more common vocabulary (symmetric pairs) they share. It also reflects relative similarity of the corpora from which the original lists have been derived as well as approaches to vocabulary selection.

The highest number of symmetric sets enjoys the pair Swedish-Norwegian: both languages belong to the same language family, subgroup and branch (Indo-European family, Germanic Subgroup, Northern branch). Both lists have been derived from web corpora. Swedish-English pair comes next. Both these languages belong to the same family and subgroup, the difference lies in the branch (Northern versus Western). English list has been derived on a combination of different corpora since there are many more available for English than for Swedish.
The least number of symmetric pairs is shared by Swedish and Arabic, which reflects distance between languages (Germanic vs Afro-Asiatic language families) and the principles of tokenization, lemmatization and vocabulary selection.

Unique vocabulary in this context means the items present in the monolingual list that were not used in any of the translations from other languages to the target language.

There are 501 words in the list of unique Swedish words. They represent 118 words marked for domains, while 370 come from the "exclusion list". The latter ones are kept for the reasons described later, among them are Swedish-specific words like *midsommar, pingst, nobelpris, kvällsmål, fika (*Eng. *Mid-summer, Treenity, Nobel Prize, supper, coffee break).*

The lists of universal, common and unique vocabulary may present certain interest for lexicographers, comparative linguists and other language-interested user groups and have a potential for being further exploited in linguistic analyses. The Swedish list is available for download at the Swedish Language Bank.

3.2.2. Inclusion and exclusion candidate lists

Apart from that, the Kelly database facilitated generation of the following lists necessary for post-translation editing and validation of the monolingual master lists (M):

- Candidates for exclusion for each individual language, i.e. words present in the target monolingual list but not used in any of the translations from other languages to the target language.
- Candidates for inclusion, i.e. words that have been used as translations to the target language, but are not present in the target language monolingual list.
- Multiword expressions not present in the original monolingual list, but given as translations into the target language from other languages.

3.3 Embedding the evidence

The Swedish M2 list sent for translation contained 6000 items. After processing the candidate lists generated from the Kelly DB it expanded to 8425 items. This confirmed our intuitions that translations from other languages could enrich each language with approximately 2000-3000 items.

The deletion candidate list for Swedish contained 644 candidates for exclusion, i.e. 644 lemmas that have not been used as translations into Swedish from any of the eight partner languages. We went through the deletion candidates manually, deleted 137 items from the monolingual list and kept 507, guided by the principles described in Johansson Kokkinakis & Volodina (forthcoming), the most important one being the domain of importance to language learners, e.g. *veckodag* (Eng. *weekday*), *väster om* (Eng. *to the left of*); and culturally important words for Swedish, e.g. *midsommar* (Eng. *midsummer holiday*), *fika* (Eng. *coffee break*).

We deleted items from the Swedish M2 list if the deletion candidates were words that had functional word classes, e.g. particles, determiners, pronouns; historical terms, e.g. *stalinistisk, bolsjevik, marxistisk, koncentrationsläger;* adverbs if they were "t"-derived forms from an adjective present in the M2 list, and some other groups as described in Johansson Kokkinakis and Volodina (forthcoming).

Inclusion candidates list comprised 3430 base forms. Of those, 2630 lem-pos have been added. The 3430 candidates were first automatically checked against a SweWAC lemma list, and all possible POS-tags for each item and their WPM frequencies were collected. A number of items did not match any of the lemmas in the SweWAC and were discarded as illegitimate ones. Among the latter ones there were non-lemmatized items e.g. *dikter* (Eng. *poems*), non-existent or misspelled word forms.

Due to the collected SweWAC wpm frequencies, it was possible to place all inclusion candidates relative to the items already on the Swedish list. Most of the added candidates ended up in the last two proficiency levels on the Swedish list.

Out of 530 candidate multiple word expressions (mwe), examples (as opposed to headwords) were added to 115 headwords, to 44 of those – multiple examples. Altogether 194 mwe were added to the list. We discarded non-idiomatic and unlemmatized candidate mwes e.g. *bära in* (Eng. *bring in*), *bära ut* (Eng. *take out*); *jag kan* (Eng. *I can*). We avoided inclusion of mwe as new headwords since we did not have the frequency for those.

As for the lacking domain specific vocabulary, only frequency justified topical words from the 8 languages were added in the Swedish list, thus making the selection of domain vocabulary also based on the frequency principle.

4. Coverage

4.1 General on vocabulary distribution in the Swedish Kelly-list

The 8425 headwords on the Swedish Kelly-list have been equally assigned to CEFR levels according to their frequency range in the following way:
A1, A2, B1, B2, C1 – 1404 headwords per level
A6 – 1405 headwords
With respect to their sources, the headwords are distributed in the following way:

- 85 have been added manually. They constitute 1% of the list, all belonging to CEFR A1 and cover 0,44% of SweWAC.
- 2564 headwords come from T (translation lists). They constitute 30,4 % of the Kelly-list and cover 1,7% of SweWAC texts. Approximately 2500 of those items appear in the last two proficiency levels C1 and C2, as shown in table 4
- 5776 headwords come from SweWAC. They constitute 68,5 % of the Kelly-list and cover 77,98% of the total SweWAC texts. They appear evenly (between 1305 and 1377 headwords per level) in the first four CEFR levels, and disappear at all from the last CEFR level C2, as shown in table 4.

CEFR level	**Nr of T2 words**	**SweWAC coverage, %**	**Nr of SweWAC items**	**SweWAC coverage, %**
1 (A1)	14	0,7	1305	68,9
2 (A2)	27	0,0909	1377	5,3198
3 (B1)	53	0,0882	1351	2,26
4 (B2)	69	0,12	1335	1,16
5 (C1)	996	0,495	408	0,2686
6 (C2)	1405	0,2476	0	0
Total	**2564**	**1,6739**	**5776**	**77,98**

Table 4. SweWAC coverage by T2 and SweWAC items.

Word classes distribution is presented in table 5.

POS	Total count (% of Kelly-list)	Coverage, SweWAC
Adjective	1354 (16,07%)	6,43%
Adverb	569 (6,75%)	7,6%
Aux.verb	5 (0,06%)	0,14%
Conjunction	19 (0,23%)	0,41%
Determiner	10 (0,12%)	3,6%
Interjection	24 (0,28%)	0,1%
Noun	4607 (54,68%)	14,51%
Numeral	56 (0,66%)	1,19%
Participle	1 (0,01%)	0,001%
Particle	29 (0,34%)	0,45%
Preposition	108 (1,28%)	11,14%
Pronoun	61 (0,72%)	11,4%
Proper name	13 (0,15%)	0,24%
Subjunction	31 (0,37%)	1,8%
Verb	1538 (18,26%)	16,9%

Table 5. Kelly POS distribution in SweWAC

61 pronouns covered 11,4% of SweWAC; 108 prepositions covered 11,14%; whereas 4607 nouns covered only 14,51% compared to 1538 verbs which covered 16,9%. Verbs, pronouns and prepositions therefore appears more “beneficial” to learn than of nouns in terms of text coverage, or so it would seem from statistics.

4.2 Corpora coverage by Kelly-items

We have performed coverage tests on three corpora: the core corpus SweWAC, and two control corpora - Parole and SUC.

Both Parole and SUC are well-annotated general-purpose corpora of written Swedish. Texts in Parole date from 1976-1997 and comprise newspaper texts and imaginative prose. SUC dates from 1990’s, and is a balanced corpus of written language coming in 9 genres. SUC has been manually proofread for errors in lemmatization and part-of-speech tagging.

Parameter	**SweWAC**	**Parole**	**SUC**
Size	114 mln	25,7 mln	1,16 mln
Language	2010’s	1976-1997	1990’s
Type of corpus	web-acquired	general-purpose (written) language	general-purpose (written) language
Annotation (POS, lemma)	Yes	Yes	Yes
Punctuation	10,7%	12,7%	11,5%
Infinitive marker	1,26%	1,01%	1,1%
Proper names	4,87%	8,67%	3,6%
Kelly-words	79,65%	62,75%	68,87%
Total coverage	96,5%	85,14%	85,07%

Table 6. SweWAC, Parole and SUC coverage in %.

Coverage calculations indicates that words from the Swedish Kelly-list cover 80% of the total of SweWAC, punctuation, infinitive markers and proper names stand for 16%. However, coverage calculations of the two other corpora have shown that Kelly words cover only 62,75% of the Parole corpus and 68,87% of the SUC corpus as illustrated in table 6.

A number of Kelly-items got zero-matches in the control corpora: 653 items didn't appear at all in SUC and 224 had no match in Parole. Reasons might be: (1) differences in tagging and lemmatization; and (2) difference in text genres constituting the three corpora.

(1). Lemmatization and pos-tagging of the two control corpora differ from the SweWAC-based Kelly-list. Even though Parole was tagged and lemmatized the same way as SweWAC, the headwords in the Kelly-list have undergone manually introduced changes. As a result a number of items were corrected for word class tags or lemma, for example *själv* (Eng *self*) changed pos from *adjective* to *pronoun* in the Kelly-list. In Parole *själv* is alternatively tagged (in certain cases erroneously!) as *adjective, noun* or *adverb*. Tagging differences can also be seen in POS-mismatches in such highly frequent words as *ett, det, sin, annan,* etc. that are tagged as *pronouns* in the Kelly-list as opposed to *determiner* in SUC.

A number of headwords in the Kelly-list have been modified to make them more user-friendly for L2 learners. For example, the reflexive verb *te sig* had originally been lemmatized and POS-tagged as *te, verb*, but was manually corrected during the work on the Kelly-list to *te_sig, verb*. Thus, none of the lemmas in Parole matched the Kelly-item *te_sig,* nor any other reflexive verbs for that matter. Generally, verbs appearing among zero-matches fall into two categories: the above-mentioned group of reflexive verbs (e.g. *te_sig*); and -s verbs that originally have been lemmatized without the final "-s", but have been manually corrected in the Kelly-list, e.g. *vista* vs *vistas* (Eng. *to stay*).

A big group of POS-mismatches are items tagged as *adjectives* in the Kelly-list, while having *participle* tag in SUC and Parole, among them *nuvarande, anställd, växande,* (Eng. *present, employed, growing*).

Some multiword expressions have been manually corrected by us in the Kelly-list and did not find any correspondences in either Parole or SUC, e.g. *till_slut, på_sistone, i_närheten_av, varken…eller* (Eng. *in the end, of late, in the vicinity of, either…or*).

(2). The second difference lies in the type of texts used in different corpora. Since SweWAC is a web corpus of more modern language than SUC or Parole, it shows vocabulary development of the recent decade:

- The zero-matches reflect recent "hot" political events and technological innovations, e.g. *piratparti, svininfluensa, alliansregering, islamist, taliban, reporänta, fildelare, sms* (Eng. *pirate party, swine flu, alliance government, Islamist, Taliban, funding rate, file sharer, sms*);
- The zero-matches make it obvious that the domain of web-related texts and computer technologies dominate in SweWAC, e.g. *blogga, bloggare, blogginlägg, textstorlek, postning, webbläsare, webbsida,* (Eng. *to blog, a blogger, blog entry, font size, posting, web browser, website*);
- Some other vocabulary absent in SUC and/or Parole is very colloquial in its nature and can be taken as evidence of more colloquial character of online conversation that constitute a part of SweWAC (blogs, chats, forums), e.g. *toppen, jävla, tryne* (Eng. *great, damn, snout*);
- Absence of down-to-earth learner-specific domain vocabulary in SUC can be demonstrated by the words coming to Kelly-list from translation lists, such as *krabba, socka, huva, sparv, sesam, aprikos, brorsdotter* (Eng. *crab, sock, hood, sparrow, sesame, apricot, niece*)
- One more group of zero-matches is constituted by widely spread loaned words such as *shopping, klick, mejl, kidnappning, designer, server.*

This type of check has confirmed our hypothesis about the text genres that are typical of SweWAC, namely newspaper texts, web- and computer related texts as well as blogs and forums.

To sum it up, we can claim that, had it not been for lemmatization and POS-tagging mismatches, the coverage numbers would have been increased for both Parole and SUC. Moreover, the vocabulary absent in SUC and Parole as shown in (2) above is both modern and relevant vocabulary for L2 learners.

Thus, assuming that the learner who knows words from the Swedish Kelly-list would have no difficulty coping with punctuation and infinitive markers, his/her vocabulary competence will allow understanding of approximately 90% of the texts.

5. Concluding remarks

5.1 Time aspect

The linguistics part of the project described included generation of mono- and bilingual lists during a period of 4 months of full-time work for the Swedish team. The five-step process for generation of the Swedish list took time as shown below:

1. Corpus creation and tagging– 2 months
2. Frequency lists generation via SketchEngine – 1,5 weeks full-time work
3. Working on headwords – 6 weeks full-time work
4. Translation – 4 months
5. Validation – 7 weeks full-time work

Using automatic methods is necessary when dealing with large corpora, but some automatic processes are not fully satisfactory, e.g. lemmatization, identification of multiword expressions, phrasal verbs and lexeme differentiation into the first version of the frequency list. Various types of error correction of the first version of the vocabulary list was time consuming but necessary.

5.2 The source corpus

The process of creating learner-oriented word lists should start with a well-composed and balanced corpus. The best approach is to use some available balanced representative corpus of modern language that is large enough for the task. If such corpus is not available, the web-corpus is the best and fastest alternative, though in that case we suggest that the language team be asked to provide a list of seed words. It is then possible to "design" a balanced web-corpus with seed words selected for different genres. The list of genres can be complemented as necessary; seed words for each genre carefully preselected manually or generated automatically from a shorter existing balanced corpus that contains a number of genres. Genre corpus will presumably prevent obvious gaps in learner-specific domain vocabulary, e.g. lack of words like *orange, elbow* or *alphabet*.

5.3 Multiword expressions and lexeme differentiation

Phrasal verbs, idioms and multiword expressions are definitely valuable items on any list, to say nothing of the learner-oriented lists. The question is whether existing NLP tools display sufficient accuracy.

As far as word sense disambiguation and lexeme-based frequency calculations are concerned, we are back to the fact that there are no reliable tools for Swedish at the moment that can either disambiguate word senses and collect frequency statistics per lexeme or differentiate between homography within the same word class with sufficient accuracy. However, we can hypothesize that having the same lem-pos several times in the list in different proficiency levels (i.e. homographs or different lexemes) might be confusing for a language learner. A learner who identifies a token "sentence" in a text and who has for the reason of frequencies learned only one meaning of this token, let's say within the domain of linguistic meta-language, will be baffled when he sees the item in the "legal" context: *He had his prison sentence reduced*. It is probably better to inform the learner of other possible meanings of the lem-pos the first time they come across it, so that they know they need to go back to that item and check additional meanings when they encounter it in an unknown context.

6. Future plans

We can conclude by saying that we plan to continue working with the Swedish KELLY list in the future. The way it has been compiled, it addresses a number of target user groups, including language teachers, test producers, lexicographers, comparative linguists, computational linguists, etc. In the near future we plan to set up a dynamic lexical database where different types of word lists can be extracted, e.g. items per domain, per CEFR-level, items shared by different language pairs, words that have received multiple translations etc. The users will be able to add corpora examples and translations to the items in a dynamic way. Linking this database to other lexical resources available through the Swedish Language Bank (<spraakbanken.gu.se>) the intention is to provide for automatic analysis of morphological constituents of each item and experiment with other interesting options.

Another path we want to pursue is within language teaching, among other things we plan to test how many words learners of different CEFR levels know; whether the words are assigned to the appropriate CEFR-levels; and run coverage tests on language course text books used in language courses using the CEFR.

7. References

Berg, S., Cederholm, Y. (2001). Att hålla på formerna. Om framväxten av Svensk morfologisk databas [On the creation of the Swedish Morphological Database]. In *Gäller stam, suffix och ord*. Publication in honor of Martin Gellerstam, October 15, 2001. Meijerbergs arkiv för svensk ordforskning 29. pp. 58-69.

Borin, L., Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Odense.

Council of Europe. (2001). *The Common European Framework of Reference for Languages*.

Gardner, D. (2007). Validating the Construct of Word in Applied Corpus-based Vocabulary Research: A Critical Survey. *Applied Linguistics* 28(2), pp. 241-265.

Johansson Kokkinakis, S., Volodina, E. (forthcoming). En svensk ordlista för språkinlärning – ett korpusbaserat angreppssätt i EU-projektet Kelly. *NFL 2011*, Lund, Sweden.

Kilgariff A., Reddy S., Pomikálek J. & PVS Avinesh (2010). A Corpus Factory for Many Languages. *Proceedings of LREC 2010.*

Kilgariff A., Rychly P., Smrz P. & Tugwell D. (2004). The Sketch Engine. *Proc EURALEX 2004*, Lorient, France.

Kokkinakis, D., Johansson Kokkinakis S. (1997). *A Robust and Modularized Lemmatizer/Tagger for Swedish Based on Large Lexical Resources*, Inst. F. svenska språket, Göteborgs universitet.

Källgren, G., Gustafson-Capková, S. & Hartmann, B. (2006). Manual of the Stockholm Umeå Corpus version 2.0. 85 p.

Savický P., Hlavácová J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics, 9*:, pp. 215-231.

Volodina, E., Johansson Kokkinakis, S. (submitted). Introducing Swedish Kelly-list, a new free e-resource for Swedish. *LREC 2012*, Turkey.

Volodina, E., Johansson Kokkinakis, S. (forthcoming-b). Swedish Kelly: Technical Report. The Swedish Language Bank, Gothenburg University.

What is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using Termania website

Mojca Kompara*, Peter Holozan‡
*University of Primorska, Faculty of Humanities
Titov trg 5, 6000 Koper, Slovenia
‡Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik, Slovenia
E-mail: mokopt@yahoo.com, peter.holozan@amebis.si

Abstract

The paper presents what is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using *Termania* web site. At a first step an algorithm for the automatic recognition of abbreviation-expansion pairs in newspaper texts has been used. After manual cleaning genuine pairs obtained with the algorithm were automatically included in *Termania* editing software. The editing process was automatised, non-nominal expansions were converted to nominal and language qualifiers were added automatically. The precision of the algorithm for recognition of abbreviation-expansion pairs is 96%, the algorithm for lemmatisation is successful in 70.9% of cases, the algorithm for adding language qualifiers in 95.9% of cases. Out of the obtained results we can say that simple entries are produced automatically, whereas complex "semi" automatically, because translations or explanations of abbreviation-expansion pairs can be added, at the present stage, just manually. The same goes for encyclopaedic data. For that reason we can say that the first Slovene online dictionary of abbreviations is produced "semi" automatically but further attempts to automatize translations and encyclopaedic data will be done in the future.

Keywords: abbreviations; dictionary; automatic production of dictionary entries

1. Introduction

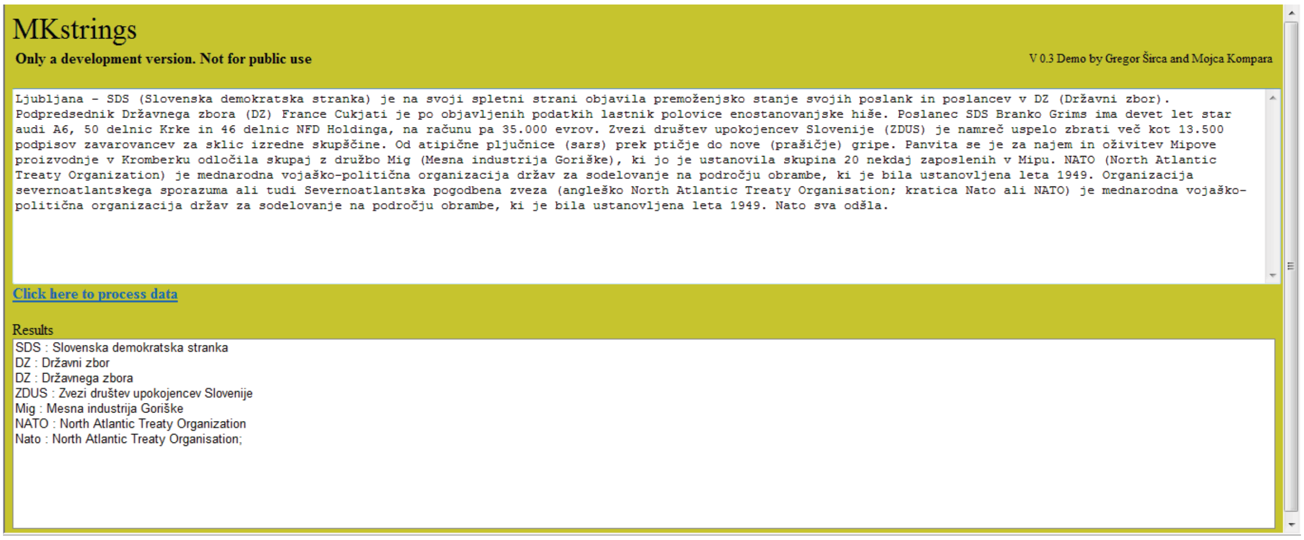
Abbreviations are difficult to deal with (Gabrovšek, 1994) and represent a growing phenomenon present in all languages. The scope of this article is to present what is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using *Termania* web site.

2. The beginning

Kazem Taghva (1998) is the pioneer in automatic recognition of abbreviations and abbreviation's expansion. Automatic recognition of abbreviations was dealt also by Yeast (1999), Larkey *et al.* (2000), Pustejovsky *et al.* (2001), Schwartz and Hearst (2003), Park and Byrd (2001), Chang *et al.* (2002) and Zahariev (2004), his approach is considered special due to the fact that he is not limiting just to one language recognitions. In the early stage of the research words up to 5 capital letters written in brackets were used as abbreviations' candidates, e.g. *(NATO)*. Words of five letters with the first letter capitalised, e.g. *(Mig)* were also used. Symbols and abbreviations such as *itd., npr., ipd., itn.* etc. were not included. In the fist stage the reference was the Slovene online newspaper *Delo* from 2007. *Delo* had 25,588 such candidates and some occurred more than once. In order to come across a genuine amount of candidates for abbreviations words that are not abbreviations, such as proper names, names of places etc. were excluded, using the *Slovene monolingual dictionary*, after the exclusions the database had over 2,500 candidates. The second step covered candidates for expansions. In order to obtain the abbreviations' expansions from newspaper *Delo,* left context was observed, because expansions are usually placed before the abbreviations, e.g. *European central bank (ECB),* but still not excluding the possibility of right context. A genuine abbreviation is determined by the expansion/s and is divided into official and/or non-official. An abbreviation can have several expansions and the algorithm took into consideration all the expansions an abbreviation could have. To recognize expansions 4 types of abbreviations were used. The first type are the so called *covered abbreviations* where letters match the words in left context, e.g. *FF* with the expansion *Filozofska fakulteta*, *Mig* with the expansion *Mesna industrija Goriške*. The second type are *abbreviations with expansions containing prepositions and conjunctions*, e.g. *FDV Fakulteta za družbene vede*. The algorithm takes into consideration also expansions with one additional word e.g. *za*. The third type concerns *abbreviations composed of the first two letters*, e.g. *NAMA Narodni magazin*. The fourth type covers *abbreviations with prepositions*, e.g. *DZU Družba za upravljanje* where prepositions appear in the abbreviation and also in the expansion. In the final list containing abbreviation-expansion pairs several problems were observed, such as the occurrence of cases, the multiple occurrences of the same expansion and abbreviations without expansions. Abbreviations without the matching expansion in the text were automatically deleted. In the manual revision that followed the most neutral case was preserved and all identical pairs appearing more than once were deleted. Considering the above mentioned criteria 1,800 expansions matched the abbreviations and formed genuine abbreviation-expansion pairs.

3. The development

In the second stage of the development the number of letters in the abbreviation was extended to 10 left and right contexts were observed. All four types of pattern: *(abbreviation) expansion, (expansion) abbreviation, abbreviation (expansion), expansion (abbreviation)* were used. Abbreviations with more than 10 words were not included. Abbreviations identical to legal words (lexicalized abbreviations), e.g. *Nama, Kad, Sod* etc. are very common in Slovene. *Nama* can be both an acronym for *Narodni magazin* or a personal pronoun at the beginning of a sentence. Such abbreviations are usually well-known but problematic and misleading for the algorithm. They were included in the algorithm rules via the dictionary of abbreviations *Slovarček krajšav*. After the newly established rules for recognition a demo version of the algorithm was produced. The system called *MKstrings* is composed of two windows, in the first one we add text rich in abbreviations, after clicking *Click here to process data* in the second window abbreviations and expansions occur as seen from Figure 1.



MKstrings

Only a development version. Not for public use

V 0.3 Demo by Gregor Širca and Mojca Kompara

```
Ljubljana - SDS (Slovenska demokratska stranka) je na svoji spletni strani objavila premoženjsko stanje svojih poslank in poslancev v DZ (Državni zbor).
Podpredsednik Državnega zbora (DZ) France Cukjati je po objavljenih podatkih lastnik polovice enostanovanjske hiše. Poslanec SDS Branko Grims ima devet let star
audi A6, 50 delnic Krke in 46 delnic NFD Holdinga, na računu pa 35.000 evrov. Zvezi društev upokojencev Slovenije (ZDUS) je namreč uspelo zbrati več kot 13.500
podpisov zavarovancev za sklic izredne skupščine. Od atipične pljučnice (sars) prek ptičje do nove (prašičje) gripe. Panvita se je za najem in oživitev Mipove
proizvodnje v Kromberku odločila skupaj z družbo Mig (Mesna industrija Goriške), ki jo je ustanovila skupina 20 nekdaj zaposlenih v Mipu. NATO (North Atlantic
Treaty Organization) je mednarodna vojaško-politična organizacija držav za sodelovanje na področju obrambe, ki je bila ustanovljena leta 1949. Organizacija
severnoatlantskega sporazuma ali tudi Severnoatlantska pogodbena zveza (angleško North Atlantic Treaty Organisation; kratica Nato ali NATO) je mednarodna vojaško-
politična organizacija držav za sodelovanje na področju obrambe, ki je bila ustanovljena leta 1949. Nato sva odšla.
```

Results

```
SDS : Slovenska demokratska stranka
DZ : Državni zbor
DZ : Državnega zbora
ZDUS : Zvezi društev upokojencev Slovenije
Mig : Mesna industrija Goriške
NATO : North Atlantic Treaty Organization
Nato : North Atlantic Treaty Organisation;
```

Figure 1: Demo version of the algorithm

As seen from Figure 1, the algorithm is not taking into consideration abbreviations such as *sars, Mipu, NFD, A6*, which is expected according to the rules stated above. Although at first sight the obtained results look really well, the algorithm was improved. Randomly selected texts rich in abbreviations (from the website *24ur.com*) were used in order to observe how the algorithm behaves. Problems occurred mainly in examples containing the abbreviation *e.g. RS* in the expansion. But also after taking into consideration this step the problem was still not solved. Prepositions *za* and *v*, represented a problem too, because at the present stage the algorithm was able to consider just one preposition or additional word in the expansion. Problems occurred also in some copy-pasted examples e.g. *Urada za varstvo konkurence (UVK)*, recognized when retyped. An interesting issue are also patterns composed of a foreign abbreviation and a Slovene expansion, e.g. *Združenje evropskih avtomobilskih proizvajalcev (ACEA)*. Such patterns were not observed in the present article and will be recognized in the future. After applying modifications and improvements the software was enlarged in order to be able to filter larger amounts of data. A larger corpus composed of 60 million words (newspaper *Delo* from 2005 to 2009) was used. The algorithm filtered the corpus in 30 minutes and gave 5,820 abbreviation-expansion pairs. The obtained pairs were manually revised and verified using *Google*. The precision of the algorithm is 96%. Recall was not retrieved because the corpus was the whole newspaper and not just a corpus made of just texts with abbreviations. Manual check up of the whole corpus would take too much time and a smaller sample would not show the real situation in the text. Among the good expansions many occurred more than once and/or with tiny modifications, e.g. usage of different cases or spelling as seen in Table 1. In Table 1 just 3 expansions out of 6 are genuine but also the genuine ones are not lemmatised (provided in nominal form) and can not be included in a dictionary as such.

MNZ 1 ministrstva za notranje zadeve 2 medobčinskih nogometnih zvez 3 ministrstvom za notranje zadeve 4 Medobčinske nogometne zveze 5 Muzeja novejše zgodovine 6 Muzej novejše zgodovine	MNZ 1 ministrstva za notranje zadeve 2 medobčinskih nogometnih zvez 3 Muzej novejše zgodovine

Table 1: Good expansions for MNZ entry

After the exclusion of false pairs (4%), verification and revision of good pairs, 2,665 genuine abbreviations-expansion pairs (in different cases) occurred. Among the good pairs there were also some foreign pairs although the recognition focused only on Slovene texts. Among the foreign some problems occurred in misrecognition of parts of expansions, e.g. *FEE for Environmental Education*, where *Foundation* is missing.

4. Termania editing software

After the genuine abbreviation-expansion pairs were obtained from the corpus using the algorithm, the next step was to include the pairs automatically into a dictionary editing software and edit the entries automatically or semi automatically. *Termania*[1] was used as exiting software. It is a free on-line dictionary portal with integrated dictionary browsing and editing tools developed by *Amebis* software company (Kamnik, Slovenia) in cooperation with *Trojina*, the Institute for Applied Slovene Studies. It provides an interface for dictionary browsing and a simple but reasonably versatile on-line dictionary editing tool. The portal is intended for general public users with no specialized computer or lexicographic knowledge, but with an interest in sharing terminological or general language knowledge, either by offering translations in a bilingual or multilingual environment or providing definitions in a monolingual context. The portal is intended to serve as the central terminology data and opinion exchange node for Slovene terminology. Access is free of charge, but does require registration (Krek, 2011). After the automatic inclusion of genuine pairs in *Termania* the editing phase started. The aim of the project was to provide an automatic editing process in simple and complex entries of the first Slovene online dictionary of abbreviations.

5. Conversion of expansions into nominal form

The main problem in automatic production of simple and complex dictionary entries are abbreviations' expansions that appear in non-nominative cases. Slovene has six cases: nominative, genitive, dative, accusative, locative and instrumental, but for dictionary purpose only nominative case is used. Another problem is the number, which can be singular, plural or dual, as well as the occurrence of other languages.

5.1 Presis Interlingua

Presis[2] is a machine translation program developed by Amebis. It supports Slovene to English, English to Slovene and German to Slovene translations and it is part of iTranslate4.eu project[3].

It is a rule-based system consisting of analyzers and generators. Analyzers translate text in Slovene, English or German to Presis Interlingua, while generators translate Presis Interlingua to Slovene or English (German to English translation is also possible, but it is used just for testing and it is not commercially available).

5.1.1 Interlingua sample

Presis Interlingua for the expansion *Plesne zveze Slovenije* is as follows:
(-POV:(-STC:(-PR2:(-SFR:(-DSF:(-PFR:(-DPF:(-PRVo:{4b0199;21f42f1}[0]<2044>))),(-JED:(-SAMe:{8d42;309123d,309128d,dfccf8,2c9be,1ba3a0d,195d56f,2dc88fe,30912e1}[1]<111c>)),(-SFR:(-DSF:(-JED:(-SAMe:{9147;28ec956}[2]<c508>)))))))))

Each element is in parentheses. It starts with status character (in this case always '-' for sentences, these are used to connect parts of a sentence to a verb template), then there is a three-letter name of element followed by colon (there can be some parameters before the colon, e.g. element *SAM* (noun) has information about number (e – *ednina* (singular)).

For complex elements there is a list of included elements, simple elements (elements connected to actual words) have the following information: inside {} is first the ID of lemma followed by semicolon which is followed by a list of IDs of possible meanings (senses) separated by commas. Inside [] is an index of a word in the original text and inside <> is the ID of morphosyntactic descriptor. All IDs are in hexadecimal.

A detailed description of Presis Interlingua and all its elements can be found in Holozan (2011). Table 2 lists just elements from the sample.

[1] http://www.termania.net/

[2] http://presis.amebis.si

[3] http://itranslate4.eu

POV	sentence
STC	single sentence element
OSB	subject
PR2	object in genitive
PRD	direct object (in English)
SFR	noun phrase
DSF	part of noun phrase
PFR	adjectival phrase
DPF	part of adjectival phrase
PRV	adjective
JED	centre of noun phrase
SAM	noun

Table 2: Some elements of Presis Interlingua.

5.2 The procedure

The idea is to use Presis Slovene analyzer to translate the expansion in question into Presis Interlingua. A special version of the analyzer, which allows only subjects and objects in various cases, is used.

If the result is a subject, it is already in nominative case and no further work is needed. If the result is an object, the translation Interlingua is changed and the object becomes the subject (e.g. *(-PR2* is changed to *(-OSB*) and then fed to the Slovene Presis generator). The resulting Slovene "translation" is the same noun phrase in nominative case.

An important part of this procedure is the meaning information that has to be removed from the translation Interlingua before sending it to the generator otherwise some words in generated nominative form can become replaced by synonyms.

5.3 Results

The expansions were processed using the above mentioned procedure and then the results were manually corrected.

A program was made to count the number of differences between automatic results and manually corrected ones. The first test showed 2531 correct expansions and 655 mistakes. Without the conversion to nominal form, the number of mistakes was 1486, in this way conversion solved 55.9% of cases (and not all mistakes were due to wrong cases, so the result is even better).

5.4 Problems

Most of the problems found can be assigned to one of the following categories.

5.4.1 Number

There are quite a few problems in the cases where nominative plural form is chosen instead of some non-nominative singular form. However, if the program changed all plurals into singulars, new problems appeared where the required expansions really should be in plural e.g. *konvencionalne sile v Evropi*, *cestno prometni predpisi*, so the final result is worse than it was without this change.

5.4.2 Definite forms of adjectives

In some cases an indefinite form of adjectives was generated instead of the definite e.g. *nacionalen energetski program*. The problem is that in Slovene definiteness is only present in nominative and accusative of singular masculine forms, for all other cases the definite and indefinite form is the same. However, for abbreviation's expansions included in a dictionary entry it is essential to use only definite forms. For that reason definiteness was added to all adjectives in Interlingua (in the form of definite article, which doesn't exist in Slovene, but forces definite forms in Slovenian generator). This update solved 27 cases.

5.4.3 Capitalisation

Capitalisation represents a problematic issue. On one hand, there are many expansions, which should be capitalized, on the other hand, even more expansions are not capitalized e.g. *socialno varstveni center*, *obnovljiv vir elektrike*, *indeks telesne mase*, and for that reason the test for automatic capitalization of all the expansions has worsened the results.

In the second run, 16 typical beginnings to be capitalized were added to the program (e.g. *zakon* (act), *fakulteta* (faculty), *slovenski* (Slovene)) and it solved 117 cases.

5.4.4 Wrong adjective/noun disambiguation

In some cases e.g. *Slovenskega ljudskega gledališča*, *Slovenskega svetovnega kongresa*, *Primorskega poletnega festivala* the analyzer made a mistake with disambiguation in considering the first word as a (proper) noun instead of an adjective. The analyzer was updated and now adds penalty for disambiguation for noun phrases where a genitive noun phrase follows proper noun. This update solved 24 cases.

5.4.5 Doubling of results

Conversion of expansions into nominal forms caused doubled expansions in some cases. For the abbreviation *RS*, the following expansions (alongside 5 others) were found and then converted to nominal form:

Računskim sodiščem	Računsko sodišče
Računsko sodišče	Računsko sodišče
računskega sodišča	računsko sodišče

Table 3: Nominal forms for the abbreviation RS.

For the abbreviation *RP*, the expansions were (alongside 3 others):

razdelilnih postaj	razdelilne postaje
razdelilno postajo	razdelilna postaja

Table 4: Nominal forms for the abbreviation RP.

In this case, *razdelilne postaje* is the plural form of *razdelilna postaja*.

5.5 Improved results

After the applied changes for conversion into nominal form and some corrections in manually corrected results (the mistakes were noticed during testing), the result is: 2661 correct expansions, 433 mistakes. Conversion solved 70,9% of the cases (compared to tests without conversion). The majority of the remaining mistakes concerns capitalization and number, but there are also some problems with unknown words.

6. Language identification

Some abbreviations have expansions in other languages, not just in Slovene. Statistical methods are often used for determining the language of a text (Dunning, 1994). However, these methods do not work well on very short text (less than 20 characters) and expansions are generally very short. We can assume quite safe, that if Presis analyser can produce the noun phrase analysis for the given text in some language, this text is in that language (assuming the number of unknown and guessed words do not exceed some percentage of text). Presis has analysers for English and German besides Slovene and this covers most of the expansions.

6.1 The procedure

In the first stage the expansion is sent to the Slovene analyzer. If the analyzer is successful, the language code “sl” is assigned; if not the language code “sl-x” is assigned (Slovene – to be manually checked).

In cases where there is still no code the English analyzer is applied and if it is successful, the language code “en” is assigned, if not the language code “en-x” is assigned (English – to be manually checked).

In the final step, if there is still no code, the German analyzer is applied. If it is successful, the language code “de” is assigned; if not the language code “de-x” is assigned (German – to be manually checked).

If no analyzer was successful and all found unknown words, the language code “xx” (unknown – to be checked and assigned manually) is assigned.

6.2 Results

Table 5 shows the required manual corrections of language tags with some examples.

xx→en	28	Codex Alimentarius Commission Bharatiya Janata Party Eco Management and Audit Sheme
xx→sl	24	Državnem inštitutu za fizkulturo Adriatica Slovenice Slovenske narodne podporne jednote
xx→fr	12	Federation Internationale des Echecs
xx→it	10	Associazione Sportiva
xx→de	6	Sport Events Steinforth Financial Times Deutschland
xx→sr	3	Crveni zvezdi
xx→hr	3	Istarsko narodno kazalište
sl→en	5	Super Proton Synchrotron New York Times John Fitzgerald Kennedy
sl→hr	3	Hrvatska stranka prava
sl→sr	3	Duvanska industrija Niš
sl→it	1	Parti socialiste
en→de	5	Deutsche Bank
en→fr	5	Electricité de France

Table 5: Corrections of language tags.

Out of 3094 language tags, 126 were assigned wrong. 95.9% of selected tags were correct. In Table 5, the final numbers of expansions for each language are shown.

Slovene and English made up most of the cases; French might be interesting to be added to the language identification since there already exists alpha version of analyzer for French for Presis.

For Croatian, Serbian, Bosnian and Montenegrin it will be difficult to make automatic language identification because of their similarity.

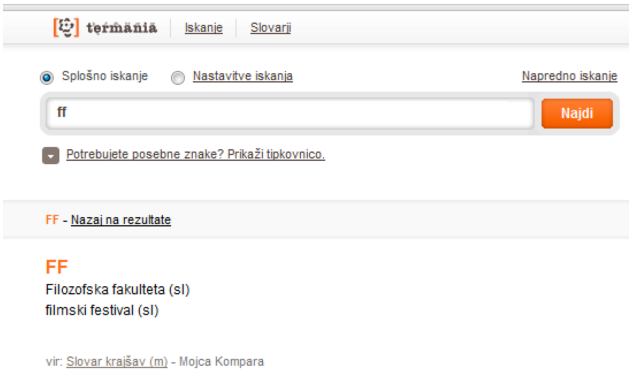
Slovene (sl)	2335
English (en)	680
German (de)	23
French (fr)	18
Italian (it)	11
Croatian (hr)	7
Serbian (sr)	7
others (es, ro, bs, eu, hu, ne, pl, mno)	13

Table 6: Number of expansions by language.

One problem that should be solved in future versions concerns foreign names included in Slovene dictionary e.g. *John Fitzgerald Kennedy*. One possibility is to try both Slovene and English analyzer and if they both produce result, the program should decide which is more probable.

7. Entries in Termania

After applying above-mentioned changes the first Slovene online dictionary of abbreviations was available online free of charge on *Termania's* web site. The entries in *Termania* can be divided into simple and complex. Simple are those composed of a Slovene pair *e.g. FF, Filozofska fakulteta* where language is provided and the expansion is in nominative case. At present the algorithm provides such entries entirely automatically, as seen from Figure 2. Such entries work perfectly well in a Slovene dictionary of abbreviations, the only embellishment we could add to such entry is some encyclopaedic data. Such data give some additional information to the user but are not essential in simple dictionary entries and can easily be omitted.



ff

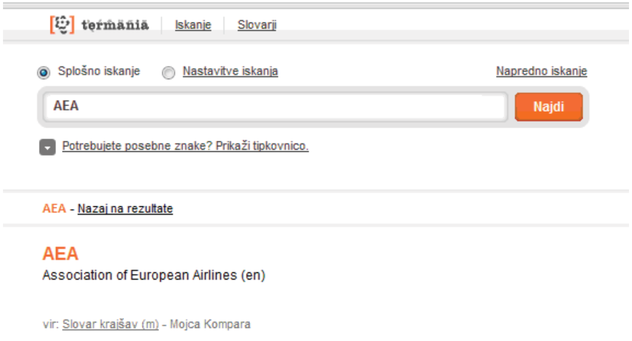
FF - Nazaj na rezultate

FF
Filozofska fakulteta (sl)
filmski festival (sl)

vir: Slovar krajšav (m) - Mojca Kompara

Figure 2: Simple dictionary entry

Complex entries are those containing foreign abbreviations where language is provided and the expansion is checked, but the Slovene translation is missing, as seen from Figure 3. A user friendly dictionary of abbreviations should include translations of foreign abbreviations and also some encyclopaedic data or a description, if there is no official translation. For now translations, descriptions and encyclopaedic data can be included only manually and for that reason such complex entries are not produced automatically, as are the simple ones. In the future we will make an attempt to provide automatic translations, descriptions and encyclopaedic data in complex entries.



AEA

AEA - Nazaj na rezultate

AEA
Association of European Airlines (en)

vir: Slovar krajšav (m) - Mojca Kompara

Figure 3: Complex dictionary entry

8. The future

In the future we will try to cope with new challenges, the first one being automatic recognition of translations or automatic translation of foreign expansions into Slovene. We will also focus on patterns composed of a foreign abbreviation and a Slovene expansion, e.g. *Združenje evropskih avtomobilskih proizvajalcev (ACEA)* and try to recognise them out of the text. In such cases lexical recognition will not work and will be switched with statistical recognition. Such patterns are frequent in newspaper texts and were not observed in the present article but will be dealt with in the near future. The last automatised procedure will be done on encyclopaedic data. Not all abbreviations will be provided with encyclopaedic data, but a selection will be made. The input of lexicographers will be essential in all phases of our dictionary development and the automatised process is used just to help the lexicographer and could never replace them.

9. Conclusion

The answer to the question what is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using *Termania* web site is: algorithms, a good knowledge of the topic and fresh ideas. In the paper we present how we automatically extract abbreviation-expansion pairs out of newspaper texts, how we clean manually and obtain genuine pairs, how we cope with the automatic editing phase and add language qualifiers to expansions and transform non-nominative expansions into nominative. We placed our work online, free of charge, on the web site of *Termania* editing software in order to share it with the users. It is the first dictionary that is produced "semi"[4] automatically from newspaper articles with the help of algorithms. Algorithms give the possibility to create a semi automatic dictionary of abbreviations and such dictionary represents the future of electronic lexicography. Algorithms for automatic recognition of abbreviations, lemmatization and adding language qualifiers present a link between the text and the semi automatic production of a dictionary of abbreviations. That is why the production and further development of the algorithm is essential and useful for lexicographers.

10. References

24ur.com. Accessed at: http://24ur.com.

Amebis Termania. Accessed at: http://www.termania.net.

Chang, J.T., Schütze, H. & Altman, R.B. (2008). Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of American Medical Informatics Association (JAMIA)*, IX(VI), pp. 612-620.

Dunning, T. (1994). Statistical Identification of Language. Technical Report MCCS 94-273. New Mexico State University, USA.

Gabrovšek, D. (1994). Kodifikacija angleškega jezika v

[4] Some procedures were done manually, such as selection of genuine pairs and cannot be automatised for now.

specializiranih enojezičnih slovarjih: Too much of everything?. (Codification of Englih in specialised monolingual dictionaries). *Vestnik*, XXVII/I-II, pp. 150-180.

Google.si. Accessed at: http://www.google.si.

Holozan, P. (2011). Samodejno izdelovanje besedilnih logičnih nalog v slovenščini (Automatic generation of textual logic puzzles in Slovenian). Master of Science thesis. Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Ljubljana, Slovenia.

Krek, S. (2011). Termania – Free On-Line Dictionary Portal. Retrieved from: http://www.simonkrek.si/Objave/5Krek_SD011.pdf.

Larkey, L., Ogilvie, P., Price, M.A. & Tamilio, B. (2000). Acrophile: An Automated Acronym Extractor and Server. In *Proceedings of the fifth ACM conference on Digital libraries 2000, San Antonio, 2-7 June 2000*. San Antonio, Texas, USA.

Park, Y., Byrd, R.J. (2001). Hybrid Text Mining for Finding Abbreviations and Their Definitions. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. pp. 126-133.

Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M. & Morrell, M. (2001). Extraction and Disambiguation of Acronym-Meaning Pairs in MEDLINE. *Medinfo*, 2001, 10, pp. 371-375.

Schwartz, A.S., Hearst. M.A (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing, 3-7 January 2003*. Lihue, Hawaii, USA.

Slovarček krajšav – Dictionary of abbreviations. Accessed at: http://bos.zrc-sazu.si/kratice.html.

Taghva, K., Gilbreth, J. (1998). Recognizing acronyms and their definitions. *IJDAR*, I(IV), pp. 191-198.

Yeates, S. (1999). Automatic extraction of acronyms from text. In *Proceedings of the Third New Zealand Computer Science Research Students' Conference, 6 April - Friday 9 April 1999*. University of Waikato, New Zealand.

Zahariev, M. (2004). A (Acronyms). PhD thesis. School of Computing Science, Simon Fraser University, Ottawa, Canada.

Understanding How Users Evaluate Innovative Features of Online Dictionaries – An Experimental Approach

Alexander Koplenig

Institute for German Language (IDS), Mannheim

E-mail: koplenig@ids-mannheim.de

Abstract

Compared with printed dictionaries, online dictionaries provide a number of unique possibilities for the presentation and processing of lexicographical information. However, in Müller-Spitzer/Koplenig/Töpel (2011) we show that – on average - users tend to rate the special characteristics of online dictionaries (e.g. multimedia, adaptability) as (partly) unimportant. This result conflicts somewhat with the lexicographical request both for the development of a user-adaptive interface and the incorporation of multimedia elements. This contribution seeks to explain this discrepancy, by arguing that when potential users are fully informed about the benefits of possible innovative features of online dictionaries, they will come to judge these characteristics to be more useful than users that do not have this kind of information. This argument is supported by empirical evidence presented in this paper.

Keywords: dictionary use; empirical research; online dictionary

1. Introduction

In contrast to the long tradition of printed dictionaries, online dictionaries are a very recent phenomenon in lexicography. In addition to the classical criteria of reference books (such as reliability of content), online dictionaries have unique and innovative features resulting from the possibilities of the electronic medium, e.g. multimedia applications or a user interface that is customizable (de Schryver, 2003). Thus, it is not clear *a priori* how users judge the potential usefulness of those features. In this paper, we present and discuss the findings of an interdisciplinary research project that tries to investigate this question by applying various methods of empirical social research (Müller-Spitzer et al., 2011).

The first studies carried out as part of our project combined online surveys with experimental elements. Throughout this project, we plan to conduct some lab usability tests including an eye-tracking study. Based on this research agenda, we aim to illustrate these points by showing how empirical research can foster new insights and increase our understanding of dictionary use.

This paper is structured as follows. In section 2, we give a short overview of the project background and the hypothesis to be tested. Sections 3 and 4 describe the experimental approach and the methodological procedure, while section 5 presents the result. Finally, this study concludes with a brief discussion of the implications of the findings (section 6).

2. Background and Hypothesis

To identify different user demands, we conducted two online surveys in English and German in 2010. A total of 1,074 respondents participated. Among other questions, respondents of the first study (N = 684) were asked to rate ten aspects of usability with respect to their importance regarding the use of an online dictionary and then to rank these aspects in order of importance. The classical criteria of reference books (e.g. reliability, clarity) were both rated and ranked highest, whereas the unique characteristics of online dictionaries (e.g. multimedia, adaptability) were rated and ranked as (partly) unimportant[1] (Müller-Spitzer et al., 2011). This result conflicts with the general lexicographical request both for the development of a user-adaptive interface and the incorporation of multimedia elements to make online dictionaries more user-friendly and innovative (de Schryver, 2003; Müller-Spitzer, 2008).

We assume that one possible explanation for this result is the fact that respondents are not used to online dictionaries incorporating those features. Thus, respondents currently have no basis on which to judge their potential usefulness. This line of reasoning predicts a learning effect. That is, when users are fully informed about possible multimedia and adaptable features, they will come to judge these characteristics to be more useful than users who do not have this kind of information.
To test this assumption, we incorporated an experimental element into our second survey.

3. Experiment

The participants of our second online study were presented, both visually and linguistically, with several possible multimedia applications and various features of an adaptable online dictionary in a set of statements (S1). Each feature was explained in detail (cf. Table 1) and/or supplemented by a picture illustrating its potential function. The participants were then asked to rate each feature with respect to three different characteristics regarding the use of an online dictionary (importance/benefit/helpfulness).

[1] Analysis of correlation revealed a significant association between importance and ranking; r = 0.39 [0.20; 0.56]; p < .01.

Domain	Feature	Explanation
Multimedia	Audio pronunciations	In contrast to a printed dictionary, an online dictionary can include audio files illustrating the pronunciation of a word, a phrase or a whole sentence.
	Collocation graphs	An online dictionary can represent collocations, i. e. frequently occurring word combinations, in a visual form.
	Illustrations	An online dictionary can contain illustrations.
User adaptability	Customized user interface	To facilitate access to relevant personal information, the user interface of the online dictionary automatically adapts to the user's preferences depending on the item classes used in previous search requests. In this process, the online dictionary "remembers" a particular item class. For example, if the user frequently consults an online dictionary to search for synonyms, then a special search window for this kind of request appears on the homepage. A widely known commercial example is the homepage of the mail-order company Amazon, which changes according to the user and his/her previous shopping preferences.
	Alternative profiles	This means that the user of the online dictionary can choose between different profiles that optimally adjust the content according to the user's needs. For this purpose, the user first chooses between different user types and/or different usage situations. Certain defaults are then used to structure the mode of content presentation.
	Dynamic visual representations	This refers to the possibility of creating a personalised user view of the online dictionary. This can be done by choosing between different item classes, e. g. paraphrase, sense relations, information on grammar or citations.

Table 1: Features of user adaptability and multimedia presented in the survey

In a second set (S2), participants were asked to indicate how much they agree with the following two statements: *The application of multimedia and adaptable features ...*

(A) ... makes working with an online dictionary much easier.
(B) ... in online dictionaries is just a gadget.

To induce a learning effect, we randomized the order of the two sets: participants in the learning-effect condition (*L*) were first presented with the examples in S1. After that, they were asked to indicate their opinion in S2. Participants in the non-learning-effect condition (*N*) had to answer S2 followed by S1. Thus, to judge the potential usefulness of adaptability and multimedia, the participants in the learning-effect condition could use the information presented in S1, whereas the participants in the non-learning-effect condition could not rely on this kind of information. If our assumption is correct, participants in the learning-effect condition *L* will judge adaptability and multimedia to be more useful compared with participants in the non-learning-effect condition *N*.

Another objective of the experiment was to assess whether the size of this difference depends on further variables, especially the participants' background (linguistic vs. non-linguistic[2]) and the language version of the online survey as chosen by the participants (German vs. English).

4. Method

Three hundred and eighty-one people participated in the bilingual online survey on the use of online dictionaries. Participants were randomly assigned to one of the conditions of the 2 (learning vs. non-learning-effect) factorial design.

The dependent variables were measured as described above (S2). Both ratings were made on 7-point Likert scales (1 = *strongly disagree*, 7 = *strongly agree*). The answers to these two items were averaged and oriented in the same direction to form a reliable scale of adaptability and multimedia benefit judgments ($\alpha = .75$), with higher values indicating more benefit.

5. Results

An ANOVA yielded a significant effect of outcome, $F(1, 379) = 12.27$, $p < .001$. As hypothesized, the results showed that participants in *L* judged adaptability and multimedia to be more useful ($M = 5.02$, $SD = 1.30$, $N = 175$) than participants in *N* ($M = 4.50$, $SD = 1.54$, $N = 206$).

[2] We asked the participants whether they work as a linguist and whether they study linguistics (yes/no).

In order to better interpret these results, we conducted a three-way ANOVA with condition, background and language version as independent factors. The statistical analysis revealed significant main effects for condition (F(1, 373) = 18.29, p < .01), for background (F(1, 373) = 8.75, p < .05), and for language version (F(1, 373) = 13.56, p < .01). In addition, a significant three-way interaction between experimental condition, background, and language version was found (F(3, 371) = 7.59, p < .05); cf. table 2.

Post hoc comparisons using the Tukey HSD test indicated that the mean difference in the German language version between the conditions was significant for the non-linguists (p < .05) and insignificant for the linguists (p = .99), whereas the difference between the two conditions was highly significant (p < .00) for the linguists and insignificant for the non-linguists (p = .41) in the English language version.

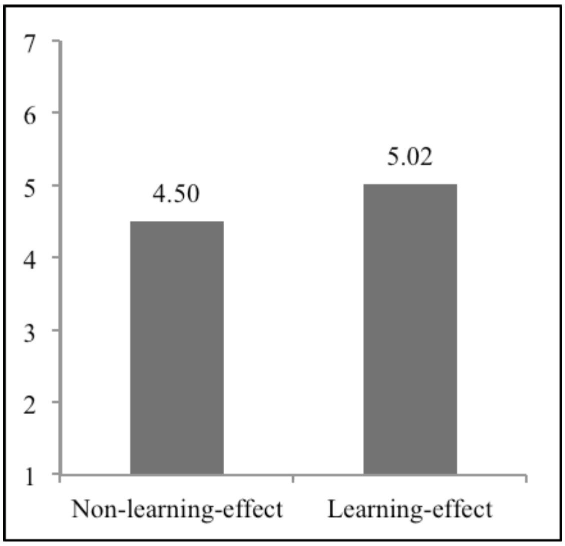


Figure 1: Means of adaptability and multimedia benefit judgments as a function of the learning-effect condition. *Note*. Means are on 7-point scales with higher values indicating higher levels of judgments of benefit

German Language Version		
	Background	
	Linguistic	Non-Linguistic
Condition		
Non-learning-effect	5.02 (1.47)	**4.45** (1.66)
Learning-effect	5.02 (1.18)	**5.09** (1.35)
English Language Version		
	Background	
	Linguistic	Non-Linguistic
Condition		
No-learning-effect	**4.23** (1.47)	4.12 (1.63)
Learning-effect	**5.15** (1.26)	4.45 (1.50)

Table 2: Means of adaptability and multimedia benefit judgments as a function of condition, background and language version. *Note.* Means are on a 7-point scale, with higher values indicating higher levels of benefit. Significant differences in bold. Standard deviations in parentheses

6. Discussion

As predicted, the results revealed a learning effect. This effect turned out to be modest in size (about half a point on a 7-point scale), but highly significant: from a Bayesian point of view, the probability of observing this difference, given that it is just due to a random variation, is 0.05%. Furthermore, it should be noted here that we implemented only a weak manipulation of the learning effect. Due to the nature of our survey design, we simply presented several features of multimedia and adaptability. It seems plausible to assume that, if the participants had the opportunity to actually use the presented features, the observed learning effect would be even bigger.

However, a closer inspection (cf. Table 2) showed that this difference is mediated by linguistic background and language version: while there is a significant learning effect in the German version but only for non-linguists, there is a highly significant learning effect in the English version but only for linguists. This leaves room for further studies focusing on the reasons for this interaction effect.

7. Acknowledgements

This project is financed by the Gottfried Wilhelm Leibniz Scientific Community ("Joint Initiative for Research and Innovation"). I thank one anonymous reviewer for a helpful comment on the title of this paper.

8. References

de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, *16*(2), 143-199.

Müller-Spitzer, C. (2008). Research on Dictionary Use and the Development of User-Adapted Views. In A. Storrer, A. Geyken, A. Siebert, & K.-M. Würzner (eds.) *Text Resources and Lexical Knowledge Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008.* Berlin: de Gruyter, pp. 223-238.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2011). What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. Presented at the eLexicography in the 21st century: new applications for new users, organized by Trojina, Institute for Applied Slovene Studies, Bled, Slovenia, November 10-12, 2011.

GDEX for Slovene

Iztok Kosem[1], Milos Husak[2], Diana McCarthy[2]
[1]Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia
[2]Lexical Computing Ltd., Brighton, UK
E-mails: iztok.kosem@trojina.si, milos.husak@sketchengine.co.uk, diana.mccarthy@sketchengine.co.uk

Abstract

Good Dictionary Examples or GDEX is a tool in the Sketch Engine designed to help lexicographers with identifying dictionary examples by ranking sentences according to how likely they are to be good candidates. The ranking is done automatically using various syntactic and lexical features. So far, only GDEX for English has been available. This paper presents the design and evaluation of Slovene GDEX, which was used for finding good examples for the new lexical database of Slovene, one of the activities in the Communication in Slovene project. Several different GDEX configurations were designed, evaluated and compared. The evaluation involved examining sentences of lemmas belonging to different word classes. Good sentences were logged for subsequent analysis with external data-mining software, WEKA. The observed behaviour was then used to adjust the parameters of the GDEX classifiers. We believe that the procedure of identifying features of good examples and their values, described in this paper, can be used for the development of GDEX for any language.

Keywords: dictionary example; GDEX; lexical database for Slovene; Sketch Engine; data mining

1. Introduction

Examples are a very important part of a dictionary entry, as they illustrate how the word is used in a particular meaning, construction or pattern. Examples provide additional support to the definition, which is sometimes hard to understand without reading the examples (Atkins & Rundell, 2008). Furthermore, examples can be of great help with navigating through longer entries, where the users can “identify the particular sense they are seeking by finding examples that are similar to the one they need or have in front of them” (Fox, 1987:137).

Good dictionary examples have to be natural and typical, informative and intelligible (Atkins & Rundell, 2008). Taking all these criteria into account makes finding a good dictionary example a time-consuming task for a lexicographer, as the search for a good example requires the inspection of a number of different features. These features include sentence length, full-sentence form, non-complex structure, and lack of rare words and/or anaphora. As corpora grow bigger and bigger, lexicographers have more sentences to choose from, so there are more likely to find good examples; on the other hand, this also means they need to inspect more examples.

Good Dictionary Examples or GDEX (Kilgarriff et al. 2008) is a tool in the Sketch Engine (http://the.sketchengine.co.uk) designed to help the lexicographers identify dictionary examples by ranking of sentences according to how likely they are to be good example candidates. The ranking is done automatically using various syntactic and lexical features. The usefulness of GDEX for English has been confirmed on an actual dictionary project, namely when selecting additional examples for the online version of the Macmillan English Dictionary. However, there were still parts of the heuristics that were identified as open for improvement. Furthermore, as most of the GDEX settings were English-specific, the usefulness of GDEX for other languages was limited.

This paper presents the development of GDEX for Slovene that was used in the building of the new lexical database of Slovene. First, the basic information on GDEX is presented, including an overview of the characteristics of examples that can be measured. Then, GDEX for English is discussed in more detail. Next, the design of GDEX for Slovene is provided, including a description of our approach for devising the heuristics. Also, the evaluation process of GDEX for Slovene is presented in detail, from the comparison of different configurations to the evaluation of the effectiveness of GDEX at words from different word classes. In conclusion, the lessons learned during the design of GDEX for Slovene are summarized and future plans are laid out.

2. GDEX

GDEX is a tool for ranking sentences according to specified criteria. It was designed for use by the Sketch Engine (Kilgarriff et al, 2004) to sort concordances in a way that is useful to lexicographers when creating dictionaries. The aim is to separate good candidates for dictionary examples from the bad candidates.

The most important criteria of good dictionary examples are usage typicality, informativeness and intelligibility (Kilgarriff et al., 2008), however these are difficult to describe and measure directly, therefore GDEX circumvents the problem by measuring observable features, such as sentence length, word length, presence/absence of black/whitelisted words/non-words (urls, numbers, etc.) which are related to the more covert criteria. Using corpora, GDEX can take into account word frequencies, common collocations and available word attributes (e.g. part-of-speech, lemma, grammatical tags, etc.). Depending on the resources available for the language or domain of the text, it is also possible to use other sources of linguistic information such as the degree of ambiguity of words contained in the sentences.

Originally, GDEX was developed as a set of classifiers for specific features which each performed normalization and scoring in a fixed way and returned a score in the range from 0 to 1. These scores were then combined in a weighted average to provide a single score for each sentence. The number of parameters for each individual classifier was limited to facilitate their automatic optimization based on the training data. GDEX has subsequently been adapted so that the individual values of features (for example, the number of words in a sentence) can be accessed directly and any normalization and aggregation is now fully customizable. This makes it possible to manipulate GDEX more precisely.

The normalization can be performed with respect to sentence length, corpus size or a fixed interval of values. In case of measuring features of individual tokens (e.g. word frequency, collocation score) it is also necessary to decide whether to take an average, minimum, maximum or sum of all values of the tokens in a sentence. The aggregation functions are used to combine the measurements into a single value that is used for sorting the sentences according to their suitability to serve as a dictionary example. The specification of measured features and the way they are combined together is defined in files called GDEX configurations.

2.1 GDEX for English

The original GDEX configuration for English was based on a set of concordances with good sentences manually annotated. Using this data, a set of various classifiers was optimized so as to rate the sentences which were marked as good higher than the others.

It became apparent that the most successful criterion was the number of long words, which tend to be harder to understand. This single classifier improved the relative position of good sentences in the test concordances by 34% (100% is the case where all marked good sentences are at the top positions higher than any non-selected sentence, and 0% improvement corresponds to a random ordering of sentences).

After several experiments with different combinations of analysed features, the following subset of all classifiers was chosen with an overall improvement of 49% over the random baseline. Besides the long words penalization, the following classifiers were used in the collection:

- penalization for interpunction marks, brackets and apostrophes;
- preference for sentences with length within an optimal interval (between 6 and 28 tokens (i.e. words and interpunction marks));
- penalization for proper nouns (based on capital letters);
- penalizaton for multisense words (based on the number of WordNet (Fellbaum, 1998) synsets that the word belongs to);
- fraction of low frequency words within the sentence (the best results were achieved with a threshold frequency of 107 in the British National Corpus (Leech, 1992), which corresponds roughly to sentences focusing on the 35,000 most common words);
- relative keyword position in the sentence (in the training data the best results were achieved with the keyword at the beginning of the sentence);
- penalization for mixed symbol words (such as email address, urls, etc.);
- penalization for anaphoric expressions (based on a word-list);
- preference for complete sentences (starting with a capital letter and ending with ".", "?" or "!");
- since GDEX was intended to be used on corpora collected from the internet, a classifier that completely banned sentences mistakenly containing html or xml markup;
- for practical reasons, as the sentences proposed by GDEX were intended for publication in dictionaries, we also added a blacklist for sensitive and offensive nouns.

Values of all classifiers were then balanced one-by-one using an eager algorithm to obtain weights for a weighted average that was used as the main aggregation function.

The presumption was that after adjusting the parameters of individual classifiers, a similar set of classifiers would give good results for Slovene as well. Therefore, we used this as a start point for the project, but without language dependent blacklists and the polysemy classifier, for which no data were available.

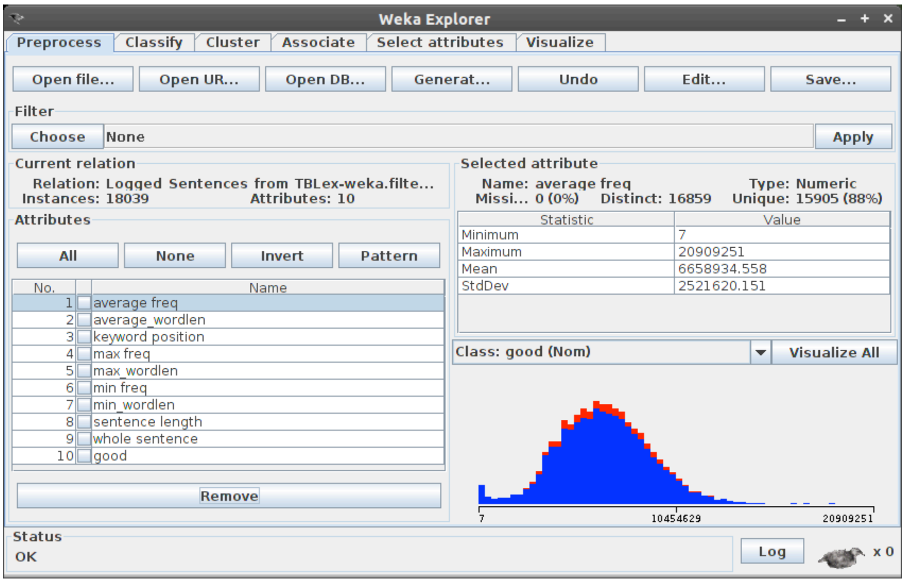
3. GDEX for Slovene

There are two ways for devising a GDEX configuration for a new language. The first, which was pursued in the original work (Husak, 2008), depends on training classifiers on annotated data. The human input in that case is limited (besides annotating data) to the choice of classifiers.

The other approach, pursued in this project, arose from the assumption that humans experienced in lexicography can provide a useful set of heuristics based on their intuition or knowledge. We used a small amount of annotated data, far smaller than would be required for machine learning, with an external tool that helps visualize the data and provide the user with additional statistical information.

The tool of choice was WEKA (Hall et al., 2009), because not only can it help in visualizing the data and providing the statistics, but it also contains extensive functions for data filtering and manipulation. The other reason was that machine learning algorithms implemented in WEKA open many new possibilities for future experiments.

WEKA itself does not work with corpora, therefore we use GDEX analysers to do the measurements, which we export into an attribute-relation file format (ARFF) that can be opened by WEKA. In WEKA, we can immediately see the minimum, maximum, mean value, standard deviation and value distribution for each of the analysed features (Figure 1). More importantly, any two measurements can be plotted in a 2D chart that shows how well they separate good and bad sentences (Figure 2). Using this information, one can make sense of the data and make better decisions when creating a GDEX configuration.



Filter: None

Relation: Logged Sentences from TBLex-weka.filte...
Instances: 18039 Attributes: 10

Name: average freq Type: Numeric
Missi... 0 (0%) Distinct: 16859 Unique: 15905 (88%)

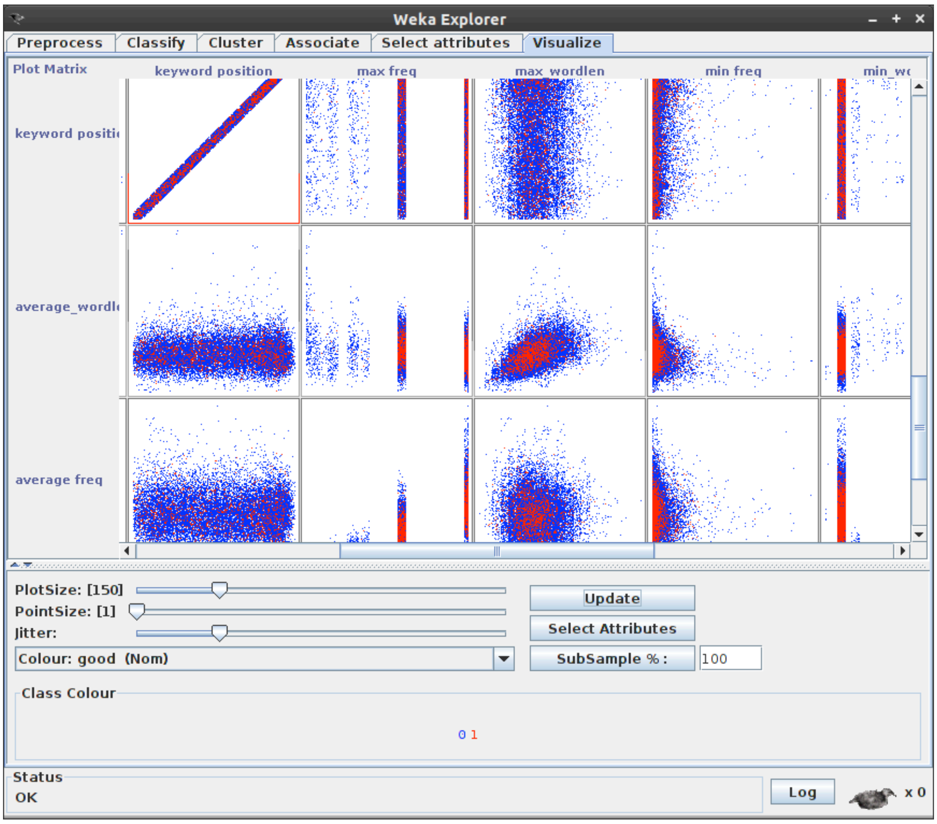
Statistic	Value
Minimum	7
Maximum	20909251
Mean	6658934.558
StdDev	2521620.151

No.		Name
1	□	average freq
2	□	average_wordlen
3	□	keyword position
4	□	max freq
5	□	max_wordlen
6	□	min freq
7	□	min_wordlen
8	□	sentence length
9	□	whole sentence
10	□	good

Class: good (Nom)

Status: OK

Figure 1: Different features of good examples (left) and the statistics for average frequency (right) in WEKA



PlotSize: [150]
PointSize: [1]
Jitter:
Colour: good (Nom)
SubSample % : 100

Status: OK

Figure 2: WEKA 2D visualisation charts of good (red) and bad (blue) examples according to different features

3.1 Design

The design of GDEX for Slovene was motivated by the needs of lexicographers working on the entries of the new lexical database for Slovene (Gantar & Krek, 2011)[1]. Their work involved selecting a great number of examples for each entry from the 620-million-word FidaPLUS corpus – each construction, pattern, phrase etc. had to be attested with at least one example from the corpus, though more than one was preferred.

Initially, the GDEX for English settings were used when selecting the examples, however this proved to be ineffective as good examples were rarely found among 10-15 examples (this setting was most commonly used) for a collocate in the word sketch. The survey among the lexicographers working on the project showed that the use of GDEX was even counter-productive – most of them switched the GDEX feature off or stopped using the TickBox Lexicography feature in Word Sketch. It was thus decided to devise a GDEX configuration that would take into account the linguistic characteristics of the Slovene language when ranking the examples. An important criterion considered when designing and evaluating configuration(s) was the needs of the project; namely, the purpose was to find good examples for a lexical database rather than a dictionary.

The design of GDEX for Slovene consisted of the following stages: selecting classifiers for the first GDEX for Slovene configuration, determining the values of classifiers, evaluating the configuration on the word sketches of selected lemmas, devising an improved configuration, evaluating the two configurations and comparing their results, devising the third configuration based on the findings, evaluating and comparing the results of the configurations, etc.

The classifiers used in configurations for GDEX for English were used as a point of departure, excluding English-specific classifiers such as polysemy (WordNet synsets), blacklists of offensive nouns, and lists of anaphoric expressions. The values of certain classifiers, e.g. preferred sentence length, were determined with the WEKA tool, using the existing examples in the lexical database for Slovene, which were selected manually by lexicographers, as a benchmark.

The first configuration, named Slovene1, had the following heuristics:

- preferred sentence length: 8 to 30 words;
- threshold of low frequency words: 104;
- keyword position: beginning of the sentence (in the first 20% of the tokens);
- penalty for words containing regular expressions;
- penalty for sentences containing urls, email addresses, etc.;
- penalty for sentences containing capital letters;
- penalty for proper nouns;
- penalty for pronouns;
- a good example had to be a whole sentence;
- a good example could not contain words that occur less than three times in the FidaPLUS corpus.

Higher weight (value of 2 rather than the normally used 1) was attributed to the preferred sentence length, threshold of low frequency words, proper noun penalization, and pronoun penalization as these features were reported by the lexicographers as both the most crucial criteria in identifying good examples, and the most indicative for identifying elements of bad examples.

When selecting examples, lexicographers are also looking for diversity, so that each selected example offers a different type of information. So if GDEX produces 10 good examples, which are all very similar, the lexicographers will probably select only one of them – the other examples are treated as "bad" ones due to their similarity to the selected example. To ensure diversity of examples offered by GDEX, a script was included in GDEX that ensured that the difference between the examples was at least 30% measured in the Levenshtein distance (Levenshtein, 1966).

Four other configurations (Slovene1b, Slovene2, Slovene3, and Slovene3b) were devised in succession during the process of evaluation. These configurations were basically variants of the initially devised Slovene1, as they included minor incremental tweaks to the classifier values and/or weights.

3.2 Evaluation

Evaluation was an important part of the GDEX for Slovene design process, since it helped to determine the efficiency of the configuration, and to identify frequent features of good and bad examples and suggest further improvements (i.e. tweaking) to the configuration.

Evaluation consisted of the manual examination and selection of good examples of collocates in the word sketches of selected lemmas. The evaluation was conducted on word sketches because it allowed us to log the selected examples using TickBox Lexicography, analyse them and identify any required changes in the values of classifiers, or create completely new classifiers. Another reason for using word sketches was to simulate the actual conditions in which GDEX for Slovene would be used. Word sketches are the main source of entry information in the lexical database for Slovene; once the

[1] The new lexical database for Slovene is part of the Communication in Slovene project (http://www.slovenscina.eu) which is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. The operation is being carried out within the operational programme Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013.

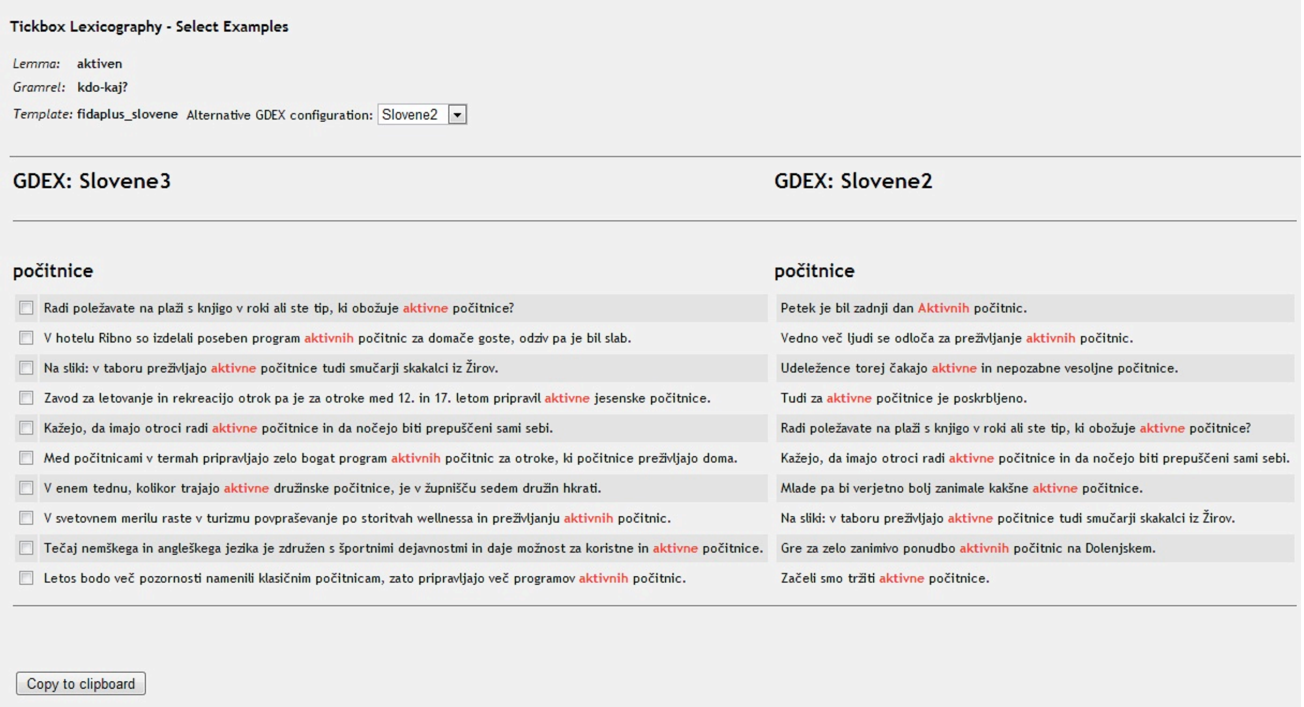
lexicographers make an initial meaning division of the headword, they analyse the word sketch to obtain information on grammatical constructions, collocates, and patterns of the headword, and extract examples.

The default Word Sketch settings were adapted for evaluation purposes. The number of examples per collocate for the TickBox Lexicography view was set to 10, as this was the recommended setting, and the most frequently used setting of lexicographers working on the lexical database. On a related note, the minimum frequency of a collocate, i.e. the number for the word sketch from which the TickBox Lexicography selection is made, was set to 15, since there was little sense in examining examples of collocates with frequency of less than 15, given that the lexicographers can quickly examine (all) the examples of such collocates, and that the only change made by GDEX in such cases is the order in which the examples are provided.

Lemmas used for the evaluation were selected from the list of existing entries in the lexical database for Slovene. As the lexical database currently contains only nouns, verbs, adjectives, and adverbs, the list of lemmas included only words from these four word classes. The aim was to make the selection as heterogeneous as possible, so the lemmas included both abstract and concrete nouns, monosemous and (highly) polysemous words, words with few and many constructions, patterns, etc.

The evaluation was conducted by two people who examined and selected the examples offered by GDEX, and wrote their observations, comments and suggestions in a shared online document. When making comments and suggestions, the evaluators had to consider which measurable characteristics of examples, i.e. the classifiers of GDEX configurations, needed improvement, had to be given less/more weight, or had to be added to the configuration. This was then taken into account when devising new configurations.

It was considered important that the evaluators were able to compare the results of different configurations in order to quickly determine any improvements in results, if any, produced by newer configurations. For this reason, a special setup of GDEX in Tickbox Lexicography was designed to allow side-by-side comparison of example ranking by two different configurations (Figure 3), the left-hand column showing the results of the currently selected GDEX configuration, and the right-hand the results of the configuration used for comparison. Good examples could be selected for the current configuration only. The configurations, and their results, in the right-hand column could be quickly changed by selecting another configuration from the drop-down menu.



Tickbox Lexicography - Select Examples

Lemma: aktiven

Gramrel: kdo-kaj?

Template: fidaplus_slovene Alternative GDEX configuration: Slovene2

GDEX: Slovene3	GDEX: Slovene2
počitnice	**počitnice**
☐ Radi poležavate na plaži s knjigo v roki ali ste tip, ki obožuje aktivne počitnice?	Petek je bil zadnji dan Aktivnih počitnic.
☐ V hotelu Ribno so izdelali poseben program aktivnih počitnic za domače goste, odziv pa je bil slab.	Vedno več ljudi se odloča za preživljanje aktivnih počitnic.
☐ Na sliki: v taboru preživljajo aktivne počitnice tudi smučarji skakalci iz Žirov.	Udeležence torej čakajo aktivne in nepozabne vesoljne počitnice.
☐ Zavod za letovanje in rekreacijo otrok pa je za otroke med 12. in 17. letom pripravil aktivne jesenske počitnice.	Tudi za aktivne počitnice je poskrbljeno.
☐ Kažejo, da imajo otroci radi aktivne počitnice in da nočejo biti prepuščeni sami sebi.	Radi poležavate na plaži s knjigo v roki ali ste tip, ki obožuje aktivne počitnice?
☐ Med počitnicami v termah pripravljajo zelo bogat program aktivnih počitnic za otroke, ki počitnice preživljajo doma.	Kažejo, da imajo otroci radi aktivne počitnice in da nočejo biti prepuščeni sami sebi.
☐ V enem tednu, kolikor trajajo aktivne družinske počitnice, je v župnišču sedem družin hkrati.	Mlade pa bi verjetno bolj zanimale kakšne aktivne počitnice.
☐ V svetovnem merilu raste v turizmu povpraševanje po storitvah wellnessa in preživljanju aktivnih počitnic.	Na sliki: v taboru preživljajo aktivne počitnice tudi smučarji skakalci iz Žirov.
☐ Tečaj nemškega in angleškega jezika je združen s športnimi dejavnostmi in daje možnost za koristne in aktivne počitnice.	Gre za zelo zanimivo ponudbo aktivnih počitnic na Dolenjskem.
☐ Letos bodo več pozornosti namenili klasičnim počitnicam, zato pripravljajo več programov aktivnih počitnic.	Začeli smo tržiti aktivne počitnice.

Figure 3: Side-by-side comparison of two GDEX configurations in TickBox Lexicography (Sketch Engine)

3.2.1. Evaluation criteria

The main evaluation criterion was the number of good examples per collocate. As the selected examples were logged, such information was easy to obtain. Initially, the aim was to devise a configuration that would yield at least 5 good examples out of 10 per collocate. However, after evaluating several different configurations, it became clear that 3 examples per collocate was a more realistic aim. Such a figure was also considered acceptable because the number of examples per collocate, construction, or pattern in the lexical database rarely exceeds two.

The evaluators selected good examples considering the criteria of naturalness, typicality, and intelligibility. Informativeness was attributed less importance due to its close relation to the meaning division – "an informative example is one that complements the definition and helps the user understand it better" (Atkins & Rundell, 2008:460) – given that automatic word sense disambiguation is not possible (yet), we could not use meaning division as a criterion for selecting examples in GDEX evaluation; in other words, all meanings of the word had to be considered.

Examples were also considered good (for the lexical database) if they showed the potential to be turned into good dictionary examples. Consequently, minor breaking of the "good dictionary example" principles was tolerated, especially the ones related to the length and complexity of examples – e.g. sentences were allowed to be longer, i.e. they were allowed to have more context that could be reduced or removed, sentences could have "removable" relative clauses with less frequent words or proper nouns, etc. This of course meant that we made the decision to allow for subsequent modification of corpus examples for dictionary purposes; in fact, as evidence suggests, such practice cannot be completely avoided even in corpus-driven dictionary projects (Atkins & Rundell, 2008; Landau, 2001; Krishnamurthy, 1987).

3.2.2 Findings

The evaluation showed that certain classifiers played a much more significant role than others in the production of good examples by the configurations. These classifiers were preferred sentence length, relative keyword position in the sentence, penalty for keyword repetition, penalty for words exceeding the prescribed maximum length, and penalty for sentences exceeding maximum length. The parameters of these classifiers were consequently given more focus and were subject to changes when devising new configurations. The values of the aforementioned classifiers in different configurations, which also point to the differences between the configurations, are shown in Table 1.

The most significant classifier for good example identification was sentence length. In the first three configurations, the preferred sentence length was between 8 and 30 words, but the evaluation pointed to the lack of good examples, mainly on account of examples being too short. Shorter sentences often proved to lack context, whereas longer sentences were more often better examples, or at least had more potential to be turned into good examples. Once the preferred sentence length was increased (configurations Slovene3 and Slovene3b), and more importantly, once the minimum length was increased to 15, the average length of examples increased (e.g. compare the examples of two configurations in Figure 3), and the number of good examples per collocate improved considerably. The improvement was observed at almost all the lemmas used in the evaluation, regardless of word class.

Another key classifier for good example detection was relative keyword position. The initial criterion specifying that the position of the keyword should be at the beginning of the sentence (first 20 % of the sentence) frequently promoted bad examples as they were not informative enough or lacked the necessary context. In the Slovene2 configuration the condition was removed, but this did not considerably improve the ratio of good vs. bad examples. It was observed that in good examples, the keyword almost always occurred between the middle and the end of the sentence. Once the new span for keyword position was implemented, the number of good examples per collocate improved, in certain cases significantly. However, with certain verbs, the preferred keyword position towards the end of the sentence actually proved to be even more problematic as the verb lacked the context necessary for understanding its meaning.

	Slovene1	**Slovene1b**	**Slovene2**	**Slovene3**	**Slovene3b**
sentence length	min 8 max 30	min 8 max 30	min 8 max 30	min 15 max 35	min 15 max 35
keyword position	0–20% of the sentence	0–20% of the sentence	not used	40–100% of the sentence	40–100% of the sentence
penalty for keyword repetition	NO	NO	NO	YES	YES
maximum word length is 18 characters	NO	NO	YES	YES	YES
maximum sentence length is 60 tokens	NO	NO	YES	YES	YES

Table 1: Key differences between the GDEX configurations

Several comments made during the evaluation referred to the fact that examples that contained more than one occurrence of the keyword were almost never good examples, as they were too difficult to understand and/or lacked the necessary context to be a good attestation of the meaning, collocate, construction or pattern. As a result, the classifier that penalized all the repetitions of the keyword in the same sentence was added (to configurations Slovene3 and Slovene3b).

Two classifiers that banned sentences longer than 60 words, and/or containing words longer than 18 characters were added based on WEKA analysis of the examples selected during evaluation. The impact of the two classifiers was difficult to observe during the evaluation, however as far as the classifier for word length was concerned, the subsequent examination of the wordlist of lemmas from the 1,13-billion-word Gigafida corpus, an upgrade of the FidaPLUS corpus, revealed that very few lemmas consisting of 18 characters or more were actual words – the vast majority of lemmas were websites, parts of websites, email addresses, errors that occurred during the conversion of various file formats into txt (e.g. several words joined into one word), etc., i.e. mainly items that were also penalized by other classifiers.

Weights of certain classifiers were also subject to experimentation (see Table 2), however the evaluation showed that the original weight setting produced the best results. For example, when the weight of classifiers penalizing proper nouns and pronouns respectively was lowered to 1, the evaluators observed more cases of bad examples containing proper nouns and, to a lesser extent, pronouns. Similarly, changing the weight of relative keyword position in the sentence configuration did not have any significant effect on results – in the case of this classifier, the parameters were much more relevant for identifying good examples.

	Slovene1	**Slovene1b**	**Slovene2**	**Slovene3**	**Slovene3b**
keyword position	1	1	/	1	2
penalty for proper nouns	2	1	1	2	2
penalty for pronouns	2	1	1	2	2

Table 2: The classifiers and configurations where changes in weights were made

It is also noteworthy that the evaluation proved the usefulness of other classifiers, such as the classifier allowing only whole sentences, and classifiers penalizing for regular expressions, urls, email addresses, etc. Examples that were not whole sentences were found only at collocates with low frequency that lacked better examples, i.e. most examples were not whole sentences. Similarly, regular expressions were rarely encountered in the provided examples, while urls and email addresses never appeared in the examples, which was, at least at configurations Slovene2, Slovene3, and Slovene3b, also related to the introduction of the maximum word length limit.

In the end, Slovene3 was selected among all the GDEX configurations since it produced the best results for different types of lemmas (nouns, verbs, adjectives, adverbs). The configuration was implemented in the Sketch Engine and used by the lexicographers working on the lexical database for Slovene. Several lexicographers soon reported a significant improvement in the helpfulness of GDEX when searching for good examples.

3.2.3 Remaining issues

Some evaluation findings and observations could not be fully addressed during this particular development of the GDEX for Slovene, and we list them here as they indicate which direction the further development of the configurations for Slovene might take. Moreover, these findings may be useful for the developers of GDEXes for other languages.

One common feature of bad examples was the occurrence of the sentence initial adverb (e.g. *nato*, *tako, torej, potem, poleg tega, zaradi tega, zato ker*) [2] that linked the examples with the preceding sentence. This often meant that the example did not contain enough context to be understandable. The planned solution is to devise a blacklist of sentence-initial adverbs and any other words that feature in bad examples, based on the frequency list of sentence-initial words from the corpus.

Another issue was that examples ending with something other than full stop, question mark or exclamation mark, the punctuation marks allowed by the whole sentence classifier, were still offered in the results. The sentence-ending punctuation mark that proved problematic was ellipsis; sentences ending in ellipsis were often among the top 10 offered by GDEX (all the configurations), if not even at the very top. This was not caused by the lack of good example candidates or errors in tokenisation (i.e. ellipsis treated as three full stops).

[2] The English translations of the sample adverbs are *after that, so, then, in addition, because of that, because*.

Further fine-tuning of the whole sentence classifier will be required, based on the analysis of sentence-ending punctuation in bad examples.

As mentioned in 3.2.2, the parameters of certain classifiers (e.g. relative keyword position) did not work well for all types of lemmas, and the same was true for different configurations. For example, Slovene3 and Slovene3b produced much better results for nouns and adjectives than Slovene1, Slovene1b, and Slovene2. For verbs, the differences between the results of different configurations were much smaller, however sometimes Slovene1 and Slovene1b produced better results than the other settings. Furthermore, the differences between the effectiveness of different configurations were also observed at the level of grammatical relations in word sketches. For example, for some nouns, the average number of good examples in the grammatical relations containing preposition collocates was sometimes considerably lower than in grammatical relations containing lexical words. These findings reveal the disadvantage of using a single GDEX configuration for different types of lemmas, and suggest that perhaps GDEX configurations should be tailored more narrowly, e.g. to a particular word class or even a category of lemmas within a word class.

4. Conclusion

GDEX is a very helpful tool for any lexicographer, considering how many examples need to be examined and selected during dictionary compilation. By ranking examples according to their potential to be good example candidates, GDEX acts as a sort of a sieve, pushing bad example candidates towards the bottom of the list and thus making it less likely that the lexicographers will waste time inspecting them during selection. On the other hand, by pushing good example candidates towards the top of the list, GDEX saves lexicographers' time by making it more likely that good examples will be found quickly.

The experience with English GDEX has shown which characteristics of examples are important for automatically determining whether an example is good or bad. As the experience in designing the GDEX for Slovene has shown, certain classifiers are not language-specific, such as the classifier banning sentences that are not complete, and the classifier penalizing sentences containing urls, email addresses, and regular expressions. Also classifiers penalizing proper nouns and pronouns are easily transferred to other languages assuming the appropriate distinctions are made with the part-of-speech tagger. More language-specific appear to be classifiers such as relative keyword position, where at least for Slovene, the optimal setting seems to be almost opposite to the setting for English. Classifiers such as preferred sentence length in particular, but also the threshold of low frequency words are less language-specific and more project-specific. It is also noteworthy that despite the fact that the criteria of good examples for Slovene GDEX were somewhat different to the criteria for English GDEX, the weights attributed to classifiers did not change significantly.

One of the important contributions of this project to further development of GDEX is its methodology. The process of using a data visualisation tool (WEKA) for determining and improving the values of classifiers, based on analysis of (logged) good and bad examples, combined with manual evaluation and comparison of the examples produced by different configurations has proven effective. This methodology can be used in further development of English GDEX and Slovene GDEX, as well as in the development of GDEXes for other languages. There are still improvements to be made, for example we feel that the WEKA tool can be exploited much more extensively.

Plans for the future include improving the existing GDEX for Slovene by adding blacklists, e.g. of sentence-initial adverbs, and polysemy classifier based on synsets from the wordnet for Slovene, i.e. sloWNET (Fišer, 2009). In addition, a further analysis of good and bad examples with WEKA is foreseen, in order to identify and test new classifiers.

The next project in which GDEX for Slovene will be used, will be the automatic extraction of grammatical relations, collocates, and examples for the entries in the lexical database for Slovene. For this project, the aim will be to design a GDEX configuration where the **top** two or three examples offered are always good examples. To achieve this, we plan to devise and test configurations specific to a particular category of words, e.g. nouns, verbs, adjectives, or even specific to a particular subcategory of lemmas within a word class, e.g. monosemous nouns.

5. Acknowledgements

We would like to thank Simon Krek and Polona Gantar for helping with the design and evaluation of GDEX for Slovene, the lexicographers working on the new database for Slovene for valuable feedback on GDEX, and the Sketch Engine team for technical support.

6. References

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

Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database. MIT Press.

Fišer, D. (2009). sloWNET - slovenski semantični leksikon. In M. Stabej (ed.) Proceedings of 28th symposium Obdobja. Ljubljana: University of Ljubljana, pp. 145-149.

Gantar, P., Krek., S. (2011). Slovene Lexical Database. In D. Majchráková, R. Garabík (eds.) *Natural language Processing, Multilinguality. Sixth*

International Conference. Modra, Slovakia, 20-21 October 2011. Slovenská akadémia vied, Jazykovedný ústav L'udovíta Štúra, pp. 72-80.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), pp. 10-18.

Husak, M. (2008). Automatic Retrieval of Good Dictionary Examples, Bachelor thesis. Masaryk University, Brno, Czech Republic.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress.* Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425-432.

Kilgarriff, A., Kovar, V., Rychlý, P. (2009). Tickbox Lexicography. In S. Granger, M. Paquot (eds.*)* *eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 411-418.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress*. Lorient: Universite de Bretagne-Sud, pp. 105-116.

Krishnamurthy, R. (1987). The Process of Compilation. In J. Sinclair (ed.) *Looking up: An Account of the COBUILD Project in Lexical Computing.* London and Glasgow: Collins ELT, pp. 62-85.

Landau, S. (2001). *Dictionaries: the Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Leech, G. (1992). 100 million words of English: the British National Corpus (BNC). *Language Research* 28(1), pp. 1-13.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady,* 10, 707–10.

The Communication in Slovene project. Accessed at: http://www.slovenscina.eu.

eLexicography, collaboration, bilingual corpora - all in one!

Vera Kuzmina
ABBYY
127273, Otradnaya 2B, Moscow, Russia
E-mail: vera_k@abbyy.com

Abstract

While working on a dictionary or glossary, a lexicographer faces the problem of choosing the right resource and convenient tools for finding usage examples, checking the meaning of words, and their translation into other languages. According to Atkins, B.T.S. & Rundell, M., it is a standard for the lexicographer today to use different types of corpora and language data material for these purposes. Each of them is stored in different databases with several access points, sometimes the query language is very sophisticated and lexicographers need additional training to start using these sources. In my paper I would like to draw your attention to the online dictionary and corpora website called ABBYY Lingvo.pro. The idea behind the website was to provide a single access point for the professional user and to meet his multiple needs, including lookup in various dictionary sources, searches in bilingual and monolingual corpora, collaboration with dictionary users, easy search, and other nice features for dictionary makers.

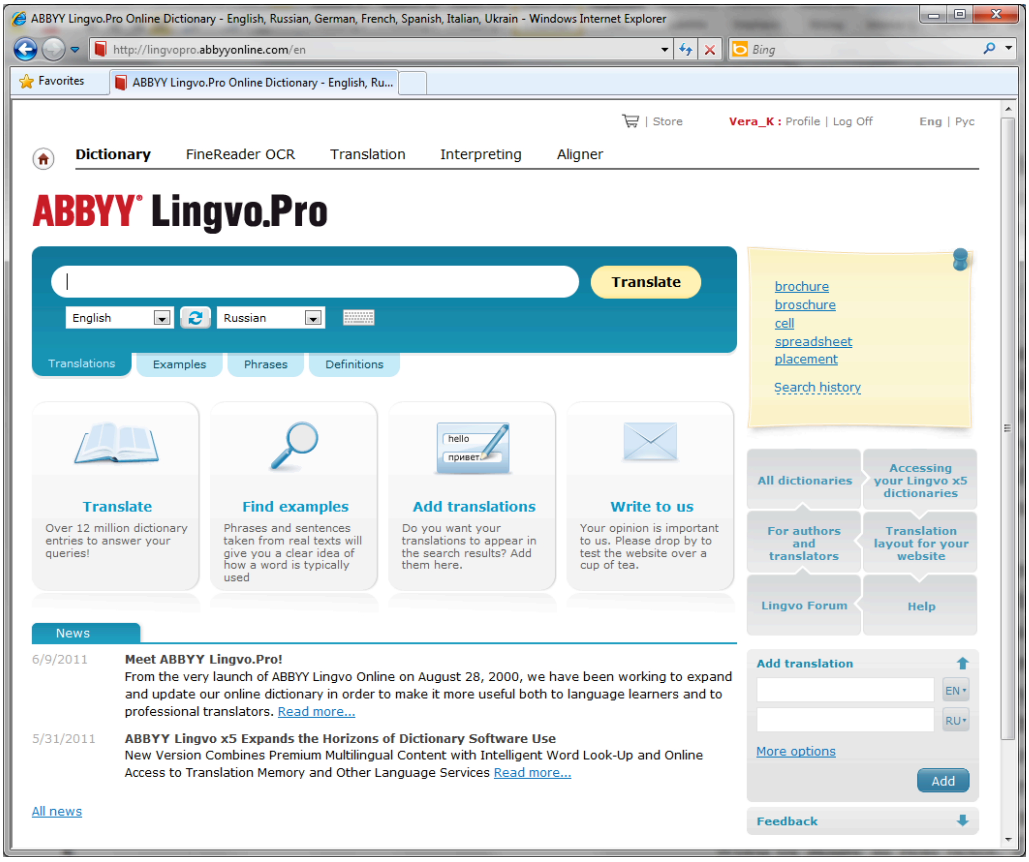
Keywords: bilingual corpora; single access point for content; term extraction tool; collaborative lexicography

Software demonstration

While working on a dictionary or glossary, a lexicographer faces the problem of choosing the right resource and convenient tools for finding usage examples, checking the meaning of words, and their translation into other languages. According to Atkins, B.T.S. & Rundell, M., it is a standard for the lexicographer today to use different types of corpora and language data material for these purposes. Each of them is stored in different databases with several access points, sometimes the query language is very sophisticated and lexicographers need additional training to start using these sources. In my paper I would like to draw your attention to the online dictionary and corpora website called ABBYY Lingvo.pro. The idea behind the website was to provide a single access point for the professional user and to meet his multiple needs, including lookup in various dictionary sources, searches in bilingual and monolingual corpora, collaboration with dictionary users, easy search, and other nice features for dictionary makers.

In general, Lingvo.pro is an online dictionary (see figure 1) which aggregates different types of content and convenient lookup and search tools, including high-quality morphology support. On the other hand, it also accumulates a large amount of bilingual parallel texts for different language pairs, for instance English, Russian, German, French, with usage examples showing words and phrases as they occur in real-life texts. This means that the user will automatically see appropriate bilingual parallel examples for each sense of the word and its translation. For instance, when looking up the word “table” in the sense “spreadsheet”, the corpus will provide sentence examples and translations only for this sense (e.g. “Tabelle” in German, see figure 2).

The above usage scenario helps lexicographers avoid multiple steps, starting from the lookup of the word in different dictionaries and making queries to a corpus. In Lingvo.pro, all these steps are merged into one quick lookup. The lexicographer may extract terms with the help of an online Terms Extraction tool and check the terms and their translations directly in this list. This list of terms can be used as a draft for the new dictionary or as an additional source to check possible translations while working on word senses. This simplicity is the main advantage for both professional lexicographers and those who only start their careers in dictionary creation. At the same time, there are also much more sophisticated tools available on Lingvo.pro.


ABBYY Lingvo.Pro Online Dictionary - English, Russian, German, French, Spanish, Italian, Ukrain - Windows Internet Explorer
http://lingvopro.abbyyonline.com/en
Favorites
ABBYY Lingvo.Pro Online Dictionary - English, Ru...
Store
Vera_K : Profile | Log Off
Eng | Рус
Dictionary
FineReader OCR
Translation
Interpreting
Aligner
ABBYY Lingvo.Pro
Translate
English
Russian
Translations
Examples
Phrases
Definitions
brochure
broschure
cell
spreadsheet
placement
Search history
hello
привет
Translate
Over 12 million dictionary entries to answer your queries!
Find examples
Phrases and sentences taken from real texts will give you a clear idea of how a word is typically used
Add translations
Do you want your translations to appear in the search results? Add them here.
Write to us
Your opinion is important to us. Please drop by to test the website over a cup of tea.
All dictionaries
Accessing your Lingvo x5 dictionaries
For authors and translators
Translation layout for your website
Lingvo Forum
Help
News
Add translation
EN
RU
More options
Add
Feedback
All news

6/9/2011 **Meet ABBYY Lingvo.Pro!**
From the very launch of ABBYY Lingvo Online on August 28, 2000, we have been working to expand and update our online dictionary in order to make it more useful both to language learners and to professional translators. Read more...

5/31/2011 **ABBYY Lingvo x5 Expands the Horizons of Dictionary Software Use**
New Version Combines Premium Multilingual Content with Intelligent Word Look-Up and Online Access to Translation Memory and Other Language Services Read more...

Figure 1. ABBYY Lingvo.pro General view

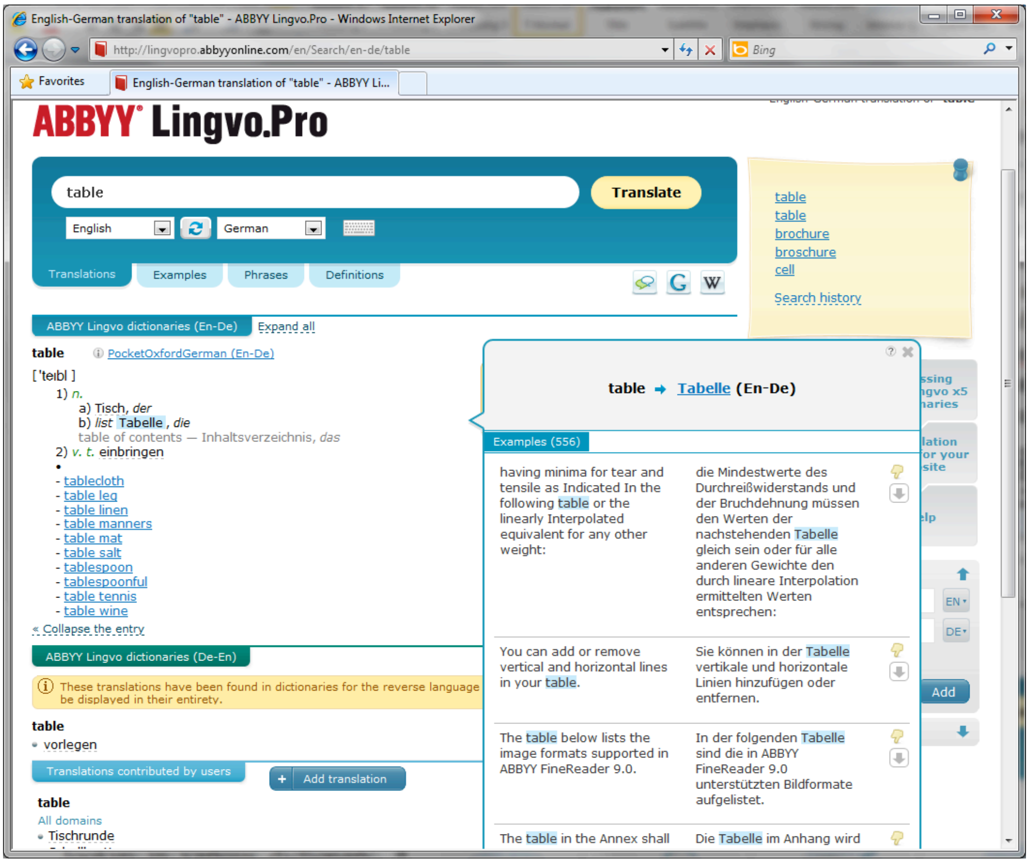

English-German translation of "table" - ABBYY Lingvo.Pro - Windows Internet Explorer
http://lingvopro.abbyyonline.com/en/Search/en-de/table
Favorites
English-German translation of "table" - ABBYY Li...
ABBYY Lingvo.Pro
table
Translate
English
German
Translations
Examples
Phrases
Definitions
table
table
brochure
broschure
cell
Search history
ABBYY Lingvo dictionaries (En-De)
Expand all
EN
DE
Add

table PocketOxfordGerman (En-De)
['teɪbl]
1) *n.*
a) Tisch, *der*
b) *list* Tabelle , *die*
table of contents — Inhaltsverzeichnis, *das*
2) *v. t.* einbringen
- tablecloth
- table leg
- table linen
- table manners
- table mat
- table salt
- tablespoon
- tablespoonful
- table tennis
- table wine

« Collapse the entry

ABBYY Lingvo dictionaries (De-En)

These translations have been found in dictionaries for the reverse language be displayed in their entirety.

table
- vorlegen

Translations contributed by users
+ Add translation

table
All domains
- Tischrunde

table → Tabelle (En-De)

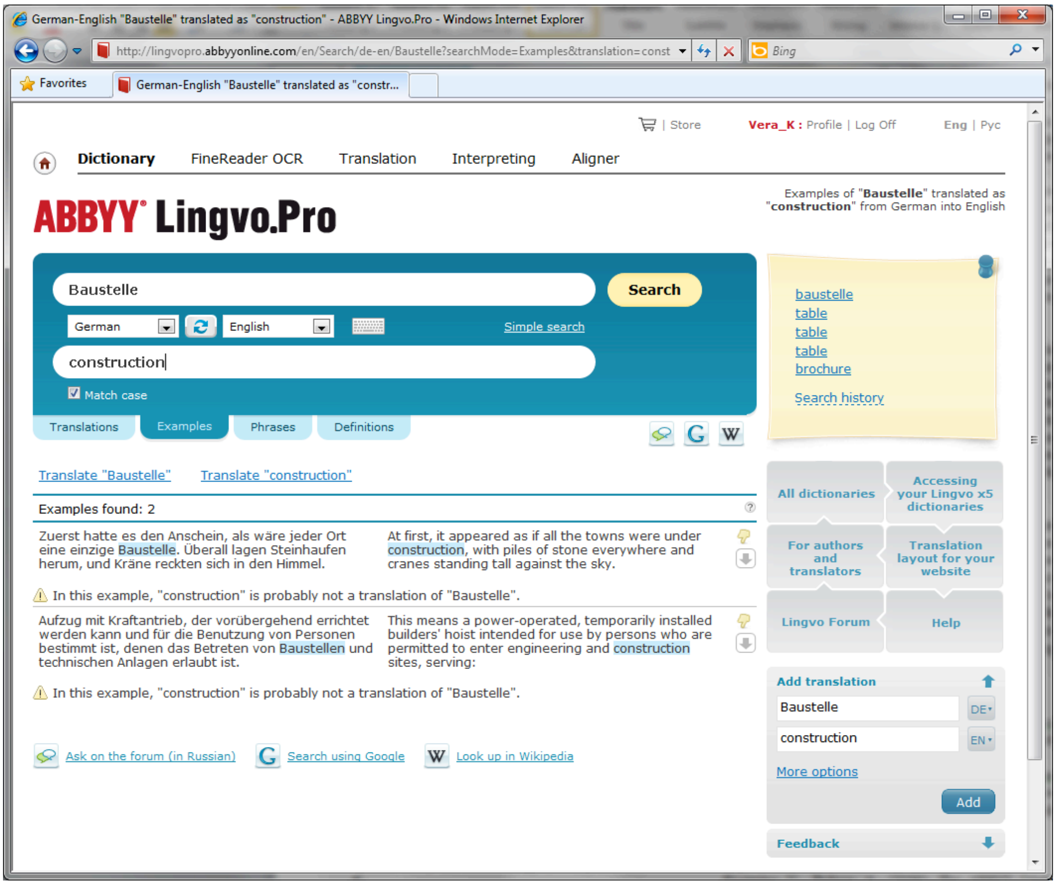
Examples (556)

having minima for tear and tensile as Indicated In the following table or the linearly Interpolated equivalent for any other weight:	die Mindestwerte des Durchreißwiderstands und der Bruchdehnung müssen den Werten der nachstehenden Tabelle gleich sein oder für alle anderen Gewichte den durch lineare Interpolation ermittelten Werten entsprechen:
You can add or remove vertical and horizontal lines in your table.	Sie können in der Tabelle vertikale und horizontale Linien hinzufügen oder entfernen.
The table below lists the image formats supported in ABBYY FineReader 9.0.	In der folgenden Tabelle sind die in ABBYY FineReader 9.0 unterstützten Bildformate aufgelistet.
The table in the Annex shall	Die Tabelle im Anhang wird

Figure 2. ABBYY Lingvo.pro Translation example

One of the useful tools is the advanced search tool, a professional, yet easy to use, corpus query system. Firstly, there is a bilingual search feature (see figure 3) with the possibility to find a word in one language and its translations into another language. Users can make queries for words in different senses and search for words and phrases in different positions. Then, there is a “do not” option, a match word order option, ability to find inflected forms, case-sensitive searches, and many other things which may make the life of dictionary maker easier. The user interface of the website and the corpus query system within are user-friendly and easy to use.

All the text corpora accessible on Lingvo.pro are morphologically and syntactically tagged so that even disambiguation becomes possible. For dictionary makers, such precise and sophisticated approach to word senses means that their everyday language investigations become simpler and the results achieved are significantly better. For example, an author may list the meanings of the word according to their usage frequency. With Lingvo.pro, a lexicographer may see the frequency of a word or sense in real texts. This is a very important and unique feature of bilingual text collections. Frequency is automatically shown next to each retrieved word sense. This makes it really easy for dictionary makers to investigate usage frequencies, which are now only a couple of mouse clicks away.



German-English "Baustelle" translated as "construction" - ABBYY Lingvo.Pro - Windows Internet Explorer

http://lingvopro.abbyyonline.com/en/Search/de-en/Baustelle?searchMode=Examples&translation=const

Examples of "**Baustelle**" translated as "**construction**" from German into English

Baustelle

German → English

construction

☑ Match case

Examples found: 2

Zuerst hatte es den Anschein, als wäre jeder Ort eine einzige Baustelle. Überall lagen Steinhaufen herum, und Kräne reckten sich in den Himmel.	At first, it appeared as if all the towns were under construction, with piles of stone everywhere and cranes standing tall against the sky.

⚠ In this example, "construction" is probably not a translation of "Baustelle".

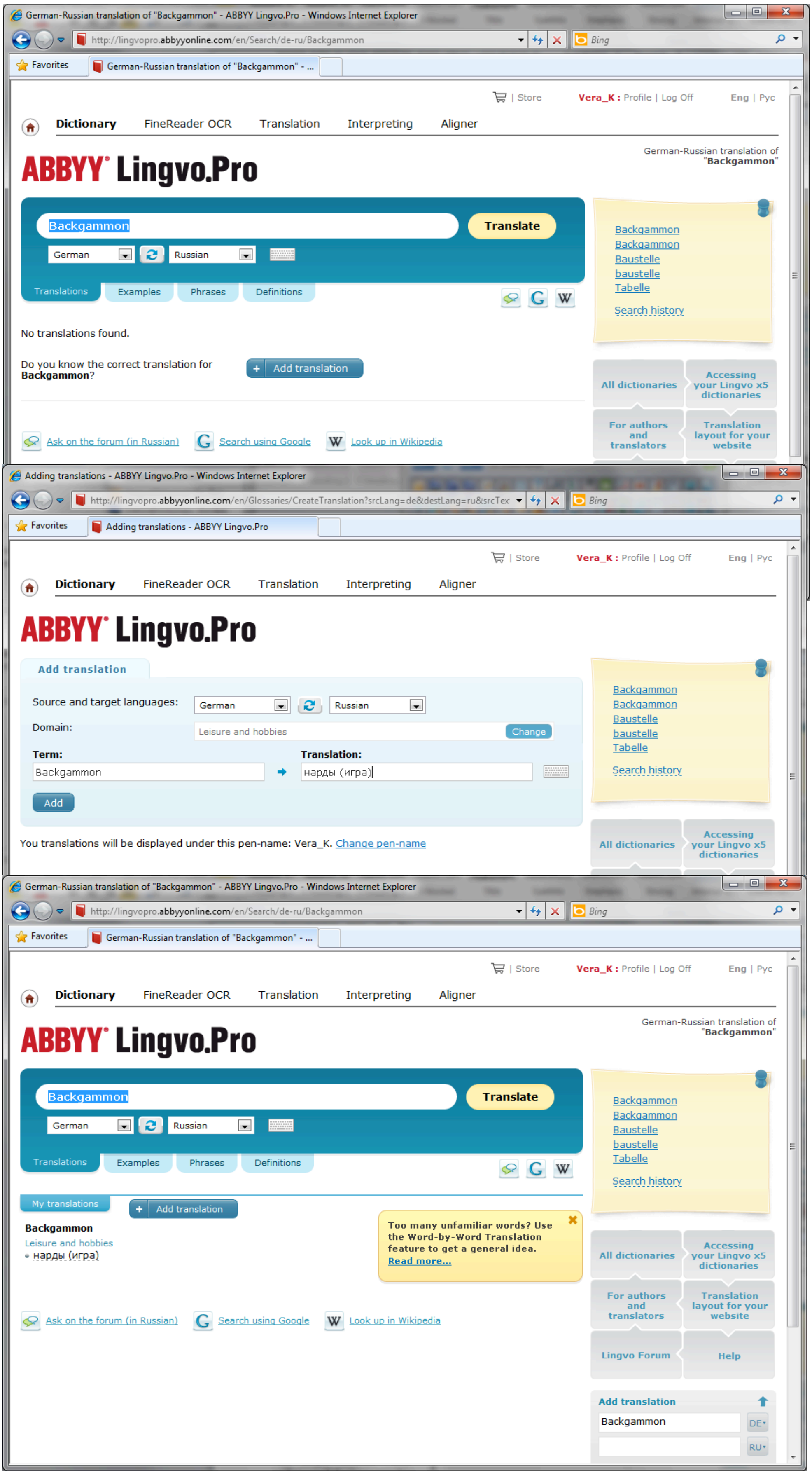
Aufzug mit Kraftantrieb, der vorübergehend errichtet werden kann und für die Benutzung von Personen bestimmt ist, denen das Betreten von Baustellen und technischen Anlagen erlaubt ist.	This means a power-operated, temporarily installed builders' hoist intended for use by persons who are permitted to enter engineering and construction sites, serving:

⚠ In this example, "construction" is probably not a translation of "Baustelle".

Add translation: Baustelle DE → construction EN

Figure 3. ABBYY Lingvo.pro Bilingual search

Another important feature is the ability to create glossaries and add new words into glossaries and dictionaries available on the website (see figure 4). Members of the community can work together, adding new words, senses, and translations. Convenient collaboration tools are available: users can propose brand-new words to be included into the dictionary and discuss the new proposals with other users or with the editors and lexicographers involved in the project.


German-Russian translation of "Backgammon" - ABBYY Lingvo.Pro - Windows Internet Explorer
http://lingvopro.abbyyonline.com/en/Search/de-ru/Backgammon
Favorites
Store
Vera_K : Profile | Log Off
Eng | Рус
Dictionary
FineReader OCR
Translation
Interpreting
Aligner
ABBYY Lingvo.Pro
German-Russian translation of "Backgammon"
Translate
Translations
Examples
Phrases
Definitions
Backgammon
Backgammon
Baustelle
baustelle
Tabelle
Search history
All dictionaries
Accessing your Lingvo x5 dictionaries
For authors and translators
Translation layout for your website
Ask on the forum (in Russian)
Search using Google
Look up in Wikipedia
Add translation
Adding translations - ABBYY Lingvo.Pro - Windows Internet Explorer
http://lingvopro.abbyyonline.com/en/Glossaries/CreateTranslation?srcLang=de&destLang=ru&srcTex
Change
Add
My translations
Lingvo Forum
Help
DE
RU

Backgammon

German → Russian

No translations found.

Do you know the correct translation for **Backgammon**?

Source and target languages: German → Russian

Domain: Leisure and hobbies

Term: Backgammon

Translation: нарды (игра)

You translations will be displayed under this pen-name: Vera_K. Change pen-name

Backgammon
Leisure and hobbies
- нарды (игра)

Too many unfamiliar words? Use the Word-by-Word Translation feature to get a general idea. **Read more...**

Add translation: Backgammon

Figure 4. ABBYY Lingvo.pro Adding new terms

In conclusion, I would like to point out that dictionary creators today are sometimes overloaded with different sources and software tools which enable access to these sources. There is demand for one simple solution which will meet all the lexicographic needs at once and will be easily accessible from any computer even outside the office, because most dictionary creators are teleworkers. There is also a growing trend to make complicated things simpler, which is exactly the purpose of Lingvo.pro.

References

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

Hockey, S.M. (2000). Dictionaries and lexical databases. *In Electronic Texts in the Humanities: Principles and Practice.* Oxford & New York: Oxford University Press, pp. 146-171.

http://www.abbyy.com/

http://www.lingvo.com/

Kuzmina V., Rylova A. (2009) *The ABBYY Lingvo electronic dictionary and the ABBYY Lingvo Content dictionary writing system as lexicographic tools.* In S. Granger, M. Paquot (eds*.) eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Not the Word I Wanted?
How Online English Learners' Dictionaries Deal with Misspelled Words

Robert Lew
Adam Mickiewicz University
E-mail: rlew@amu.edu.pl

Roger Mitton
Birkbeck College, University of London
E-mail: roger@dcs.bbk.ac.uk

Abstract

This study looks at how well the leading monolingual English learners' dictionaries in their online versions cope with misspelled words as search terms. Six such dictionaries are tested on a corpus of misspellings produced by Polish, Japanese, and Finnish learners of English. The performance of the dictionaries varies widely, but is in general poor. For a large proportion of cases, dictionaries fail to supply the intended word, and when they do, they do not place it at the top of the list of suggested alternatives. We attempt to identify some of the mechanisms behind the failures and make further suggestions that might improve the success rate of dictionary interfaces when identifying and correcting misspellings. To see whether it is possible to do better than the dictionaries tested, we compare the success rates of the dictionaries with that of an experimental context-free spellchecker developed by the second author, and find the latter to be markedly superior.

Keywords: online dictionaries; English; spelling; spelling correction; spelling errors; misspelling; access

1. The role of spelling in dictionary consultation

One painful limitation of traditional paper dictionaries is that the primary access route — at least for the most popular semasiological (form-to-meaning) reference works and for languages with alphabetic writing systems — requires that the user (1) is familiar with the access alphabet (Nielsen, 1995) of the dictionary, and (2) knows how the target item is spelled. With reference to the first point, users of modern electronic dictionaries are indeed (if only up to a point) 'liberated from the straitjacket of ... alphabetical order' (Atkins, 1996: 516), thus making alphabetical ordering less of a critical factor in the success of the access process. However, point (2) remains a valid concern: the dictionary engine still needs to match the search term entered by the user against the available list of keywords covered in the dictionary, which include, but need not be limited to, the headwords.

Of course, dictionary users cannot always be expected to replicate standard English spelling. Further, cases of misspelling can result from a mechanical typo (performance errors) or erroneous lexical-graphemic representation (competence errors). In the second case in particular, misspelling patterns (Mitton & Okada, 2007) typical of native speakers of English may be different from those of learners of English.

Further, online dictionaries are increasingly used in conjunction with online work and entertainment. This includes the need for lexicographic assistance in the context of listening, such as when learners of English attempt to look up a word which they hear being spoken while watching a TV show on their computer. Such a lookup situation is bound to generate queries where the search term, rather than representing a specific vocabulary item from the learner's lexical repertoire, is a 'creative spelling', a shot-in-the-dark: a transcription of what the user imagines he has heard. This is a little similar to what some call *phonetic spelling* (cf. Proctor, 2002), but more complex, as here not one but at least two phonological systems are involved, with their own phonotactic regularities and spelling-to-sound correspondences. We would expect the best electronic dictionaries to be able to offer useful assistance in all of the above cases; but do they actually provide such assistance?

2. Spelling correction in e-dictionaries

No matter how rich and sophisticated the lexicographic content a dictionary, it will be completely lost on the users if they do not succeed in finding their way to the appropriate dictionary entry.

In a common type of e-dictionary interface where lookup consists in the user typing in a search term into a search box, the string entered needs to be matched against a list of keywords held in the dictionary itself, to see if it corresponds to a lemma present in the dictionary, or possibly (in more sophisticated dictionaries) is part of a multi-word unit treated under a different lemma. An exact-match algorithm would assume that dictionary users are perfect spellers, which is obviously not a realistic assumption. Ideally, a good dictionary interface should be able to guess the user's intention even if they misrepresent the orthographic form of the word. However, in a recent analysis of three online German dictionaries (Bank, 2010), only one dictionary has been found to be at all 'rechtschreibtolerant' — that is, able to deal with misspellings in any useful fashion.

A good dictionary interface — when presented with an unknown string — should make reasonable guesses as to the possible alternative forms the user may have meant. Furthermore, if the guesses are presented as an ordered list, then the best guesses should be close to the top of the list. In an ideal case, the one word actually intended by the user should be presented at the very top of the list of suggestions, but this ideal is not always achievable — even in the best possible system — due to the inherently erratic nature of misspellings. The demands on the dictionary are here similar to those on a state-of-the-art spellchecking system in a word processor, though not identical.

First, the dictionary needs to recognize that the search term entered is not a standard spelling. Then, it needs to home in on a compact set of the most likely alternatives and rank them, so that they can be presented back to the user as an ordered list. Or, less commonly, it might just take the user to the entry for the top-ranking alternative (much as the Google search engine currently does). In broad outline, the procedure is similar to that involved in checking texts; however, there are differences, such as the opportunity to use context to refine the list (typically absent in dictionary lookups), or the need to handle proper nouns.

2.1 Types of spelling errors

Many of the spelling errors in running text are single-error departures from the target word. Taking the target word *trepidation* as an example, these are usually understood as being one of the following four subcategories: a single letter is omitted in a word (*tepidation*); a single letter is wrong (*trepitation*); one extra letter is inserted (*treppidation*); two adjacent letters are transposed (*trepidaiton*). According to some studies (Damerau, 1964; Pollock & Zamora, 1984), such simple errors may account for over eighty percent of misspellings. However, this percentage is likely to be lower with a more realistic representation of poor spellers in the corpus: sixty-nine percent in Mitton (1996: 46). Many (though not all) of these simple errors tend to be the result of mistyping words. As such, they are mechanical errors of performance, rather than errors of competence, and some authors use the term *misspelling* in a narrower sense which excludes mistypings (e.g. Deorowicz & Ciura, 2005). Though it is not always possible to categorize an error as one type or the other (e.g. **accomodation* for *accommodation*, or **consistant* for *consistent*), their underlying causes are different. It is misspellings of the competence type that are our primary focus here.

At the other end of the mechanical-conceptual cline, there are non-standard formations at the lexical-morphological level, such as when a speaker actually has the word **unpolite* in their mental lexicon and uses it in place of (or as a variant of) the standard *impolite*. Though sometimes the source of genuine problems, especially for non-native users of a language, it is doubtful if such errors of lexical competence should be classified as strictly *spelling-related* (pace Deorowicz & Ciura, 2005).

2.2 Rare versus common words

Low-frequency words are, by definition, words that are used infrequently in running text. Therefore, it is reasonable to assume in a spelling correction system that an instance of a rare word (especially a very infrequent one) may be a misspelling if there is a common word to which it bears some similarity. For example, as pointed out by Mitton (1996: 96), the orthographic string *wether* when found in a running text is more likely to be a misspelling of either *whether* or *weather* than the rare word meaning ‘a castrated ram’. Spelling correctors working with text can use this information to detect and flag such potential real-word errors. However, in a corpus of strings being looked up in an online dictionary, the frequency distribution of word forms is less skewed than in running text (De Schryver et al., 2006), so that even quite rare words have a fair chance of being looked up. This makes perfect sense: when someone reads a text, they will not usually be troubled by all the familiar common words, but the occasional rare word is likely to be looked up. So, although De Schryver et al.’s study of log files presents only a single piece of evidence, it is reasonable to assume that native speakers, and to a lesser extent advanced learners, often consult their dictionaries for less frequent words.

2.3 The role of context

Most work in spellchecking and spelling correction so far has been done with reference to forms embedded in textual context, and some of the more advanced systems attempt to utilize contextual information to improve the accuracy of guessing at the form intended. However, when online dictionaries are consulted, it is at present most usually by typing a word into the search window. In such a setup, no contextual information would normally be available to the dictionary application. Still, most spellcheckers designed for the correction of texts do not use context either, and yet achieve good success rates nevertheless (Kukich, 1992; Deorowicz & Ciura, 2005; Mitton, 2009).

3. The study

3.1 Aim

The aim of the study is to assess the performance of the leading monolingual learners’ dictionaries of English in their online versions at guessing the intended headword when presented with their misspelled versions produced by foreign learners. By *performance* we here mean the particular ability to recover the intended lemma and suggest it back to the user as a plausible alternative to what the user has actually typed in the search box. Ideally, the intended word should be offered as the only suggestion, but usually several alternatives will remain plausible, so dictionaries will customarily provide not just one suggestion but a short list. In such a case, the nearer the top of the list the intended headword appears, the better the performance of the spelling correction mechanism.

More specifically, we would like to find out whether the level of performance of the most prestigious dictionaries is in general satisfactory, to what extent the different dictionaries perform similarly or differently, and how specific dictionaries compare with the others.
Our corpus (see 3.2 below) includes misspellings by learners of varying linguistic backgrounds (Polish, Japanese, and Finnish), and it might also be interesting to see if some dictionaries are perhaps better equipped to cope with misspellings typical of learners speaking a given native language.

In view of the preliminary results indicating that the tested dictionaries performed below expectation, a further aim was added during the course of the study, and for this, the original author was joined by the second author. This further aim was to see if an experimental context-free spelling corrector designed by the second author (Mitton, 1996) would be able to perform better than the dictionaries tested.

3.2 Corpora of misspellings used in the study

The corpus of spelling errors used in the present study is made up of 200 misspellings broken down into three subcorpora, each representing attempts at spelling English words by native speakers of three different languages that are typologically very distant, as they all represent different language families: Polish (100 items), Japanese (50), and Finnish (50). A brief description of the three sets of misspellings follows, and a sample of ten items from each is given in the Appendix.

3.2.1 Polish misspellings

The most substantial part of the corpus of misspellings used in this study came from a Polish subcorpus, collected in 2010 by the first author, with the help of two student assistants as experimenters.

The data were collected by way of oral elicitation. A set of English words known to be frequently misspelled was taken from *The 200 Most Commonly Misspelled Words in English*[1] reported by Richard Nordquist, and these were used as elicitation triggers (target words). One by one, the words from the list were played back in audio form to one of two Polish learners of English at first year of college (one female from Szczecin University, one male from Gdańsk University), using the built-in audio pronunciation capability of the popular bilingual English-Polish dictionary Diki.pl, known for its decent audio quality. Thus, a target word would be played back to the participant without disclosing its orthography, and the participant would respond by typing the word into the computer. The experimenter would wait until the participant indicated that they were done, and then proceed to play back the next target word. Participants had been instructed in the warm-up sessions to proceed as if they were looking up words just heard in an online dictionary. All the typed wordlike strings were logged. Correctly spelled words as well as obvious mistypings, which in all likelihood would not have challenged the spellchecking algorithms of the dictionaries, were subsequently removed, with the remaining strings yielding the Polish subcorpus of 100 misspellings. This elicitation technique is believed to mimic dictionary lookup behaviour for stimuli perceived aurally (i.e., while listening).

3.2.2 Japanese misspellings

The 50 Japanese misspellings were taken from the SAMANTHA Error Corpus created by Takeshi Okada at Tohoku University, Japan. In order to collect the misspellings, Japanese students had been asked to write down an English word based on its definition in Japanese and an approximate representation of English pronunciation in the Japanese moraic (or, more loosely, syllabic) script katakana. For this study, the most common misspelling was selected from the corpus which was not a single-error type (and thus not challenging enough for spellcheckers). Up to a point, though perhaps not as much as for the Polish sample, the elicitation technique used would be likely to produce misspellings influenced by Japanese orthotactic and phonotactic rules (i.e., the typical sequencing of letters and sounds, respectively), as well as the native spelling-to-sound correspondences.

3.2.3 Finnish misspellings

The set of Finnish misspellings was obtained from the Birkbeck spelling error corpus (Mitton, 1985) via the Oxford Text Archive. The Finnish data themselves were collected by Suomi (1984) as part of her MA research. Her corpus also included data from native speakers of Swedish, but for this study, only the data from native speakers of Finnish were used. We also discarded most obvious mistypings, as for the Polish corpus. This resulted in a list of 50 misspellings.

3.3 Dictionaries tested

Each of the misspelled words in the corpus was looked up manually in each of the following seven dictionaries, all except the Google Dictionary being dictionaries for advanced learners of English, and all but one freely available online. The seven dictionaries tested were (URL's are given in the **References** section):

1. *Longman Dictionary of Contemporary English*, free online version (henceforth, LDOCE Free);
2. *Longman Dictionary of Contemporary English*, premium subscription version (LDOCE Premium);
3. *Merriam-Webster's English Learner's Online Dictionary* (MWALED);
4. *Macmillan English Dictionary Online* (MEDO);
5. *Cambridge Advanced Learner's Dictionary* (CALD);
6. *Oxford Advanced Learner's Dictionary* (ALD), and
7. *Google English Dictionary* (GoogleED).

The general idea was to test English monolingual

[1] http://grammar.about.com/od/words/a/misspelled200.htm

dictionaries for learners of English available freely online. The set of leading English monolingual learners' dictionaries is actually well defined, and is frequently referred in the lexicographic literature as the Big Five, and includes: ALD, LDOCE, COBUILD, CALD, and MEDO. Of these, COBUILD has not been tested as it does not currently offer a free online version. For LDOCE, two versions were tested: the free online version, and also a Premium version. This version is available by subscription, with time-limited access granted to buyers of paper and DVD-Rom copies. It was included in order to see if paying users were being served better than users of the free version (in fact, quite the reverse turned out to be the case, as we shall see below).

In addition to these four British learners' dictionaries, we also included MWALED. Even though in terms of lexicographic content this American-made learner's dictionary may still not compare very favourably with the Big Five (Hanks, 2009; Bogaards, 2010), its web interface does offer some commendable features (Lew, 2011).

Finally, GoogleED was also included in the study. GoogleED used to be a learners' dictionary of sorts, with the core lexicographic content apparently based on COBUILD. In August 2010, GoogleED switched over to the *Oxford American College Dictionary* (Lindberg, 2006), which is not a dictionary targeted at language learners, but primarily at American college students speaking English as their native tongue. However, four factors spoke in favour of including GoogleED in the sample.

First, being associated with Google, the unquestioned leader in search engines, it was reasonable to expect it to become a very significant player also as an online dictionary of English for non-English-speaking netizens.

Second, its history as an online version of COBUILD, one of the Big Five, is in itself significant, and may have attracted a number of learner users who remained regular users even after the switch.

Third, although the *Oxford American College Dictionary* is a native-speaker dictionary, it is largely based on the *New Oxford American Dictionary* (McKean, 2005), which, in turn, grew out of the *New Oxford Dictionary of English* (Hanks & Pearsall, 1998). This latter dictionary benefited from Patrick Hanks' prominent involvement with the COBUILD project, and so in many ways it is closer to the learner dictionary model than a traditional dictionary for native speakers of English.

Finally, Google has become a sort of a synonym for data search and access. We therefore wanted to challenge the experts, as it were, and see if GoogleED would perform better than the 'regular' dictionaries.

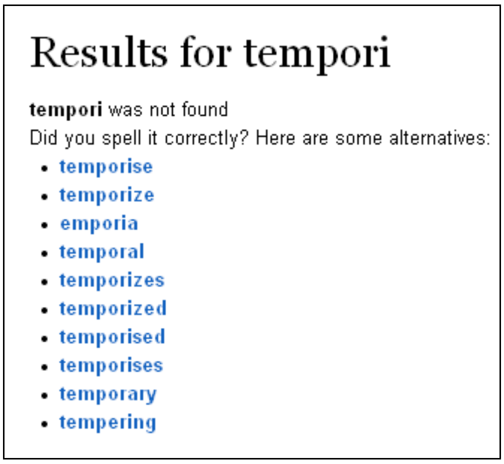
Somewhat surprisingly, GoogleED is no longer officially available online as of this writing (2 Sept 2011). Apparently, it was discontinued without warning in August 2011. However, much of the functionality can still be accessed by using the *define: term* syntax in a general Google search, and then clicking on *more* within the top item on the results list, which selects the *Dictionary* tab from the sidebar currently appearing to the left of the Google search user interface. Alternatively, the same effect can be achieved more directly by appending a parameter value of *tbs=dfn:1* to a Google search. For example, to get directly to the Google dictionary entry for the word *bay*, one would at this time use the following URL: http://www.google.com/search?q=bay&tbs=dfn:1. In some browsers (Opera, for example), it is possible to define customized search shortcuts of this type, so that lookups in the Google English Dictionary can still be performed conveniently from the address bar.

3.4 Procedure

All lookups were performed manually online by the first author, between January 16 and 19, 2011. For each misspelled word, the misspelling was pasted into the search box of each of the dictionaries. In every case, it was noted whether the dictionary was able to identify the correct target word, and, if the dictionary provided a list of alternatives, what was the position of the target word relative to other, irrelevant, hits. The word (or non-word string, as was sometimes the case) presented at the top of the suggestions list was also noted, as well as any other striking suggestions further down the list.

As an illustration of the procedure, consider

Figure 1 below, taken from a test lookup in CALD. The intended word was *temporary*, and it was misspelled as **tempori*. The dictionary returned a list of ten suggestions. The top suggestion (number 1 on the list) was *temporise*, which was not the intended word. However, the correct target word *temporary* was found further down the list: in this case it was listed ninth. So, position 9 was noted for this misspelling in CALD.



Results for tempori

tempori was not found

Did you spell it correctly? Here are some alternatives:

- temporise
- temporize
- emporia
- temporal
- temporizes
- temporized
- temporised
- temporises
- temporary
- tempering

Figure 1: Example suggestions list in CALD for

the target word *temporary* misspelled as **tempori*

This example is quite representative of six of the seven dictionaries tested; the exception was GoogleED, which did not provide a longer list of suggestions, but only a single alternative (if any).

Data for all dictionaries and misspellings were keyed into a database and analyzed so as to evaluate the relative performance of the seven dictionaries.

3.5 How well the dictionaries performed

Aggregated results for the complete corpus (i.e. Polish, Japanese, and Finnish) are presented in Table 1 and Figure 2 below. Percentage figures in the table cells indicate what proportion of the 200 target words were found in the respective positions within the individual suggestions lists returned by the dictionaries.

The figures under the heading *First* cover those cases where the target word was presented at the very top of the suggestions list. *Top 3* means that the target was listed as first, second, or third, and so on. These figures are cumulative, so if a target was listed at the top of the list, it was automatically counted under all four categories (i.e. *First*, *Top 3*, *Top 6*, and *Top 10*). Figure 2 conveys the results in a more visually appealing form.

	Target word listed in position:			
Dictionary	First	Top 3	Top 6	Top 10
LDOCE Free	51%	65%	75%	79%
LDOCE Premium	50%	59%	60%	62%
MWALED	47%	57%	63%	65%
MEDO	25%	44%	52%	55%
CALD	26%	44%	51%	55%
ALD	22%	42%	47%	52%
GoogleED	44%	(44%)	(44%)	(44%)

Table 1: Success rates for the seven dictionaries across all data. Figures indicate the proportion of target words found in the respective positions in the suggestions list.

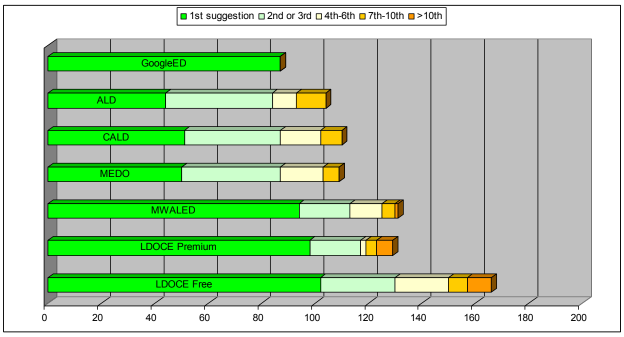


Figure 2: Performance of the seven dictionaries for all data (N=200). Colour bars indicate the number of target words ranked in the respective positions in the suggestions list.

Two things are immediately obvious in the results: the relatively wide variation between the different dictionaries, and the generally disappointing performance of most of the dictionaries tested. To get some perspective on these figures, it is worth remembering that our corpus of misspellings was designed to be challenging. Unlike some other studies, we did not focus on typos, most of which are simple errors that can be corrected with unsophisticated algorithms. Still, the very wide disparities between the success rates do indicate that at least some dictionaries are not doing the best job possible, to put it mildly.

There is a very clear gap between ALD, CALD and MEDO on the one hand and the two versions of LDOCE and MWALED on the other. The first three dictionaries only get between one-fifth and one-fourth of the target words right in the sense of placing the target at the very top of the suggestions list. In contrast, LDOCE and MWALED succeed in guessing the target word about half of the time, with LDOCE being marginally better than MWALED. GoogleED does only slightly worse than LDOCE and MWALED in this respect.

If we now lower the standard and include all suggestions in the top ten, then ALD, CALD and MEDO catch up somewhat, largely thanks to being able to include more of the target words in second or third place (pale green bars in Figure 2). But even with the top ten items on the list included, these three dictionaries only succeed in between 52% and 55% of the cases, which is comparable to the success rate of the better dictionaries for their *first* suggestion only. On the *top ten* measure, MWALED gets slightly ahead of LDOCE Premium, but it is LDOCE Free that really surges ahead, with a lot of accurate guesses in its lists found between the ranks of 2 and 6. It clearly outperforms all the other dictionaries, including — surprisingly — its sister LDOCE Premium. GoogleED has the lowest *top ten* score, but it has effectively thrown in the towel by failing to offer anything beyond the first suggestion.

3.6 Where the dictionaries failed

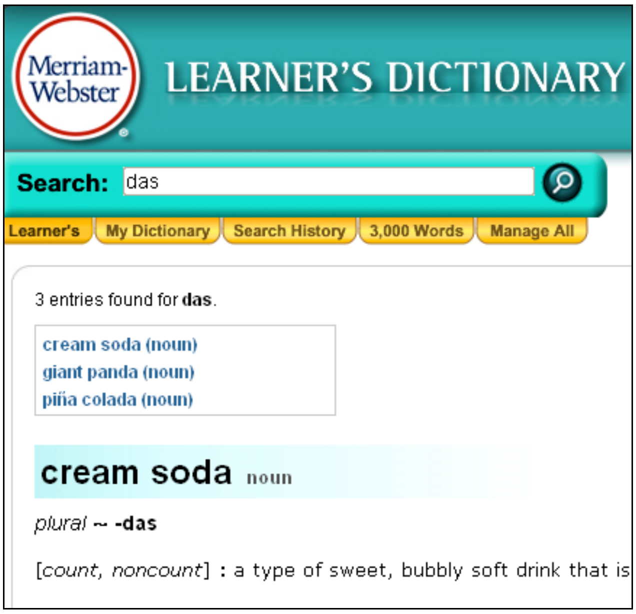
Since we have access to records of top suggestions offered by the respective dictionaries, it may be interesting to look at some of the problematic cases and offer comments as to what may have caused the less-than-optimal guesses, and how these could have been avoided.

Starting with the ALD, it seems this dictionary attaches too much weight to substring matching. This might explain why it would offer *apology* for **sakology* (a phonetically-motivated misspelling of *psychology*). Apparently, the dictionary homes in on the -*ology*, and then repeats the process with what remains, finding that *ap*- and *sak*- share the letter 'a'. The remaining items on the suggestions list are as follows: *sexology*, *sinology*, *ecology*, *zoology*, *ufology*, *urology*, *geology*, *cytology*, and *tautology*, in this order, and one wonders why *apology* was listed first. In general, ALD does not seem to give much regard to the first letter, even though research has shown that people generally get the first letter right (Yannakoudakis & Fawthrop, 1983; Mitton, 1996). For instance, it offers *masons* for **laysons* (misspelling of *license*), *newbery* for **lajbery* (*library*), and *deferens* for **referens* (*reference*).

A particular oddity of the suggestions served up by ALD, CALD, and MEDO alike is their tendency to offer words with an *–s* at the end, even though there is no indication in the misspelling that one is required. Thus, all three suggest *citizens* for **sitizen*, with the correct *citizen* only appearing in second place. Similarly, we get at the top of the list: *recommends*, *repetitions*, *disappoints*, *forwards*, and even *spaghettis* (for **spagetti*) — that despite the fact that the dictionaries mark the noun as UNCOUNTABLE, and so not usually plural. This mysterious tendency loses the three dictionaries quite a few easy points for top suggestion, at the same time inflating their *top 3* counts, as the reasonable suggestion tends to appear second in such cases. Why would all of ALD, CALD, and MEDO be affected by this overeagerness to tag on *–s*? Perhaps this has something to do with the software for dictionary compilation and publication that all three use: the DPS Writing System, developed and maintained by the company IDM. However, as far as we know, LDOCE also uses the DPS system, and yet it does not exhibit the *–s* problem.

At times, the suggestions offered by our dictionaries can be downright bewildering. A case in point are MWALED's offerings for **das*, a misspelling of *does*. Admittedly, this is indeed a challenging item, but the suggestions are puzzling, to say the least. The dictionary's output is given in Figure 3 below, and it includes three suggestions: *cream soda*, *giant panda*, and *piña colada*. Only a closer look at the entry can reveal why MWALED should come up with such a list of suggestions. As it turns out, in the comment on form section, the plural for these compounds is given in a traditional compressed form as '~-das', and apparently it is this string that the dictionary has homed in on. Obviously, such a suggestion is a complete red herring. Another surprise from MWALED, though this time with no apparent explanation, is the suggestion *archdiocese* for **ridicyles* (a misspelling of *ridiculous*).

It is difficult to see why MWALED would have a problem with the misspelling **spagetti* — probably the easiest item in the whole corpus, which all the others get perfectly right (except ALD, which only lists the intended word *spaghetti* in second place, following the pluralized *spaghettis*). MWALED offers here no less than 16 alternatives (*spigot*, *spectate*, *spotted*, *spotlight*, *speculate*, *spectacle*, *septet*, *aseptic*, *sabotage*, *septic*, *sceptical*, *sceptic*, *seepage*, *sceptically*, *slippage*, *spatula*), but the obvious *spaghetti* is not among them, even though, to be sure, the entry for it is in the dictionary.



3 entries found for **das**.

cream soda (noun)
giant panda (noun)
piña colada (noun)

cream soda noun

plural ~ **-das**

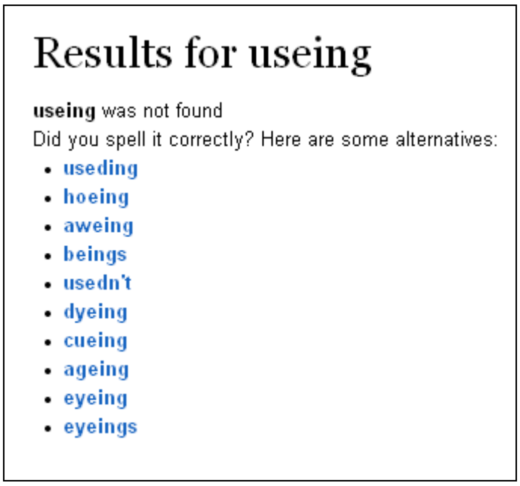
[*count*, *noncount*] : a type of sweet, bubbly soft drink that is

Figure 3: MWALED's suggestions for **das*, a misspelling of *does*

MWALED's algorithm seems to focus excessively on transpositions — it tends to rearrange the original letters: it offers *heir* for **hier* (*here*), *tire* for **trie* (*try*), but also *grade* for **gread* (great) and *crane* for **crean* (*clean*).

Life is made difficult for the spellchecker by the oddity of some of the entries in the dictionary. This is to some extent true of all our dictionaries, but especially of GoogleED. In the absence of any data on word frequency — and it does not seem to be using any — these odd words just enlarge the set of (apparently) plausible corrections, and so we find the following unhelpful suggestions among the 'best' guesses: *deferens* (probably from *vas deferens*), *etyma*, *xylem*, *inf*, *umbrae*, *commis*, as well as proper names like *Du Bois*, *Tok Pisin* and *Wat Tyler*. On top of that, GoogleED would not infrequently provide suggestions that are clearly not genuine words, and often only partially closer than the misspelling to any real English words. Thus, GoogleED offered **petryszyn* for **repetyszyn* (a Polish misspelling of *repetition*), **sejfy* for **sejfty* (*safety*), **trulli* for **truli* (truly), **sinirli* for **sinsirli* (*sincerely*), **temprecher* for **tempreczer* (*temperature*), **bicikli* for **beisikli* (*basically*), **existens* for **egzistens* (*existence*), **identiti* for **aidentiti* (*identity*).

The *–ing* ending seemed to be another cause of difficulty for these dictionaries. Of the lot, only GoogleED is able to correct **useing* to the intended *using*. Instead, LDOCE Free and MEDO offer *unseeing* (true: not entirely unlikely), LDOCE Premium suggests *suing*, MWALED *seeing*, and — strangest of all — CALD proposes the nonce form *useding* (see Figure 4), apparently as a hypothetical inflected form of *used to*, as this is the entry to which the user is taken upon clicking on *useding*.



Results for useing

useing was not found

Did you spell it correctly? Here are some alternatives:

- useding
- hoeing
- aweing
- beings
- usedn't
- dyeing
- cueing
- ageing
- eyeing
- eyeings

Figure 4: CALD suggestions for the target word *using* misspelled as **useing*

Another easy case is **diging*, a straightforward misspelling of *digging*. As for *useing* above, GoogleED gets it right, and so does ALD this time. LDOCE Free suggests *dining* (and, in third position, *diggings*, but never *digging*!). MWALED insists on *Diegan*, MEDO would like *dinging*, and CALD — *ziging*.

A rather striking feature of LDOCE (especially the free version) is that it likes to make two correct words by sticking a space in the middle of the misspelling, thus: *of fen* for **offen* (*often*), *inter fir* for **interfir* (*interfere*), *so rid* for **sorid* (*solid*), *back en* for **backen* (*bacon*), *be course* for **becourse* (*because*), *ail and* for **ailand* (*island*). This strategy may be occasionally successful when checking running text, but it does not work well for isolated dictionary query strings, especially if the spellchecker does not care whether the resulting pair is a likely combination.

Apart from that, LDOCE's offerings, among the dictionaries tested, tend to be the most respectful of the misspellings. The suggestions tend to retain the first letter and the general word structure.

4. Can the dictionaries do better? Mitton's experimental spellchecker

As the online dictionaries clearly performed below expectation, the first author wondered if there were context-free spellcheckers capable of outperforming, if not all, then at least some of the dictionaries. As a result of a literature search, a promising context-free experimental spelling correction system was identified (Mitton, 1996, 2009). Consequently, the second author was contacted and offered to run the same data through his spellchecker.

There is no space here to describe Mitton's spellchecker in detail, and the interested reader is invited to consult Mitton (2009) or, for even greater detail, Mitton (1996). At its heart is a dictionary primed with information about the quirks of English spelling. If faced with, say, **morgage*, it would consider *mortgage* a likely candidate because the entry for *mortgage* contains the information that the *t* is likely to be omitted. It also makes use of word frequency in ordering its list of suggestions.

Table 2 compares the success rates (as in Table 1) of Mitton's experimental spellchecker with the best-performing online dictionary (LDOCE Free), and Figure 5 compares it with all the dictionaries graphically.

	Target word listed in position:			
Dictionary	First	Top 3	Top 6	Top 10
Mitton	73%	87%	91%	93%
LDOCE Free	51%	65%	75%	79%

Table 2: Success rates of the best-performing dictionary compared with Mitton's experimental spellchecker, for all data

Mitton's spellchecker was able to place the intended target word among the top ten of its list of suggestions for 93% of the misspellings. The best dictionary in our set, LDOCE Free, performed significantly worse, achieving a success rate of 79%. The gap is even greater if we consider the spellchecker's ability to place the target word in the most valuable top portion of the list of suggestions. Here the experimental spellchecker outperforms LDOCE Free by over 20 percentage points (*First* and *Top 3*). Mitton's spellchecker manages to identify the intended word as the top candidate for 73%, as against 51% for LDOCE Free.

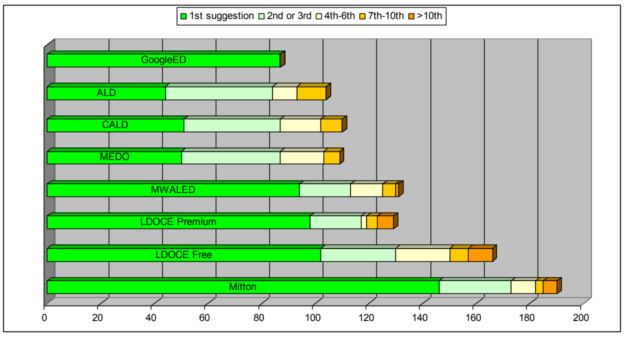


Figure 5: Performance of the seven dictionaries compared with Mitton's experimental spellchecker, for all data (N=200)

In comparison with the other dictionaries (Figure 5), the gains are still greater. The top-of-the-list success rates of ALD, CALD, and MEDO are only a third of that of Mitton's spellchecker. From another perspective, the experimental spellchecker was able to guess perfectly 23 items that none of the seven dictionaries got right.

5. Polish, Japanese, and Finnish misspellings compared

The results we have presented so far are based on aggregated data from the three subcorpora. Now we will take a closer look at the role of the native language.

Our corpus included misspellings from native speakers of three different languages — Polish, Japanese, and Finnish. Figure 6 gives the language-specific success rates in terms of the target word appearing at the top of the suggestions list, while Figure 7 includes percentages for the target word appearing in the top ten of the list.

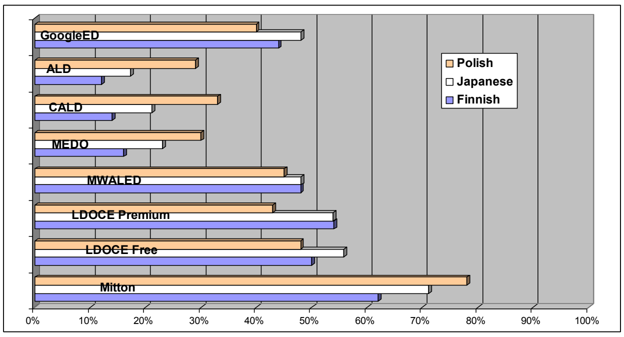


Figure 6: Percentage of Polish, Japanese, and Finnish misspellings for which the target word appeared as the first suggestion

In comparing these language-specific results with aggregated figures, we need to bear in mind that the Polish subcorpus contributed the most to the overall figures, as it represents half the data, with the Japanese and Finnish subcorpora accounting for a quarter of the corpus each. In terms of the target word being offered as the best suggestion, four systems (ALD, CALD, MEDO, and Mitton's spellchecker) seem to cope better-than-average with the Polish misspellings, while for the remaining four (GoogleED, MWALED, LDOCE Premium, and LDOCE Free), the reverse is the case. Since the Polish data were elicited via audio stimuli, this may have to do with the inclusion or otherwise of phonological awareness, explicit or implicit, rather than with specifically L1-induced misspelling patterns. Still, it is also true that part of the Finnish data came from written responses to spoken dialogue, and the Japanese misspellings were partially inspired by their katakana representations, so it might be said that all three subcorpora had some sound-motivated misspellings.

What is clear, however, is that ALD, CALD, and MEDO would have done even more poorly overall, had the Polish subcorpus not been given more weight than the others: their success with the Finnish misspellings was only half — at best — of that with the Polish data, with the Japanese figures in-between the two.

In placing the required word at the top of the list, Mitton's spellchecker did very well with the Polish and Japanese data, and not quite as well with the Finnish misspellings (though it still outperformed all the dictionaries).

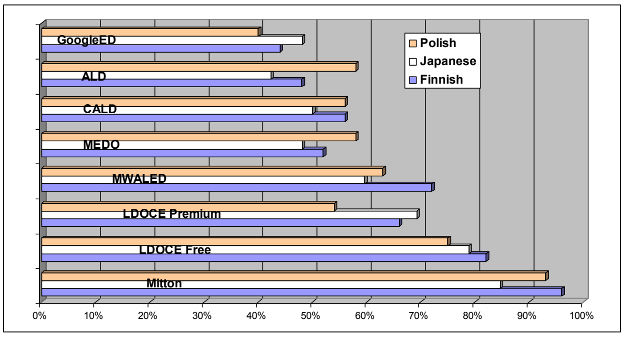


Figure 7: Percentage of Polish, Japanese, and Finnish misspellings for which the target word appeared among the top ten suggestions

Moving on now to the results for the top ten suggestions (given in Figure 7 above), we can see that there is no longer that much difference due to the native language, even for ALD, CALD, and MEDO. Apparently, the three dictionaries can still capture about half of the target words in the top ten suggested items, though somehow they find it much harder for Polish to guess the best suggestion correctly than for the other two languages.

Mitton's spellchecker performed reliably for all three languages, getting 93% of the Polish target words in the top ten, and no less than 96% of the Finnish items (actually, all of the 96% also made it into the top six suggestions).

6. Ways to improve spelling correction in e-dictionaries

6.1 Customization

While for many years the primary focus of research into spelling correction has been on native writers, recently the needs of non-native users of language speakers, particularly English, have begun to receive some attention (for an overview, see e.g. Heift & Rimrott, 2008). It is now recognized that the patterns of misspelling of non-native speakers differ both in quality and quantity from those of native users of a language. Thus, if the L1 of the user is known to the system (be it based on the Accept-Language http header, IP Geolocation, or individual user profile), the dictionary interface might use an algorithm optimized for that native language. In fact, Mitton's spellchecker used in this study has already seen a successful adaptation to better handle the typical misspellings of Japanese learners writing in English (Mitton & Okada, 2007).

However, we would not expect the influence of L1 to be uniform across a wide range of L2 proficiency levels. To account for this variation as well as for individual idiosyncrasies, customization might in the future go even further: it might be possible to design an adaptive spelling corrector, capable of tuning in to the particular areas of spelling problems exhibited by a given user.

6.2 Greater phonological awareness

A large proportion of the items at which all the dictionaries failed are recognizable as attempts at rendering the pronunciation of the English word through the spelling conventions of the native language. This is particularly evident in the case of the Polish data, no doubt partially as a result of using audio stimuli for data elicitation. Evidence for this 'phonetic access' strategy (here largely subconscious, cf. Sobkowiak, 1999) is seen in the use of L1-specific letter combinations (such as, for Polish, <sz>, <aj>, or <ej>) to approximate English pronunciation. Mitton's spellchecker handled many of these cases quite well, perhaps thanks to its level of phonological awareness, even though it has not been made aware of any Polish-specific letter-to-sound correspondences. Making provision for a few of the most common such correspondences could significantly improve a spellchecker's performance.

6.3 Dealing with real-word errors

In section 2.2 we discussed the issue of rare words. To use a specific example from the study, one of the misspellings in the corpus was **wold* for *would*. As it turns out, *wold* is also an English word, albeit very rare. Consequently, most occurrences of *wold* in text will be misspellings, and a text spellchecker would do well to flag it as a possible error. However in a dictionary look-up situation, unlike in text spellchecking, it would be risky to withhold a rare-word entry from the user and offer instead similarly-spelled frequent words. Even though the core vocabulary of a few thousand words (De Schryver et al., 2006) are looked up more commonly than the rest, it is also true that the less frequent items have a reasonable chance of being looked up (see the discussion in 2.2 above). How should a dictionary respond to such a query?

The answer need not necessarily be the same for *any* dictionary. A user of the online version of, say, the OED is much more likely to want an entry for this obscure word than a user of an intermediate learners' dictionary. The latter dictionary might not hold the word in its wordlist at all, in which case the issue would not arise. But if it did, a happy compromise might be to take the user to the rare word entry, but at the same time alert them in a sidebar saying something like 'Did you perhaps mean *world'*?

6.4 First things first

We have suggested possible avenues to improve success in correcting misspelled dictionary search terms. However, it needs to be stated emphatically that it would be misguided to pursue any such attempts at tweaking the interface before more basic problems are addressed. This study has revealed that such fundamental problems are numerous and grave, and they affect the most authoritative of English monolingual learners' dictionaries.

7. Conclusion

Our study has shown that the leading monolingual English learners' dictionaries are inadequate when it comes to correcting misspelled input from non-native users. When challenged with a misspelling, far too often the dictionaries are unable to include the word actually intended in their list of suggestions, and if they do include it, the ordering of the alternatives is often less than optimal. While the individual dictionaries vary substantially in performance, there is much room for improvement for even the best ones, and we have shown that an experimental spellchecker achieves much greater success rates than any of the dictionaries, even though it has not been designed with non-native speakers in mind.

8. Acknowledgements

The first author wishes to thank his student assistants, Marta Dąbrowska and Aleksandra Lasko, for their help in collecting the Polish corpus of misspellings.

9. References

9.1 Online dictionaries tested

ALD. *Oxford Advanced Learner's Dictionary*. http://www.oxfordadvancedlearnersdictionary.com/

CALD. *Cambridge Advanced Learner's Dictionary*. http://dictionary.cambridge.org/

GoogleED. *Google English Dictionary*. At the time of collecting data: http://www.google.com/dictionary; at the time of writing the present version: http://www.google.com/search?q=%s&tbs=dfn:1 (where %s stands for the search term)

LDOCE Free. *Longman Dictionary of Contemporary English*. http://www.ldoceonline.com/

LDOCE Premium. *Longman Dictionary of Contemporary English*. http://www.longmandictionariesonline.com/

MEDO. *Macmillan English Dictionary Online*. http://www.macmillandictionary.com/

MWALED. *Merriam-Webster's English Learner's Online Dictionary*. http://www.learnersdictionary.com/

9.2 Other references

Atkins, B.T.S. (1996). Bilingual dictionaries - past, present and future, in Gellerstam, M., Jarborg, J., Malmgren, S.-G., Noren, K., Rogström, L. & Papmehl, C.R. (eds.), *EURALEX '96 Proceedings*. Göteborg: Department of Swedish, Göteborg University, pp. 515-546.

Bank, C. (2010). Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen. M.A. Thesis, Universität Hildesheim.

Bogaards, P. (2010). The evolution of learners' dictionaries and *Merriam-Webster's Advanced*

Learner's English Dictionary, in Kernerman, I., Bogaards, P. (eds.), *English Learners' Dictionaries at the DSNA 2009*. Tel Aviv: K Dictionaries, pp. 11-27.

Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors, *Communications of the A.C.M.* 7, pp. 171–176.

De Schryver, G.-M., Joffe, D., Joffe, P. & Hillewaert, S. (2006). Do dictionary users really look up frequent words? – On the overestimation of the value of corpus-based lexicography, *Lexikos* 16, pp. 67-83.

Deorowicz, S., Ciura, M. (2005). Correcting spelling errors by modelling their causes, *International Journal of Applied Mathematics and Computer Science* 15(2), pp. 275-285.

Hanks, P. (2009). Review of Stephen J. Perrault (ed.). 2008. *Merriam-Webster's Advanced Learner's English Dictionary*, *International Journal of Lexicography* 22(3), pp. 301-315.

Hanks, P., Pearsall, J. (eds.). (1998), *New Oxford Dictionary of English*. Oxford: Oxford University Press.

Heift, T., Rimrott, A. (2008). Learner responses to corrective feedback for spelling errors in CALL, *System* 36(2), pp. 196-213.

Kukich, K. (1992). Techniques for automatically correcting words in text, *Computing Surveys* 24(4), pp. 377-439.

Lew, R. (2011). Online dictionaries of English. In P.A. Fuertes-Olivera, H. Bergenholtz, (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, pp. 230-250.

Lindberg, C. (ed.) (2006). *Oxford American College Dictionary*, 2nd edition. Oxford: Oxford University Press.

McKean, E. (ed.) (2005). *New Oxford American Dictionary*, 2nd edition. Oxford: Oxford University Press.

Mitton, R. (1985). Birkbeck spelling error corpus. Accessed at: http://ota.oucs.ox.ac.uk/headers/0643.xml.

Mitton, R. (1996). *English spelling and the computer*. Harlow: Longman.

Mitton, R. (2009). Ordering the suggestions of a spellchecker without using context. *Natural Language Engineering* 15, pp. 173-192.

Mitton, R., Okada, T. (2007). The adaptation of an English spellchecker for Japanese writers. *Symposium on Second Language Writing*. Nagoya, Japan.

Nielsen, S. (1995). Alphabetic macrostructure. In H. Bergenholtz, S. Tarp (eds.) *Manual of Specialised Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 190-195.

Pollock, J.J., Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text, *Communications of the A.C.M.* 27(4), pp. 358–368.

Proctor, E. (2002). Spelling and searching the internet: An overlooked problem, *The Journal of Academic Librarianship* 28(5), pp. 297-305.

Sobkowiak, W. (1999). *Pronunciation in EFL machine-readable dictionaries*. Poznań: Motivex.

Suomi, R. (1984). Spelling errors and interference errors in English made by Finns and Swedish-speaking Finns in the 9th form of comprehensive school. MA Thesis, Abo Akademi.

Yannakoudakis, E.J., Fawthrop, D. (1983). The rules of spelling errors. *Information Processing and Management* 19(2), pp. 87-99.

Appendix: Sample misspellings

SUBCORPUS	TARGET	MISSPELLING
PL	certain	serten
PL	easily	izli
PL	guarantee	garanti
PL	interfere	interfir
PL	interruption	interapsion
PL	library	lajbery
PL	psychology	sakology
PL	receive	reseve
PL	separate	sepret
PL	succeed	sukcid
JP	albatross	albatlos
JP	antenna	untena
JP	beautiful	butiful
JP	embarrass	enbarance
JP	enough	inaf
JP	gallery	garally
JP	graph	glaf
JP	laughter	lafter
JP	neglect	nigrect
JP	umbrella	umblera
FI	because	becourse
FI	colour	coulor
FI	delicious	delecous
FI	especially	espessially
FI	gasoline	gazolin
FI	good-bye	goodbay
FI	orchestra	orkester
FI	symphony	sinfony
FI	temperature	tempeture
FI	universities	univercitys

There And Back Again – from Dictionary to Wordnet to Thesaurus and Vice Versa: How to Use and Reuse Dictionary Data in a Conceptual Dictionary

Henrik Lorentzen, Lars Trap-Jensen
Society for Danish Language and Literature
Christians Brygge 1, 1219 Copenhagen K, Denmark
E-mail: hl@dsl.dk, ltj@dsl.dk

Abstract

This is a story of elexicographical evolution and how lexical data are used and reused to develop new products and new presentations. At eLex 2009 we demonstrated how data from The Danish Dictionary were used to construct a wordnet for Danish, DanNet, and we showed how, in a betaversion, DanNet data could be used to improve the onomasiological component of the dictionary. Since then, we have used both DanNet and dictionary data in an ongoing project to create a conceptual dictionary – a thesaurus – for Danish. In this article we will show how the thesaurus data help us overcome the major problems connected with the direct use of wordnet data in a dictionary for human users. Focus is on the editing principles of the thesaurus and how data are presented in The Danish Dictionary online.

Keywords: thesaurus; wordnet; onomasiological search; e-dictionary

1. Background

A dictionary stock of 91,500 entries with 116,000 meaning descriptions and a wordnet consisting of 65,000 synsets connected through 75,000 internal semantic relations provide the starting point for a thesaurus project which is currently being developed at the Society for Danish Language and Literature in Copenhagen. Historically, data from The Danish Dictionary (Den Danske Ordbog, DDO) were used to construct DanNet, and since both serve as input for the thesaurus, all three resources are closely interconnected. The Danish thesaurus project set off in 2010 and will appear in 2013 as a publication of its own – and even as a printed dictionary.

In this connection, focus is on the thesaurus data and how they can be used to improve the onomasiological component of the online version of the DDO. In Louvain 2009, we gave a presentation of the dictionary site *ordnet.dk*, of which the DDO is but one element (the others being a comprehensive historical dictionary and a corpus component), and we showed how data from the Danish wordnet were exploited in a new element, *Related words*, that was introduced in the online version of the dictionary (Trap-Jensen, 2010). We demonstrated how candidates for *Related words* could be automatically extracted from DanNet but it was emphasized – as it still is in the version available to the public – that it is a beta version and not without its problems. In this article, we dwell on the nature of the problems involved and how they can be solved by using data from the thesaurus instead.

2. Shortcomings of wordnet data

In Trap-Jensen (2010), some problematic areas were mentioned and possible solutions suggested. Let us briefly recapitulate the central issues as well as other shortcomings that we have encountered since.

First, there is the problem of overgeneration. This pertains to both co-hyponyms and hyponyms and is due to the fact that the categories are often too broad and the hyponymy hierarchy too shallow. From time to time, Princeton WordNet has been criticised for having too deep and detailed a hierarchical semantic structure and it is therefore important to stress that DanNet is not a translation of Princeton WordNet but was built on original Danish data primarily extracted from the DDO, mainly to ensure that the conceptual world reflected by the Danish language is maintained in the resource.

The combination of a relatively shallow semantic hierarchy and a limited number of semantic classes – DanNet operates with approximately 200 ontological types as opposed to the 900-1,000 semantic groups found in thesauri like Roget (2002) for English or Dornseiff (2004) for German – implies a built-in risk of too broad semantic categories resulting in too many and not always the most evident related words when the candidates are automatically extracted from DanNet. Examples in DanNet are the large groups of vocabulary items for persons, plants and verbal nouns with hundreds or, in extreme cases, even thousands of co-hyponyms. To some extent this can be remedied by using combinations of other relations to narrow down the number of members in each category. Our solution to the problem, as described in Trap-Jensen (2010), has been to develop an algorithmic method that ranks words from large groups, primarily based on the number and nature of shared relations. Even if the algorithm has improved usability, the overall conclusion is nevertheless that a good deal of manual effort would still be needed for this element to work properly for the human user.

A second major problem with the DanNet data relates to the fact that synsets are seldom found in more than a single place within the semantic network. For example, the assignment of multiple hyperonyms to synsets, although by no means impossible or forbidden, is still a relatively rarely used possibility: out of 65,000 synsets in DanNet, less than 500 have been assigned more than one hyperonym. This is primarily a consequence of the editorial process in DanNet: multiple hyperonyms are almost exclusively used when the dictionary definition contains more than one genus proximum. For example, in DDO *daughter* is defined as 'a girl or a woman of whom you are the father or mother' and *juniper* 'little tree or bush that bears juniper berries'. Because DanNet was constructed from the dictionary, the editor would be prompted that 'girl' and 'woman' were both possible hyperonyms of *daughter* and 'tree' and 'bush' possible hyperonyms for *juniper* but no independent routine was carried out to decide whether this was the case for other vocabulary items.

Multiple hyperonyms should not be confused with the situation known as systematic polysemy where a word, e.g. *school*, is ascribed a dual meaning: 'building' and 'institution' respectively. The difference is that the latter situation involves clearly distinct meanings belonging to different synsets. An inheritance mechanism in the editing tool helps the editor make sure that similar words and in particular hyponyms are coded in the same way.

In a traditional thesaurus, the picture is somewhat different. Here it is quite common for a word to appear in different thematic groups: for example, a *guinea pig* is at the same time a 'South American mammal', a 'rodent', a 'pet' and – at least in some parts of the world – an 'edible animal'. In other words, a word in a particular sense is not confined to occur under the nearest hyperonym alone. Sometimes it makes sense to use instead a subset of the hyperonym's hyperonym ('South American mammal' as opposed to 'rodent'), sometimes it is not the taxonomical position that matters at all but the role that the entity plays in a particular context ('pet' or 'edible'). Humans are overall more creative and flexible in the way they encode and decode meaningful categories than computers are.

People's ability to carve up the world in new ways and the assignment of multiple hyperonyms are both reflected in the index of a thesaurus: many words have several references to the systematic part. Of course, one has to remember that not all the references of a word concern the same meaning as the index only lists the form of each word and hence does not distinguish between homographs and different word senses. This is why the notion 'synset' (a set of word forms signifying a single concept) is so important in the wordnet universe whereas it has no counterpart in the common language. However, the point is here that even if we take this into account, it is not unusual for a particular word in a specific sense to be placed in several thematic groups of a thesaurus.

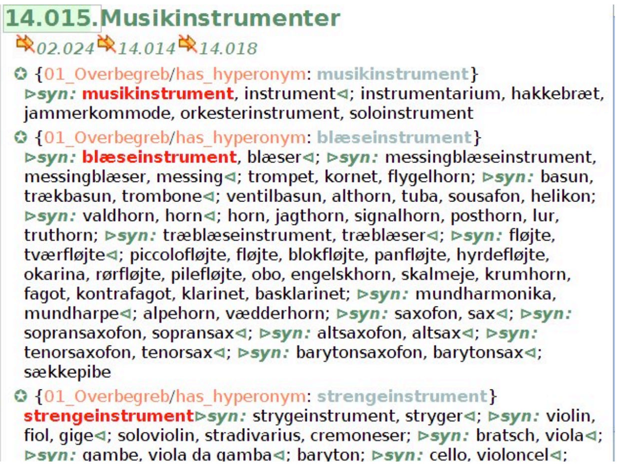
In the end we found that the difference between an NLP resource like DanNet and the human user's need for onomasiological assistance was beyond quick repair. Instead we have decided that the wordnet data should not be displayed directly in the dictionary but only indirectly – through the central role they play in building the thesaurus. How this is done is the subject of the next section.

3. Editing a thesaurus using DanNet and DDO

As a first step in the early phase of the project it was decided to copy the ontological structure of the German thesaurus *Der deutsche Wortschatz nach Sachgruppen* founded by Dornseiff, instead of building an ontology from scratch. This may seem a quick and dirty solution but considering the fact that Danish and German are closely related languages (and cultures) we concluded that the solution was justified. As a matter of fact, the predecessor of our Danish thesaurus (Andersen, 1945) was also to a large extent based on Dornseiff, which as a side effect makes it easier for us to supplement with material from the older thesaurus.

Dornseiff and consequently our thesaurus are divided into 22 main sections (chapters) ranging from *Natur und Umwelt* (Nature and Environment) over *Kunst und Kultur* (Art and Culture) to *Religion*. These chapters are further subdivided into 906 thematic groups. We are fully aware of the risk of taking over a meaning structure from German without it being fully adapted to Danish, and remind the reader that this merely serves as the starting point. Minor changes are to be expected as the editorial work progresses, for instance merging of groups or, on the other hand, splitting groups where Danish and German do not classify the world in exactly the same way.

When starting to edit a new thematic group the editor begins by extracting raw material from DanNet. Typically a central hyperonym, such as *orchestra* within the field of music, is selected as a starting point together with all its hyponyms. Relevant information about each concept such as definition, subject domain and part of speech is copied into the thesaurus database as well as ID numbers and other metadata necessary to ensure links between the two databases. The process can be reiterated any number of times in order to supply more central concepts and their hyponyms but usually this step does not exhaust the material entirely. Additional searches will then be made, notably in DDO, where relevant material can be extracted by combining search parameters like specific words in definitions and a particular subject field. Once an appropriate number of concepts has been obtained, the editor begins to sort the concepts into smaller groups. These groups are held together and labelled by a headword, often the hyperonym. The field of music can serve as a case in point: it contains groups like genre, volume, tempo, rhythm, each one consisting of hyponyms such as *baroque music*, *chamber music*, *military music* (genre), *crescendo*, *forte*, *pianissimo* (volume), *adagio*, *moderato*, *allegro* (tempo), and *beat*, *triplet*, *syncopation* (rhythm). Generally, properties (adjectives) are grouped together: *instrumental*, *vocal*, *symphonic* as well as verbs: *play*, *practise*, *compose.* Other groups unite places and institutions like *academy of music*, *concert hall*, *discotheque* and parts of musical instruments such as *string*, *keyboard* and *pedal*. In the latter case it is the holonym that serves as headword.



14.015.Musikinstrumenter

02.024 14.014 14.018

{01_Overbegreb/has_hyperonym: musikinstrument}
▷*syn:* **musikinstrument**, instrument◁; instrumentarium, hakkebræt, jammerkommode, orkesterinstrument, soloinstrument

{01_Overbegreb/has_hyperonym: blæseinstrument}
▷*syn:* **blæseinstrument**, blæser◁; ▷*syn:* messingblæseinstrument, messingblæser, messing◁; trompet, kornet, flygelhorn; ▷*syn:* basun, trækbasun, trombone◁; ventilbasun, althorn, tuba, sousafon, helikon; ▷*syn:* valdhorn, horn◁; horn, jagthorn, signalhorn, posthorn, lur, truthorn; ▷*syn:* træblæseinstrument, træblæser◁; ▷*syn:* fløjte, tværfløjte◁; piccolofløjte, fløjte, blokfløjte, panfløjte, hyrdefløjte, okarina, rørfløjte, pilefløjte, obo, engelskhorn, skalmeje, krumhorn, fagot, kontrafagot, klarinet, basklarinet; ▷*syn:* mundharmonika, mundharpe◁; alpehorn, vædderhorn; ▷*syn:* saxofon, sax◁; ▷*syn:* sopransaxofon, sopransax◁; ▷*syn:* altsaxofon, altsax◁; ▷*syn:* tenorsaxofon, tenorsax◁; ▷*syn:* barytonsaxofon, barytonsax◁; sækkepibe

{01_Overbegreb/has_hyperonym: strengeinstrument}
strengeinstrument▷*syn:* strygeinstrument, stryger◁; ▷*syn:* violin, fiol, gige◁; soloviolin, stradivarius, cremoneser; ▷*syn:* bratsch, viola◁; ▷*syn:* gambe, viola da gamba◁; baryton; ▷*syn:* cello, violoncel◁;

Figure 1: Section from the thematic group *musical instruments* in the editing tool

It appears that DanNet and DDO are by far the most important resources and the main reason why it is possible at all to develop a dictionary of this kind within a limited period of time. Once a satisfactory semantic structure has been established, however, the editor also turns to look at other sources in order to supplement the thematic group in question. These sources include the older thesaurus (Andersen 1945), a Danish dictionary of synonyms (Schultz), a dictionary of slang and informal language (Politiken) as well as other relevant reference works. Schultz is available in electronic format which allows semi-automatic comparisons with the first draft version of the thematic group.

4. Presenting the data

At the moment 145 out of the 906 thematic groups have been completed in a draft version ready for futher editorial treatment, i.e. correction of errors, supplementing missing concepts, rearranging of groups and concepts. As mentioned in section 2, the current presentation of *Related words* will be replaced by thesaurus data once these are available. Let us consider how they can be used to offer fewer but more relevant candidates as *Related words* for a particular word meaning. We can use another example from the field of music, the aforementioned word *discotheque* in the sense 'a place to dance'. This concept is found in the following four groups: *Place*, *Dance, Popular music* and *Pleasure and leisure time*. In the dictionary entry for *discotheque* in the element *Related words*, the user will be presented with four snippets along these lines (with the relevant words in English translations):

1 PLACE discotheque, bar, night club; tivoli, amusement park, theme park
2 DANCE discotheque, dancing stage, ballroom, rehearsal room; show, striptease, ball
3 POPULAR MUSIC music venue, discotheque, jazz venue
4 PLEASURE restaurant, discotheque, night disco, dance restaurant

Each word is provided with a mouse-over function giving the definition from DDO. Furthermore, the headlines of the four groups are clickable, taking the user to a thesaurus presentation where each group of concepts are expanded to the level of thematic group; in the case of *discotheque* the first group is that of *Place* and the user will see a list of concepts denoting for instance places where humans perform an activity, ranging from *workshop* and *laboratory* over *stadium* and *sports centre* to *headquarters* and *executive's office*.

The abbreviated form of *discotheque*, *disco*, can be used in the same sense but also in the sense of 'type of music and dance'. It is important that the user is made aware of this and is allowed to navigate to this new thematic group. S/he can do so either via links to related groups or via a search field present at the top of every thesaurus page. In the case of *disco* in the second sense, the relevant thematic group will contain another type of related words: *dance, disco, breakdance, hiphop, headbanging, limbo, zumba*.

The reason why only snippets are given in the dictionary entry is of course to avoid imposing too much material on a user who perhaps is looking up for entirely different reasons. Only words from the very subgroup in which the entry word appears are displayed. If the user wants to see more s/he will have to go to the thesaurus page for a complete overview.

The thesaurus page shows the entire thematic group with the search word highlighted and with references to other groups that are semantically close. A thematic group contains an average of 270 words and is divided into subgroups on two hierarchical levels. Within each subgroup the first word functions as a heading for the following words.

The structure within the thematic group as well as the group-internal order of words are semantically based so that the reader intuitively recognizes a 'natural' or 'logical' organization and progression as s/he reads through the groups.

This organization also reflects the editor's way of organizing the thematic group. There has, however, been heated discussion among the editors whether this is also the best way of presenting data. Some thesauri use alphabetical ordering at the lowest level of grouping – among them are Dornseiff and Andersen, whereas the order in Roget is 'logical' or semantical. And all thesauri that we have consulted use part of speech as a more general dividing criterion than semantics.

Following Wiegand (2004) the overall purpose of a monolingual conceptual dictionary depends on the dominant consultancy situation:

> Bei der oben formulierten Charakterisierung des genuinen Zwecks von großen einsprachigen Sachgruppenwörterbüchern wurden bei den Mögligkeiten, die Wörterbücher dieses Typs dem Benutzer eröffnen, drei verschiedene Arten [unterschieden:
>
> (i) Konsultationssituation wegen Ausdrucksfindungs-schwierigkeiten
> (ii) Konsultationssituation wegen äquivalentbezogener Angemessenheitszweifel
> (iii) Konsultationssituation wegen Wissensbedarf über einen bestimmten Aspekt des deutschen Wortschatzes]
>
> Die erste Möglichkeit … wird wahrscheinlich am meisten genutzt. Das ist auch der Grund dafür, warum die Wörterbuchartikel im Formteil nach Ausdrucksklassen und damit auch nach Wortarten gegliedert sind. Die kundigen Benutzer, die Ausdrucksfindungs-schwierigkeiten haben, werden durch diese artikelinterne Datenanordnung präferiert.
>
> Wiegand, 2004: 59 (i-iii summarized from 57-58)

If it is true that word-finding problems are the most common motive for thesaurus consultation, i.e. that the user is looking for paradigmatic alternatives to the search word, then part of speech seems well-placed as the superior criterion. If, on the other hand, knowledge needs are more common, then the semantic criterion is preferable. In this connection, Wiegand's second situation can be neglected as it refers to bilingual contexts only. In the printed dictionary, a decision must eventually be made, whereas – at least technically – it is an option to have two presentations in the online dictionary and leave it to the user to decide. This is one of the questions that have not yet been answered.

5. Perspectives

A remarkable feature of the resources in *ordnet.dk* is the fact that all the elements are closely interconnected, with mark-up of the building blocks at a fairly detailed level. A system of unique ID numbering maintains links between units, not only between lemmas in the dictionary but also at sub-lemma level between the senses of an entry in the dictionary and the synset members in DanNet and the concepts in the thesaurus. This way of organizing data is also what makes it possible for us to enrich the other resources and in that way facilitate further elexicographical evolution. When the thesaurus is complete the resulting data can be reused to improve the wordnet data. We cannot go into all the details here but just mention a few promising areas.

The thematic information contained in the thesaurus should be imported into DanNet. At present, there is little information about the degree of semantic closeness between the units in DanNet. Synonyms (members of a synset) and near-synonyms are specified if available but beyond that no finer distinctions are made. Even if the hyponymy hierarchy does say something about basic conceptual organization, it does not account sufficiently for the degree of similarity between concepts. Here, the semantic grouping on up to three different levels within a chapter represents an alternative way of categorization which could contribute significantly to the semantic richness of DanNet.

We mentioned earlier that multiple hyperonyms are under-represented in DanNet, primarily because the dictionary definitions usually do not contain more than one genus proximum. The thesaurus editors, however, often assign a second superordinate term, for example when they group objects according to their function instead of, say, physical shape. For example, the dictionary definition of *drumstick* reads 'a long wooden stick used for playing drums', and consequently *drumstick* is coded as 'a kind of stick' in DanNet but not as 'a piece of music equipment'. This situation could be remedied by using thesaurus data.

Finally, the thesaurus would be helpful to broaden the coverage of DanNet. An obvious area is the treatment of 2nd and 3rd order entities (as used by Lyons, 1977), for example properties that are not described in great detail in the current version of DanNet. In the thesaurus, however, adjectives are often grouped together on the basis of the nouns they modify: 'suntanned' is a property of humans, 'wire-haired' a property of dogs, 'carnivorous' a property of predators etc. Properties make up a surprisingly large proportion of the vocabulary – a rough estimate says that about 20 % of the vocabulary covered in the thesaurus so far pertain to properties. Another area where coverage could be broadened is the vocabulary from other parts of speech than the traditional content words as well as multiword units.

6. References

Andersen, H. (1945). *Dansk Begrebsordbog* (Danish Conceptual Dictionary). Copenhagen: Munksgaard.

DanNet. Accessed at: http://wordnet.dk.

Dansk Synonymordbog (1992). (Danish Dictionary of Synonyms), 8th edition. Copenhagen: J.H. Schultz Information Ltd.

Den Danske Ordbog (The Danish Dictionary). Accessed at: http://ordnet.dk/ddo.

Dornseiff, F. (2004). *Der deutsche Wortschatz nach*

Sachgruppen, 8th edition. Berlin/New York: Walter de Gruyter.

Lyons, J. (1977). *Semantics*. Press Syndicate of the University of Cambridge.

Politikens Slangordbog (1993). (Politiken's Dictionary of Slang), 4th edition. Copenhagen: Politikens Forlag.

Roget, P.M. (2002). *Roget's Thesaurus*, 150th anniversary edition edited by George Davidson. London: Penguin.

Trap-Jensen, L. (2010). Access to Multiple Lexical Resources at a Stroke: Integrating Dictionary, Corpus and Wordnet Data. In S. Granger, M. Paquot (eds.) *eLexicography in the 21st Century: New challenges, new applications. Proceedings of eLex 2009. Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 295-302.

Wiegand, H.E. (2004). Lexikographisch-historische Einführung. In F. Dornseiff (ed.) *Der deutsche Wortschatz nach Sachgruppen*, 8th edition. Berlin/New York: Walter de Gruyter, pp. 9-100.

Representing Nouns in the *Diccionario de aprendizaje del español como lengua extranjera* (DAELE)

Viviana Mahecha Mahecha, Janet DeCesaris
Institut Universitari de Lingüística Aplicada, Pompeu Fabra University
Roc Boronat, 138   08018 Barcelona, Spain
E-mail: viviana.mahecha@upf.edu, janet.decesaris@upf.edu

Abstract

This paper reports on the representation of nouns in the Diccionario *de aprendizaje del español como lengua extranjera*, an ongoing on-line dictionary prototype being developed at the Institut Universitari de Lingüística Aplicada. The DAELE is designed for upper-intermediate students of Spanish as a foreign language and is corpus-based. We discuss our decisions concerning the macrostructure and microstructure for noun entries, and the procedure we followed to obtain a representative sample of nouns in terms of grammatical structure and semantic field. In our project, given the expected characteristics of users, we have opted for full-sentence definitions, and we follow Lara (1998) in attempting to organize senses within entries according to semantic criteria. Issues discussed in this paper include the ordering of senses, the choice of examples, and the type of semantic label chosen to guide users.

Keywords: Spanish learner's dictionary; DAELE; nouns; lexicological issues of lexicographic importance

1. Introduction

Many decisions made by lexicographers in the past were conditioned by the space limitations imposed by printed books. It is obvious that for electronic dictionaries consulted on-line, space limitations, which *inter alia* conditioned decisions on the number of headwords defined, the defining style used, the number and length of examples provided, the presence of less frequent senses, and inclusion of illustrations and or pictures, are no longer valid. Nevertheless, it is not clear that simply providing more information is helpful to all dictionary users, many of whom turn to a dictionary to look up quite specific information. The electronic format requires lexicographers to reconsider the way they structure information because layers of information are progressively accessed. In this paper we report on some issues that have arisen in our work on the representation of nouns in the *Diccionario de aprendizaje del español como lengua extranjera* (DAELE) that are related to the relationship between dictionaries and grammar.

2. Overview of the DAELE project

The DAELE is an ongoing electronic dictionary project at the Institut Universitari de Lingüística Aplicada of Pompeu Fabra University in Barcelona, Spain. The DAELE project aims to develop a prototype for an on-line learner´s dictionary for Spanish. The DAELE went online in January 2010 with 125 entries for verbs; work on verbs has proceeded faster than work on other lexical classes. We are working with a list of some 7000 high frequency nouns and adjectives covering a wide range of semantic fields as explained below (section 2.1), and use the TshwaneLex dictionary-writing system. Our work is based on data from several different corpora, and in the case of the Spanish web corpus, we use the Sketch Engine® to help us with the analysis of corpus data. We are attempting to organize senses of all lexical classes around one or more core senses for a given word, each of which may have more or more derived senses, along the lines of the approach taken in the *New Oxford Dictionary of English* (1998) and advocated by Lara (1998).

An important feature for us to take into account is that the DAELE is not being funded by a publishing house or an official language academy; rather, our work on the dictionary is a part of research projects on Spanish, funded primarily by the Spanish Ministry of Science and Innovation, the Fundación Comillas, and to a lesser extent, the Generalitat de Catalunya through its programme to support doctoral students who, in turn, are working on the dictionary. This circumstance has several important consequences for the DAELE. First, the fact that our funding is limited and is directly tied to a research programme means that the human resources we have available are very limited. Work on the DAELE cannot be set up in the same way as it would be in a true business context, in which presumably there is a prior feasibility study to ensure conclusion and publication of a completely finished product. In an institutional setting like ours, the fact that doctoral students must write their dissertations in 3 years means that they work on the dictionary part-time for a relatively short period, and the faculty members involved do not have release time from their teaching assignments. With these constraints, it is important for us to work with a representative sample of headwords, so that we can show what should be done for a dictionary of this type for Spanish, even though we ourselves may not be able to produce a complete dictionary to compare with learner's dictionaries of other languages.

3. Obtaining a representative sample

The list of headwords that are nouns or adjectives needed to include both most semantic types known by the average native speaker with a high-school education as well as all common morphological patterns for gender and number.

3.1 List of headwords

Our list of 7069 nouns and adjectives was obtained in the following way. We initially considered using the frequency list found in Davies (2006), which includes the 5000 highest frequency words in the *Corpus del Español*. After a cursory analysis of the words on this list, however, we determined that the list did not contain nouns from several semantic fields that we thought should be included in our prototype. We therefore decided to cross the nouns and adjectives from this list with those from three other sources. The three other sources we used were: the *Corpus PAAU 1992*, which contains vocabulary used in 700 college entrance exams from 1992; the corpus study *Léxico Disponible de los Estudiantes Preuniversitarios de la Provincia de Jaén*, and the word list from the *Diccionario de Primaria de la Lengua Española Anaya-Vox* (2000), which covers the core vocabulary for Spanish. We included nouns and adjectives, and not only nouns, because in Spanish many adjectives frequently occur as nouns (e.g. *amigo/amiga* 'friend'; *director/directora* 'director'; *claro/clara* 'clear', 'clearing' (*claro*, noun), 'egg white' (*clara*, noun)) and nouns and adjectives share many morphological properties, such as gender markings, plural markings, and diminutive and augmentative formation.

We expected there would be considerable overlap across these four sources, i.e., crossing the lists obtained from these four sources would provide us with a large number of nouns that would constitute a representative inventory of nouns. Our results, however, were quite different from expected. As shown in Figure 1, only 836 lemmata of a total 16,176 were found in all four corpora, and 9107 lemmata were found in only one corpus.

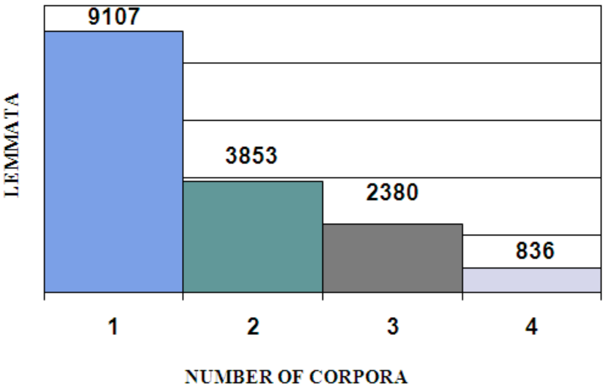


Figure 1: Total number of nouns coinciding across sources.

Clearly, 836 nouns were too few, so we decided to include lemmata that were present in at least two lists, yielding 7069 nouns and adjectives, which we believe is a large enough sample for our purposes. Of these lemmata, 5454 were classified as nouns; however, as mentioned above, this number is approximate because many lemmata classified as adjectives can also be used as nouns. It is also important to note that this headword list is not entirely closed, as we can incorporate other nouns as work proceeds.

3.2 Semantic classification

In order to identify a classificatory system that would be appropriate for descriptors and definition patterns in the DAELE, we considered both existing ontologies such as EuroWordNet and lists of superordinates that we compiled from existing dictionaries. We also took the subject labels from the work on the available lexicon (Ahumada, 2006) into account. In the end, we are creating our own system as our work proceeds, because no single classificatory system seemed appropriate for our target users. The system we are using identifies a general semantic group and then allows a maximum of three successive subgroups. For example, the word *sala* 'room' is classified most generally as a place ('*lugar'*), then as a building ('*construcción'*), then as a housing unit ('*vivienda'*) and finally, and most specifically, as a room ('*habitación'*). We expect that our classification will allow us both to establish semantic relationships between lemmata and to develop a system of more precise semantic features that is useful for definitions.

4. Corpus analysis

4.1. Role of corpus analysis in the DAELE project

The DAELE is a learner's dictionary that, as opposed to most dictionaries of Spanish, is corpus-driven. Although corpus-based lexicography is widespread in many language contexts, this is not the case for Spanish in general or for learner's dictionaries of Spanish in particular. More traditional lexicographical methods are still commonplace, although corpora are consulted. We might note that a widely available learner's dictionary of Spanish, the *Diccionario Salamanca de la lengua española* (1996), is not corpus-based. We can only agree with Atkins and Rundell (2008: 53) when they state that the advantage of using a corpus in lexicography is increased reliability of the information being included in the dictionary.

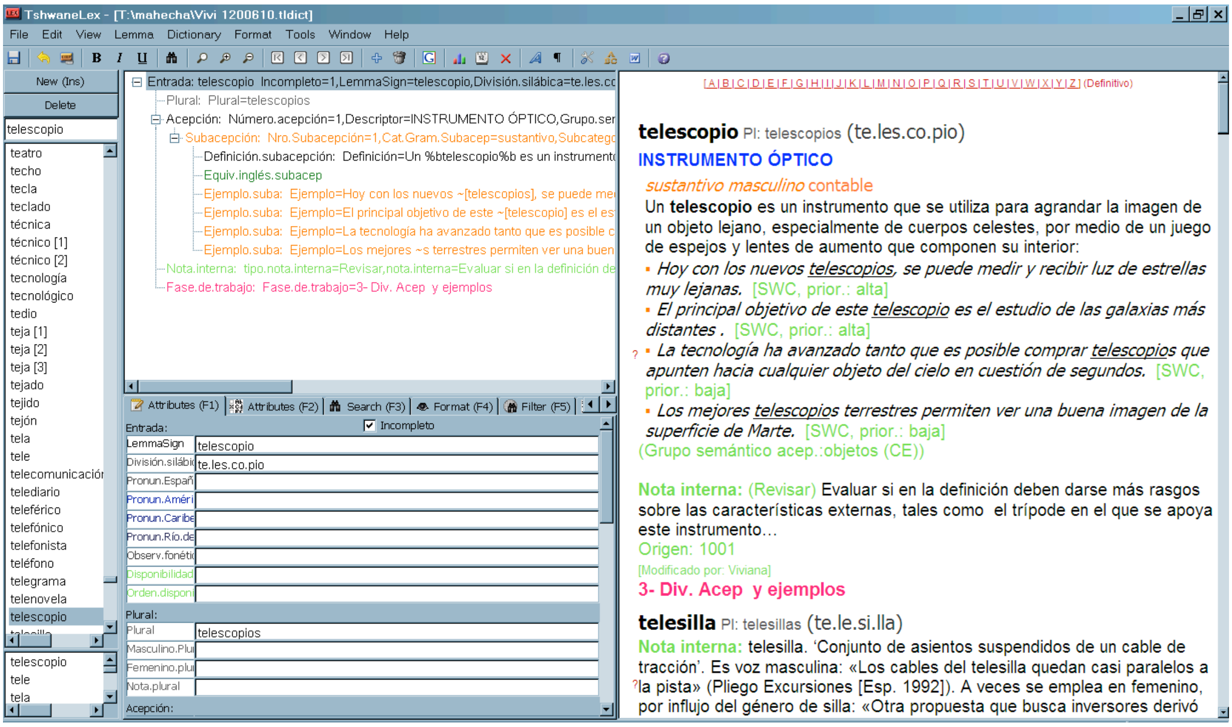
We are consulting three corpora, the CREA (*Corpus de Referencia del Español Actual*) of the *Real Academia Española*, the *Corpus del Español* compiled by Mark Davies and the Spanish Web Corpus that has been loaded into the Sketch Engine®. In practice, we study concordances from the corpora to identify the most frequent senses, the principal syntactic patterns associated with a particular sense, and any pragmatic information that might be included in the dictionary entry. In addition, the corpora provide us with examples, some of which we alter slightly to ensure they are maximally informative to non-native speakers of Spanish.

It is important to note that the extensive use of on-line sources allows the lexicographer to record large amounts of information for any specific headword, and that, obviously, not all the information recorded on

TshwaneLex platform needs to be made available to the end user.

4.2 Use of the TshwaneLex dictionary-writing system

We have loaded the 7069 nouns and adjectives into the TshwaneLex platform, in which we have defined fields for the various elements that will appear in the dictionary entry (e.g., headword, sense, subsense, example, usage note, note on syntactic structure, etc.). As seen in Figure 2, this dictionary-writing system allows the lexicographer to see several fields at once. In Figure 2, you can see the headword list to the left, the hierarchical structure of senses for the word *telescopio* 'telescope', the plural form, syllable division, a note for revision purposes, as well as the word's definition and some examples.


TshwaneLex - [T:\mahecha\Vivi 1200610.tldict]
File Edit View Lemma Dictionary Format Tools Window Help
New (Ins)
Delete
Attributes (F1)
Attributes (F2)
Search (F3)
Format (F4)
Filter (F5)

telescopio

teatro
techo
tecla
teclado
técnica
técnico [1]
técnico [2]
tecnología
tecnológico
tedio
teja [1]
teja [2]
teja [3]
tejado
tejido
tejón
tela
tele
telecomunicación
telediario
teleférico
telefónico
telefonista
teléfono
telegrama
telenovela
telescopio

telescopio
tele
tela

Entrada: telescopio Incompleto=1,LemmaSign=telescopio,División.silábica=te.les.cc
Plural: Plural=telescopios
Acepción: Número.acepción=1,Descriptor=INSTRUMENTO ÓPTICO,Grupo.ser
Subacepción: Nro.Subacepción=1,Cat.Gram.Subacep=sustantivo,Subcatego
Definición.subacepción: Definición=Un %btelescopio%b es un instrumento
Equiv.inglés.subacep
Ejemplo.suba: Ejemplo=Hoy con los nuevos ~[telescopios], se puede me
Ejemplo.suba: Ejemplo=El principal objetivo de este ~[telescopio] es el es
Ejemplo.suba: Ejemplo=La tecnología ha avanzado tanto que es posible c
Ejemplo.suba: Ejemplo=Los mejores ~s terrestres permiten ver una buen
Nota.interna: tipo.nota.interna=Revisar,nota.interna=Evaluar si en la definición de
Fase.de.trabajo: Fase.de.trabajo=3- Div. Acep y ejemplos

Entrada: ☑ Incompleto

LemmaSign	telescopio
División.silábi	te.les.co.pio
Pronun.Españ	
Pronun.Améri	
Pronun.Caribe	
Pronun.Río.de	
Observ.fonéti	
Disponibilidad	
Orden.disponi	

Plural:

Plural	telescopios
Masculino.Plu	
Femenino.plu	
Nota.plural	

Acepción:

[A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z] (Definitivo)

telescopio Pl: telescopios (te.les.co.pio)

INSTRUMENTO ÓPTICO

sustantivo masculino contable

Un **telescopio** es un instrumento que se utiliza para agrandar la imagen de un objeto lejano, especialmente de cuerpos celestes, por medio de un juego de espejos y lentes de aumento que componen su interior:

- *Hoy con los nuevos telescopios, se puede medir y recibir luz de estrellas muy lejanas.* [SWC, prior.: alta]
- *El principal objetivo de este telescopio es el estudio de las galaxias más distantes .* [SWC, prior.: alta]
- ? *La tecnología ha avanzado tanto que es posible comprar telescopios que apunten hacia cualquier objeto del cielo en cuestión de segundos.* [SWC, prior.: baja]
- *Los mejores telescopios terrestres permiten ver una buena imagen de la superficie de Marte.* [SWC, prior.: baja]

(Grupo semántico acep.:objetos (CE))

Nota interna: (Revisar) Evaluar si en la definición deben darse más rasgos sobre las características externas, tales como el trípode en el que se apoya este instrumento…
Origen: 1001
[Modificado por: Viviana]
3- Div. Acep y ejemplos

telesilla Pl: telesillas (te.le.si.lla)
Nota interna: telesilla. 'Conjunto de asientos suspendidos de un cable de tracción'. Es voz masculina: «Los cables del telesilla quedan casi paralelos a ?la pista» (Pliego Excursiones [Esp. 1992]). A veces se emplea en femenino, por influjo del género de silla: «Otra propuesta que busca inversores derivó

Figure 2: Screen in TshwaneLex for *telescopio*.

A nice feature of this platform is that is allows the lexicographer to see the headword list and the microstructure information corresponding to each entry at the same time. In our academic context, it is particularly helpful that this platform is easily adapted to a web-interface.

5. Microstructure of entries in the DAELE

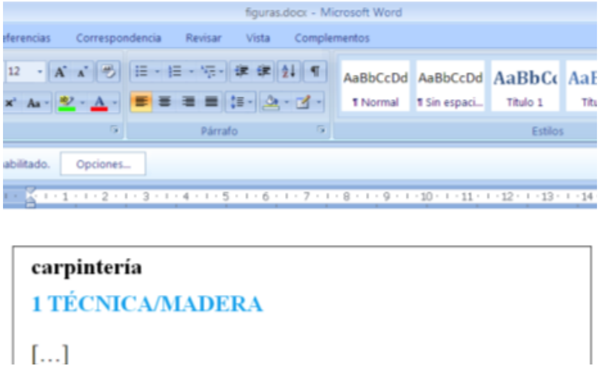
Since we are in the initial stages of writing definitions for nouns, in this paper we will concentrate on the following characteristics of noun entries: semantic labels, order of senses, defining style and examples.

5.1 Semantic labels

The DAELE makes extensive use of semantic labels to guide users through the various senses of polysemous words. There are several types of semantic label: sometimes the label is a very brief definition, sometimes it is a superordinate, other times it is a synonym.[1] The purpose of this element in the dictionary's microstructure is twofold: on the one hand, it allows non-native speakers to quickly and easily identify various senses, and, on the other, it allows us to apply a hierarchical order to senses, progressing from the core sense to derived senses, as will be discussed in section 5.2.

Many other on-line dictionaries, such as the *Macmillan English Dictionary* (MEDO), the *Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde* (DAFLES) or the *Cambridge Advanced Learner's Dictionary* (CALD), include this sort of semantic label, which appears as part of a menu with hyperlinks to entries. Our project follows suit and in the DAELE, the semantic labels are highlighted in blue and, depending on the settings the user has identified while consulting the dictionary, may appear on the screen without any further information. Figure 3 shows an example, for the word *carpintería* 'carpintery, carpenter's workshop', in which the labels may be translated as 'technique/wood', 'place', and 'wooden object or structure'.

[1] For more information about the type of semantic labels being used in the DAELE, see Estremera (2008, for nouns and adjectives) and Battaner (2010, for verbs).



carpintería

1 TÉCNICA/MADERA

[…]

Figure 3: Labels for *carpintería*.

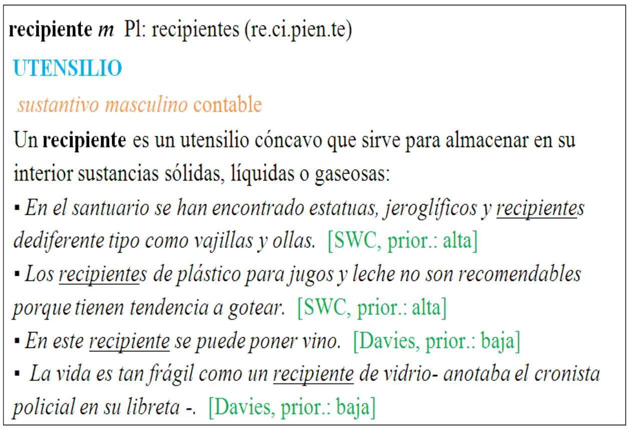
5.2 Order of senses

Establishing the criteria for the order of senses is one of the most important decisions in the DAELE, in that most of the lexical units in this project are polysemous. In preliminary work on the DAELE, DeCesaris & Bernal (2006) noted that although ordering senses of words occurring in several lexical categories according to grammatical criteria (i.e., all senses of one lexical category precede all senses of another lexical category, even though meanings might be clearly related) is a well-established practice in many lexicographic traditions, it often clashes with the notion of first defining the core meaning of a sense and subsequently derived senses. We have decided to organize senses according to semantic criteria, both in those cases in which the lemma is used in one lexical category (in our case, as either a noun or an adjective) and in those cases in which the lemma is used in two lexical categories (as both a noun and an adjective). We are convinced that this approach, in conjunction with the use of semantic labels, should help upper-intermediate students to grasp meanings better, since the information in the dictionary entry is ideally arranged in a hierarchy that establishes semantic relations between senses. As can be seen above in Figure 2 in relation to the word *carpintería*, senses are ordered in a hierarchical fashion: what we have analyzed as the core sense, that of a type of professional activity related to wood, is listed first, second comes the place where that professional activity takes place (*carpintería* in the sense of 'carpenter's workshop'), and the final sense is restricted to objects that are the result of that activity (*carpintería* in the sense of 'woodwork' or 'set of window and door frames'). We note that increased visualization of the semantic relations that hold between senses is an important difference between e-dictionaries and print dictionaries; this difference needs to be exploited, particularly in a dictionary for learners.

5.3 Definitions

We are interested in presenting noun senses in such a way so as to add to the user's ability to insert the noun into discourse. We have therefore decided to use full sentence definitions, in the spirit of Sinclair (1987). As Lew (2011) aptly notes, full sentence or 'popular style' definitions are generally more helpful to learners than the more traditional, formulaic definitions found in most print dictionaries, and the space limitations that argued against them are no longer applicable in the context of e-lexicography. This approach allows us, for example, to show the noun accompanied by an article in the definition, which identifies the noun sense as being either count or mass, in addition to reinforcing information about the noun's gender.

As shown in Figure 4, the definition of the word *recipiente* 'container' includes use of the indefinite article *un* to show that this noun is a count noun, the superordinate *utensilio* 'utensil', and information concerning its physical characteristics (a *recipiente* must be able to hold something in its interior; the definition says it is *concave* 'concave') and its function ('used to store solid, liquid or gaseous substances'). The noun being defined appears in boldface, and the definition is followed by four example sentences.



recipiente *m* Pl: recipientes (re.ci.pien.te)

UTENSILIO

sustantivo masculino contable

Un **recipiente** es un utensilio cóncavo que sirve para almacenar en su interior sustancias sólidas, líquidas o gaseosas:

- *En el santuario se han encontrado estatuas, jeroglíficos y recipientes dediferente tipo como vajillas y ollas.* [SWC, prior.: alta]
- *Los recipientes de plástico para jugos y leche no son recomendables porque tienen tendencia a gotear.* [SWC, prior.: alta]
- *En este recipiente se puede poner vino.* [Davies, prior.: baja]
- *La vida es tan frágil como un recipiente de vidrio- anotaba el cronista policial en su libreta -.* [Davies, prior.: baja]

Figure 4: Sample definition and examples for *recipiente*.

5.4 Examples

As stated in section 2, the information presented on nouns in the DAELE is based on data from several corpora, and the examples included are also taken from corpora.[2] Examples are chosen to complement the information provided by the definition, and show typical syntactic and pragmatic patterns associated with the sense.

In the DAELE we have decided to include a maximum of four examples per sense and subsense. Examples that have been chosen are classified by the lexicographer as being either 'high priority' or 'low priority' and this information is recorded in TshwaneLex (see the information in green typeface in Figure 2). On-line consultation of the DAELE allows the user to choose between full view and reduced view, although at the time of publication of this paper, the full/reduced view option for verbs does not respond to seeing more or

[2] In some cases, examples are slightly modified (e.g. spelling is adapted to conform to the current standard orthography, proper nouns are deleted, and abbreviations are avoided).

fewer examples but rather to seeing only the semantic label for each sense or the full entry for the sense.[3] It is not clear to us that this approach is the best for nouns, for which it may be preferable to break down the (current) 'full' view into parts. We have not come to a decision on this point, as we have deferred out decision until we have a larger body of entries completed.

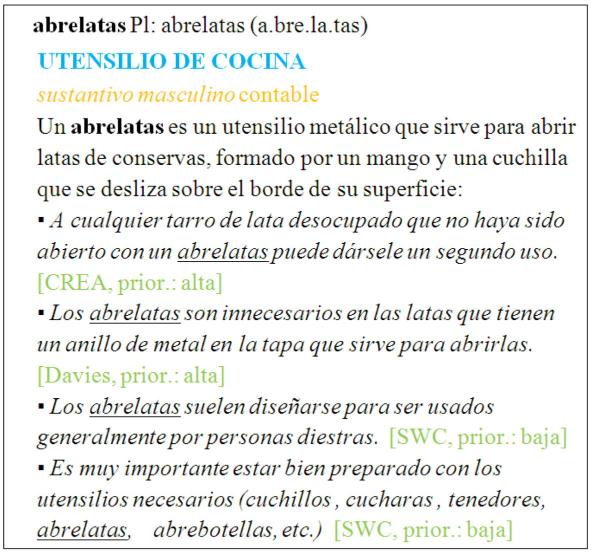
To illustrate the criteria we are using to choose examples, we again turn to the entry for *recipiente*, shown in Figure 4. Examples are differentiated from definitions by the use of italics, with the definiendum underlined. In the first example, *recipiente* is shown to be a superordinate of *vajillas* 'set of dishes' and *ollas* 'pots'. The second and third examples include substances that are typically found in *recipientes* (*vino* 'wine', *jugo* 'juice', and *leche* 'milk'). The second and fourth examples show types of material that *recipientes* are often made of (*plástico* 'plastic' and *vidrio* 'glass'). The examples have been chosen to ensure that both singular and plural forms are included and to show the noun with different determiners (*los*, the plural definite article, *este*, a singular demonstrative adjective, and *un*, the singular indefinite article).

5.5 Other information in the entry

In addition to semantic information, noun and adjectives entries in the DAELE contain other types of information that are essential in terms of grammar and which make the dictionary different from existing dictionaries of Spanish. Entries contain information on syllabification, lexical category, pluralization and grammatical gender (if applicable), and a label indicating count noun, mass noun, or both[4]. We note that including plural forms is a departure from the practice of most Spanish dictionaries, in which plural forms are not included and in which usually only partial information about gender marking is included. Clearly, in an on-line dictionary, the space-saving representation of grammatical gender that only includes the final syllable of the word (e.g. *amigo*, *-ga*) is unwarranted, and we feel that providing users with the plural form reinforces their knowledge of the word.

In Figure 5, for example, the plural for *abrelatas* 'can opener' is given; we note that the plural form is identical to the singular form, which is common for verb-noun compounds of this structure in Spanish, although in the language as a whole, it is rare for nouns to have the same form in the singular and in the plural.

Following the practice of learner's dictionaries such as the MEDO, CALD, *Oxford Advanced Learner's Dictionary* (OALD) or *Longman Dictionary of Contemporary English* (LDCE), among others, the lexical category is spelled out (note the word *sustantivo* 'noun' in Figures 2, 4, and 5), as opposed to including abbreviations. For words with more than one sense, lexical category, grammatical gender and count/noun are indicated for each sense.



abrelatas Pl: abrelatas (a.bre.la.tas)

UTENSILIO DE COCINA

sustantivo masculino contable

Un **abrelatas** es un utensilio metálico que sirve para abrir latas de conservas, formado por un mango y una cuchilla que se desliza sobre el borde de su superficie:

▪ *A cualquier tarro de lata desocupado que no haya sido abierto con un abrelatas puede dársele un segundo uso.* [CREA, prior.: alta]

▪ *Los abrelatas son innecesarios en las latas que tienen un anillo de metal en la tapa que sirve para abrirlas.* [Davies, prior.: alta]

▪ *Los abrelatas suelen diseñarse para ser usados generalmente por personas diestras.* [SWC, prior.: baja]

▪ *Es muy importante estar bien preparado con los utensilios necesarios (cuchillos, cucharas, tenedores, abrelatas, abrebotellas, etc.)* [SWC, prior.: baja]

Figure 5: Sample definition and examples for *abrelatas*.

Given that our work on nouns in the DAELE is still in progress, certain aspects of the microstructure may be revised in the future; specifically, decisions need to be taken on the role of phraseology and on what word relationships we wish to show via hyperlinks.

6. Conclusion

Work on representing nouns in the DAELE attempts to incorporate the advantages of e-dictionaries while providing learners with information that to-date has been absent from most dictionaries of Spanish. We believe that the microstructure of entries in the DAELE allows for quick, easy access to information, and provides learners with several examples of real use.

Several interesting questions have arisen in our work that need further attention. We will draw attention here to only one, namely the nature of the semantic labels. In Section 5.1, we noted that labels for nouns are of three types: superordinates, synonyms, or brief definitions. It is not clear to us at this point why one of these types is better suited to a particular set of circumstances than the others; in other words, we would like to be able to describe the conditions that should obtain for the label to be of a certain type. The study of the role of semantic labels, which are commonplace in e-dictionaries, is one of our research goals for the immediate future.

7. Acknowledgements

Work on this paper was supported by grant FFI2009-07588 "*Agrupacion Semántica y Relaciones Lexicológicas en el Diccionario*" from the Ministry of Science and Innovation of Spain (DeCesaris) and a 2009

[3] Users can also choose between seeing the full verb conjugation or not.

[4] For more information on representing the count/mass distinction in dictionaries of Spanish, see DeCesaris, Battaner & Bernal (2004) and Bernal (2010).

FI grant from the Catalan government's Agency for Management of University and Research Grants (AGAUR) (Mahecha), which we gratefully acknowledge.

8. References

8.1 Dictionaries

[CALD] *Cambridge Advanced Learner's Dictionary*. Accessed at: http://dictionary.cambridge.org/.

[COBUILD] *Collins Cobuild English Dictionary*. (1987) ed. J. Sinclair, London: Collins.

[DAFLES] *Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde*. Accessed at: http://ilt.kuleuven.be/blf/.

Davies, M. (2006). *A frequency dictionary of Spanish*. New York: Routledge.

[DPVOX] *Diccionario de Primaria de la Lengua española Anaya-Vox*. (2000) Barcelona: Vox-Biblograf.

[DSAL] *Diccionario Salamanca de la Lengua Española*. (1996) Madrid: Santillana.

[LDCE] *Longman Dictionary of Contemporary English*. Accessed at: http://www.ldoceonline.com/search/.

[MEDO] *Macmillan English Dictionary Online*. Accessed at: http://www.macmillandictionary.com/.

[NODE] *New Oxford Dictionary of English*. (1998). Oxford: Oxford University Press.

[OALD*] Oxford Advanced Learner's Dictionary*. Accessed at: http://www.oup.com/elt/catalogue/teachersites/oald7/lookup?cc=global.

8.2 Corpora

Corpus PAAU 92. Accessed at: http://www.iula.upf.edu/rec/corpus92/.

Spanish Web Corpus. Accessed at: http://www.sketchengine.co.uk.

Davies, M. Corpus del Español. Accessed at: http://www.corpusdelespanol.org.

REAL ACADEMIA ESPAÑOLA. Corpus de Referencia del Español Actual (CREA). Accessed at: http://corpus.rae.es/creanet.html.

8.3 Other

Águila, G. (2006). Las nuevas tecnologías al servicio de la lexicografía: los diccionarios electrónicos. *Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística.* León: Universidad de León. Accessed at: http://www3.unileon.es/dp/dfh/SEL/actas.htm.

Ahumada, I. (2006). *El léxico disponible de los estudiantes preuniversitarios de la provincia de Jaén*. Jaén: Servicio de Publicaciones de la Universidad de Jaén.

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

Battaner, P. (2010). El uso de etiquetas semánticas en los artículos lexicográficos de verbos en el DAELE. In *Quaderns de filologia. Estudis lingüístics 15*. València: Facultat de Filologia. Universitat de València. pp. 139-158.

Battaner, P., Renau, I. (forthcoming). El proyecto DAELE verbos: un diccionario de aprendizaje de español como lengua extranjera. *Encuentros ELE Comillas*, Comillas (Santander)*, 2010*.

Bernal, E. (2010). La gramàtica dels substantius de matèria en els diccionaris. In *Estudis de lexicografia 2003-2005*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Documenta Universitaria, pp. 61-74.

Chen, Y. (2010). Dictionary Use and EFL Learning. A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries. *International Journal of Lexicography*, 23(3), pp. 275-306.

DeCesaris, J., Bernal, E. (2006). Consideraciones previas a la representación de las formas nominales en el Diccionario de aprendizaje del español como lengua extranjera (DAELE). In *Palabra por palabra: estudios ofrecidos a Paz Battaner*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Documenta Universitaria, pp 83-92.

DeCesaris, J., Battaner, P. & Bernal, E. (2004). Representing Noun Grammar in English and Spanish Dictionaries. In *Proceedings of the 11th EURALEX International Congress*. Lorient: Euralex, Université de Bretagne-Sud, pp. 835-846.

De Schryver, G.-M. (2003) Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2), pp. 143-199.

Dziemianko, A. (2010). Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography*, 23(3), pp. 257-273.

Estremera, E. (2008). Etiqueta semántica (Proyecto nombres y adjetivos). Working paper from research project HUM2006-07898/FILO.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of 11th EURALEX International Congress*. Lorient: Euralex, Université de Bretagne-Sud, pp. 105-116.

Lew, R. (2011). Space restrictions in paper and electronic dictionaries and their implications for the design of production dictionaries. In P. Bański, B. Wójtowicz (eds.) *Issues in Modern Lexicography*. München: Lincom Europa.

Lara, L. (1998). Una hipótesis cognitiva sobre el orden de las acepciones. *Boletín de Filología. Universidad de Chile.* XXXVII, pp. 626-644.

Mechura, M.B. (2008). Giving Them What They Want: Search Strategies for Electronic Dictionaries. In *Proceedings of the XIII EURALEX International Congress Barcelona, 15-19 July 2008*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Documenta Universitaria.

Sinclair, J. (ed.) (1987). *Looking up: an account of the*

COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London: Collins.

Trap-Jensen, L. (2010). One, two, many: Customization and User Profiles in Internet Dictionaries. In *Proceedings of the XIV Euralex International Congress Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy.

Blog on Socio-Political Vocabulary as a New Reference Tool

Svetlana Manik
Ivanovo State University
Ivanovo, Russian Federation
E-mail: Svetlana_manik@yahoo.com

Abstract

The purpose of this paper is to define the socio-political vocabulary, its peculiarities and lexicographic ways of its registration. On the basis of the previous lexicographic analysis of Russian and foreign dictionaries registering words of socio-political sphere (of over 40 reference books, monolingual (Russian and English) and bilingual (English-Russian) the general tendencies and traditions of such description are stated. The Russian user's survey proved the demand in bilingual dictionary of socio-political vocabulary with the Russian students learning English. The blog on such lexis will meet their requirements and expectations as it will provide the authentic texts in English and Russian, the vocabulary on the specified issue with detailed description (translation equivalent with additional comment on its semantics; cultural note; illustrative examples; collocations), message option and forum opportunities. The ideas are illustrated by examples. The blog is easily accessible by the students and involves them in the process of creation and up-dating.

Keywords: Socio-Political Vocabulary; Blog as a reference tool; bilingual dictionary

1. Introduction: Definition of Socio-Political Vocabulary

Socio-political vocabulary (SPV) is rather diverse and covers numerous spheres of both everyday and professional life. It has been in the focus of our research for over ten years already. In our thesis the following definition of SPV was worked out: lexical units denoting events/facts/developments/phenomena of socio-political life which represent peculiarities of social society and are not be regarded, interpreted and evaluated by the members of different societies identically.

There are some features important for understanding the SPV:

- it is a culturally and ideologically meaningful lexicon; the notions *gunman* and *боевик* used in the same news context have different meanings and connotations in English- and Russian speaking societies;
- the ideology issue is important; it is understood beyond the political concept as a science/unity of ideas in a general sense and as a framework that is "assumed to specifically organize and monitor one form of socially shared mental representation, in other words, the organized evaluative beliefs — traditionally called 'attitudes' — shared by social groups" (van Dijk, 1996);
- evaluation aspect is essential since the language of politics, as W. Safire truly states in the prolegomenon of the updated and expanded edition of *Safire's Political Dictionary* (2008), "*is a lexicon of conflict and drama, of fulsome praise and fierce ridicule, of emotional pleading and intellectual persuasion. Color and bite permeate a language designed to rally support, to blast opposition, and to mold the minds of multitudes*";
- peculiarities of valued words are stipulated mostly by their functioning and lexical environment and not solely by the structure of the semantic meaning; socio-political lexis are very dynamic and much dependant on the lexical and grammatical environment;
- it denote numerous everyday and professional communication spheres and is wider than LSP of politics; some examples of the vocabulary: *Federal Security Service; aerial bombardment; bomb blast; breakaway republic; collapse; bailout; tax evasion; stimulus plan marginalised; E-coli outbreak;* etc.
- it is widely and commonly used in mass media.

2. Main Tendencies of Russian and Foreign Lexicographic Schools in Registering SPV

Over forty reliable Russian and English reference books registering words of political and social life, have been analyzed and the main tendencies of Russian and foreign lexicographic schools have been distinguished (Manik, 2010; 2011).

American and British dictionaries (mostly monolingual learner's dictionaries designed for college or University students and teachers of Political Science or for people interested in politics) rather serve as a guide to the rich technical language of political science and to the actual operations of the American/British political systems. Their great advantage seems to be the double function feature: to be a dictionary and a text book. On the one hand, terms are listed alphabetically or sometimes by subject matter and a detailed descriptive definition is provided. There is minimum or sometimes zero lexicographic labels (both linguistic and functional). On the other hand, each definition is usually followed by some additional information, like a statement of its historical and/or contemporary significance to American/British government and the citizen. There also may be some sections on important agencies, cases and statutes, reference to the political document, etc. The registered vocabulary is depicted through historical, cultural and social events and it reflects hi-tech development and new scientific discoveries. The vocabulary includes only terms of political sphere (government structure, agencies, functions; elections; international activity and negotiations; etc.), with a great number of metaphoric expressions, elliptic constructions, allusions from the politicians' speeches, debates.

Russian dictionaries registering LSP of politics are different types of dictionaries, mostly bilingual, designed for students, interpreters, journalists and other people concerned with politics and political science. There exist reference books providing: narrative and detailed description of the notions in Russian with English equivalent in brackets, on the one hand; both linguistic information (grammar, collocations, functional label, etc.) and extra-linguistic data, on the other hand. The material covers different subject areas (social, economic, political, etc. life), thus the corpus is very rich in terminological units.

According to the Russian user's needs and expectations questionnaire poll there are certain information categories that they want to get in a dictionary (cultural note, collocations) and there's no bilingual dictionary meeting all the requirements.

It should be noted that the LSP of politics dictionaries contribute greatly to the user's understanding of the political life, the governmental structure, ways of governing the country, the powerbrokers' speeches (both explicit and implicit content), the news and Mass Media reports in general. That's why recently there've come out many political Russian dictionaries concerned with the explanation of the leading party policy, being pro-Putin editions, and opposing and criticizing the present head of the country and the ruling party (Manik, 2011). The detailed review and analysis was reported in the 18th European Symposium on Language for Special Purposes, 2011.

3. LSP of Socio-Political Life in Educational Process: Idea of the Blog on SPV

It's a well-known fact that the command of the language is essential for effective communication and for carrying on useful analysis of basic problems. The latter is extremely important with the young generation to have a successful on-going community. And it may explain the learning-oriented approach of the lexicographers while compiling the LSP of politics dictionary.

LSP of politics dictionaries and LSP of socio-political life dictionaries are in the focus of the research. The latter register a wider vocabulary, add to political terms the business lexis, words of anti-terrorist movements, of ecology, medicine, social policy, etc. That's why they are of great help in the academic process when students are taught to follow the mass media reviews.

SPV is a significant part of the courses that the Russian students take up at the University because the Russian society is rather politically minded. Many people study English and follow both Russian and English-speaking news reviews, socialize and debate them in Twitter, Facebook and other social networking sites. But they face many problems in understanding, translating and then discussing the news as, on the one hand, there are a lot of terminological words entering the news, and, on the other hand, there's cultural and ideological difference in the developments interpretation.

Since there's a problem with the reference book (English-Russian, Russian-English) to be recommended to the Russian students for the academic process there appeared an idea to create a blog being a new reference tool for Russian students studying English, in particular in practical classes of English in the English Philology Department, Ivanovo State University.

Last year there was an experimenting use of the class blog because it was believed that it would be an efficient way to communicate class expectations and to engage students in conversation with teachers/professors, with their fellow classmates, and with outside resources. The advantages of the blog allowed us to insist on, and track, 100 percent participation in course discussions, to provide links to timely articles and podcasts, and to invite experts into the conversation, all to create an environment where students not only experience the subject matter but also learn how to find credible sources for learning in the future.

The blog discussed is to be a textbook and a reference tool simultaneously. The students do need up-dated news reviews on particular topic/problem/event in Russian and English and also some bilingual glossary with the detailed comments on the meaning, usage, cultural notes and illustrations.

Traditionally the students are advised to work with the BBC News, CNN News and various Russian news agencies as they are expected to make a news review or a brief presentation on the breaking news or (inter)national development demonstrating their understanding and critical thinking abilities. The typical problems with the students are:

- lack of the vocabulary on the topic;
- misunderstanding or ignorance on the issues discussed (especially specific questions);
- consultation with only one source (either Russian or foreign).

The aim of the blog as a reference resource is: to facilitate learners in translating mass media reports from English into Russian and in understanding cultural and ideological differences of the societies; to provide users with necessary SPV, its translation equivalents and required extra-linguistic information; to demonstrate illustration examples on the usage of SPV.

4. Structure of the Blog on SPV

The blog is structured according to the basic rules of social net blogs and theoretical rules of applied lexicography:

· Brief introduction indicating the aims, target users/readers, resources and explanation of the structure of every information block;

· Topic under discussion and some authentic texts in English and Russian, along with outlined words as essential vocabulary on the topic (there will be several topics like *the World Set Up; (Inter)National Organizations and Social Movements; Globalization and Integration Process; Social Problems of Today; Terrorism; Election Campaign; etc.* corresponding the currents developments in the world*)*. The texts/abstracts are selected according to the informational and vocabulary density. They are short and provide either the facts and statistics or the personal opinion of a politician/expert/powerbroker. Since they are edited, there is a reference on the original version for the students to be able to consult with it if necessary;

For example, there's a following text on terrorism:

> Police probing a ***deadly blast*** outside the Delhi High Court on Wednesday have released two sketches based on descriptions of suspects from eyewitnesses and they are focused on leads that an Islamic ***extremist group carried out the attack***.
>
> At least 11 people were killed Wednesday after a bomb inside a briefcase went off outside the Delhi High Court, India's home minister told lawmakers. The home minister's website said another 76 people were injured and officials fear ***the death toll*** will rise. It was the second explosion outside the court complex in four months.
>
> "This is ***a cowardly act*** and we will never succumb to the pressure of terrorists," Prime Minister Manmohan Singh said during a visit to Bangladesh. (New Delhi (CNN))

(http://edition.cnn.com/2011/WORLD/asiapcf/09/07/india.court.blast/index.html)

· Detailed entries of the words and word combination outlined in the text with the following information categories:

- additional comment on the lexical meaning of the key word within the definition/translation equivalent, in brackets, to explain the semantics of the notion and show the difference (if any) between the English and Russian concepts;
- cultural information in a separate paragraph as it turns out to be of vital importance in interpreting and understanding a word/word combination. It is essential in a bilingual dictionary as the translation stipulates the replacement of a source language word by a target word in the translation language and its culture. This definition is worth mentioning, especially within the context of the dictionary of the socio-political vocabulary. We mean the translation of the ideologically coloured lexis reflecting the ideas and attitude to this phenomenon of the given society. Besides, it can be called an element of the *background knowledge* or *языковая компетенция* and it contributes to the success/failure of the translation/communication;
- illustrative verbal example carrying out a big notion explanation through the context of the entrée usage;
- collocations: the words and word combinations close in the meaning and used in similar contexts.

This approach's been worked out and described in our previous works in regards to a model dictionary of Socio-Political Vocabulary (Manik, 2001; Manik, 2010).

For example, the entry on ***deadly blast***:

This word combination is not registered in the English-Russian dictionaries, but the word combination *bomb blast* is usually translated as *взрывная волна (blast wave, explosion wave)*, though there's an entry *blast (=explosion)*.

Deadly blast – смертоносный взрыв (используется вместо *bomb blast,* чтобы подчеркнуть масштабность и трагические последствия взрыва)

(killing/dead explosion (used instead of bomb blast to underline the dimensions and tragic consequences of the explosion)

Interesting (!) (=cultural data)

Это словосочетание используется в с вязи с различными террористическими актами, когда от взрыва заложенной в общественном месте бомбы погибает много людей. Наиболее широкое применение получило после событий в Москве летом 1999 г., тогда в результате взрыва в переходе погибло около 300 человек.

(*This word combination is used while talking about various terrorist act, when a lot of people are injured and killed by the explosion of a bomb in the public place. It entered the active and wide use after the Moscow summer events of 1999 when over 300 people died of the underground passageway explosion).*

Illustrative examples

A bomb blast has rocked the Indian capital, New Delhi, killing nine and injuring at least 65 outside the city's High Court.The Harkat-ul-Jihad terror group has taken responsibility for the attack (http://rt.com/news/new-delhi-blast-india-975/)

"Terrorist attack" condemned as *deadly blast* hits Ankara (http://www.euronews.net/2011/09/20/three-dead-in-ankara-explosion/)

Collocations

bomb explosion - *Взрыв бомбы.*

Deadly bomb blast - *смертоносный взрыв бомбы*

Deadly explosion – *смертельный взрыв*

Powerful blast – *сильный взрыв*
Car bomb blast – *взрыв бомбы, заложенной в машине*
Delhi bomb blast – *взрыв в Дели*
Taliban suicide blast – *взрыв от талиба террориста-смертника*
To plan bomb blast – *планировать взрыв бомбы.*
The blast kills a great number of people – *взрыв убивает огромное количество людей*

In our view such description is a crucial issue in bilingual socio-political terminology management. It allows to represent specialized concepts so as to provide the user with an adequate understanding of their meaning as well as sufficient knowledge of their cultural, ideological colouring and message, appropriate usage.

· Message option to provide personal attitude/addition to the entry described. The students are supposed to write and share their comments and opinions prior the lesson or after the discussion in class. It allows to involve the students in the academic process and in the work of a lexicographer, when (s)he learns how to describe a word, emphasize the usage notes, etc. Besides a teacher can be asked a question any time apart from the lesson.

· Forum section on the additional vocabulary on the given topic. There may come out interesting illustration and usage examples, or some new collocations, or other updating on the issue. Besides the blog format may attract the users from other educational establishments or some experts in the sphere.

The blog also possess some multimedia contents, which will be presented in the given way: an entry will have an integrated flash audio and video player. That feature allows users to watch and listen to a piece of speech, news, report, parliamentary debates, etc. connected with the topic under discussion.

5. Conclusion

The blog, used in the academic process, seems to be a new reference tool with a dictionary element. It is also important to note that it is just a tool and not the objective itself. Blogs are not for everyone or for all classes and need to be made an integral part of the course design. However, as noted by Trammell and Ferdig (2004), the use of blogs as a learning tool seems to be low-cost with high-returns. While more research needs to be done as to how blogs can more effectively be used, it is a given that technology will continue to influence learning

Our experiences in using classroom blogs have been overwhelmingly positive. While students' acceptance of technology in the classroom requires its perceived usefulness and ease of use, students do tend to learn best when they need information that they can put to use immediately. Blogs are an effective and efficient method of allowing students to access information as it is needed and to make connections between explicit knowledge from textbooks and tacit knowledge gained as students see how others can and are using the knowledge being shared. Blogs also introduce students to online learning communities so they can access and evaluate information, and construct new learning paradigms for themselves. Finally, effectively modeling ways to use blogs as a teaching, learning and reference tool is a useful skill for the students to have as they embark on their journey of life-long learning.

The work on compiling a real bilingual dictionary of SPV is in progress and it will take some more time. But even now the students may use the data collected and processed and take part in the lexicographic process themselves.

6. References

Averbukh, K.Y., Karpova, O.M. (2009). *Leksicheskije i phraseologitcheskije aspekti perevoda. (*Lexical and Phraseological Aspects of Translation). Moscow.

Carr, M. (1997). Internet Dictionaries And Lexicography. *International Journal of Lexicography*, 10(3), pp. 209-221.

Karpova, O. (2005). Russian Lexicography. *Oxford Encyclopedia of Language & Linguistics*. Oxford: Elsevier. Volume 10, pp. 704-715.

Karpova, O., Kartashkova, F. (eds.) (2007). *Essays on Lexicon, Lexicography, Terminography in Russian, American and Other Cultures*. Cambridge: Cambridge Scholars Publishing.

Karpova, O., Manik, S. (2000). Public Political Vocabulary: Model of a Dictionary. In *Lexicographica Series Major 109. Symposium on Lexicography X*. Copenhagen, pp. 173-184.

Manik, S. (2001). Socio-Political Lexis (Evaluation Aspect) in Dictionaries of Different Types. Thesis. Ivanovo.

Manik, S. (2010). Socio-Political Vocabulary: Description in Bilingual LSP Dictionary. In O. Karpova, F. Kartahkova (eds.) *New Trends in Lexicography: Ways of Registrating and Describing Lexis*. Cambridge: Cambridge Scholars Publishing, pp. 249-261.

Manik, S. (2011). LSP of Politics vs LSP of Socio-Political Sphere. In *The 18th European Symposium on Language for Special Purposes (LSP). Special Language and Innovation in Multilingual World*. Perm State University, 22-26 August 2011. Book of Abstracts, p. 52.

Safire, W. (2008). *Safire's Political Dictionary*. Oxford: Oxford University Press.

Sheigal, E.I. (2000). *Semiotika politicheskogo discursa.* (Semiotics of the Political Discourse). Moscow-Volgograd.

Trammell, K.D., Ferdig, R.E. (2004). Pedagogical implications of classroom blogging. *Academic Exchange Quarterly*, 8(4), pp. 60-64.

Van Dijk, T.A. (1996). Discourse, Opinions and Ideologies. In C. Schäffner, H. Kelly-Holmes (eds.) *Discourse and Ideologies*. Philadelphia: Multilingual Matters, pp. 7-37.

vernetziko: A Cross-Reference Management Tool for the Lexicographer's Workbench

Peter Meyer
Institut für Deutsche Sprache
Mannheim
E-mail: meyer@ids-mannheim.de

Abstract

vernetziko is an assistive software tool primarily designed for managing cross-references in XML-based electronic dictionaries. In its current form it has been developed as an integral part of the lexicographic editing environment for the German monolingual dictionary *elexiko* developed and compiled at the Institut für Deutsche Sprache, Mannheim. This paper first briefly outlines how *vernetziko* fits into the XML-based dictionary editing technology of *elexiko*. Then *vernetziko*'s core functionality and some of the auxiliary tools integrated into the program are presented from both a practical and a technological point of view. The concluding sections discuss some software engineering aspects of extending the tool to handle cross-references between multiple resources and point out some of the advantages of *vernetziko* vis-à-vis corresponding features of proprietary dictionary writing systems. The software can be adapted to interconnect off-the-shelf components (database management systems and editors), thus providing a tailor-made lexicographical workbench for a wide range of XML-based dictionaries without vendor lock-in.

Keywords: electronic dictionaries; dictionary editing software; cross-references; XML; Java

1. Introduction

The proper technical handling of cross-references within and between articles in electronic dictionaries poses several well-known problems, cf. (Joffe et al., 2003); amongst other things, the editing process must enforce and preserve the validity and consistency of cross-references as well as any required bidirectionality (symmetry) of relations such as synonymy. Many contemporary electronic dictionary systems use a semistructured markup data representation, usually based on XML (Lemnitzer et al., [to appear]), which requires specific solutions for cross-reference modeling (Müller-Spitzer, 2007; 2010a).

This paper presents and discusses the conceptual underpinnings of a modular approach to handling cross-reference structures in XML-based dictionaries. In its current form, this approach has been implemented for the German monolingual online dictionary *elexiko* which forms part of an ongoing research project of the Institut für Deutsche Sprache (Institute for the German Language) (Haß, 2005; Klosa, 2011). *elexiko* is accessible free of charge under *www.elexiko.de*. For expositional purposes, we will focus on the specific implementation chosen for *elexiko*; its overall architecture as outlined in this paper is, however, easily adaptable to other dictionary writing systems.

Section 2 is a brief survey of the overall structure of *elexiko* XML entries and the technical interplay of various components of the dictionary writing technology in *elexiko*. Section 3 focuses on the core functionality and some implementational aspects of *vernetziko*, an assistive software tool primarily designed for managing cross-references in electronic dictionaries. Section 4 presents an overview of further assistive management tools built into the program and gives some background on the database design chosen for *elexiko*. Section 5 discusses several software engineering issues that arise when extending the tool to handle cross-references between multiple heterogeneous lexicographic resources in a dictionary portal. The concluding section 6 briefly summarizes the specific advantages of the approach presented in this paper vis-à-vis monolithic dictionary writing systems with built-in reference management.

2. Background: *vernetziko* as a part of the lexicographer's software environment in *elexiko*

The lexicographic information contained in each *elexiko* entry is encoded in a single standalone XML document. A cross-reference element inside a 'source' element of one article relates to a 'target' element in the same or another *elexiko* article, usually by specifying special ID attributes of the target article and target element. In this way, cross-references are stored in a strictly local and non-redundant fashion. An important implication of this design is that cross-references assumed to be bidirectional (e.g., links between synonymous senses of two lexemes) are simply represented as two references in two separate XML articles.

Every XML document – i.e. dictionary entry – is stored in an XML-enabled Large Object (LOB) together with some metadata as a separate record (row) in an Oracle database table (Müller-Spitzer & Schneider, 2009). In order to edit an article, authors use a Content Management System (CMS) that retrieves the corresponding XML file from the database and writes the altered version back later. XML files are edited locally by lexicographers using an off-the-shelf XML editor. *vernetziko* is a Java 6 SE application that interacts with all three of the aforementioned components:

- It ‘remotely controls’ the XML editor via the editor’s API or plugin architecture; specifically, it can parse, analyze and modify the current document content.
- It has read-only access to the dictionary database via a standard JDBC interface.
- It interacts via HTTP with the CMS in order to check articles out and in. Strictly speaking, this third interdependency is not necessary; one could easily eliminate it by allowing *vernetziko* to update the database directly. This route has not been taken for the specific technical setup of *elexiko* in order to avoid duplicating code used in the CMS for authentication and data integrity verification, amongst other things.

Overall, a highly modular approach has been chosen for *vernetziko*, such that any of the three components enumerated above may easily be replaced by a different software component. On the implementation side this modularity is enforced by programming against Java interfaces that represent the functionality of the different components and abstract away from implementational details of database queries and calls to the XML editor’s API.

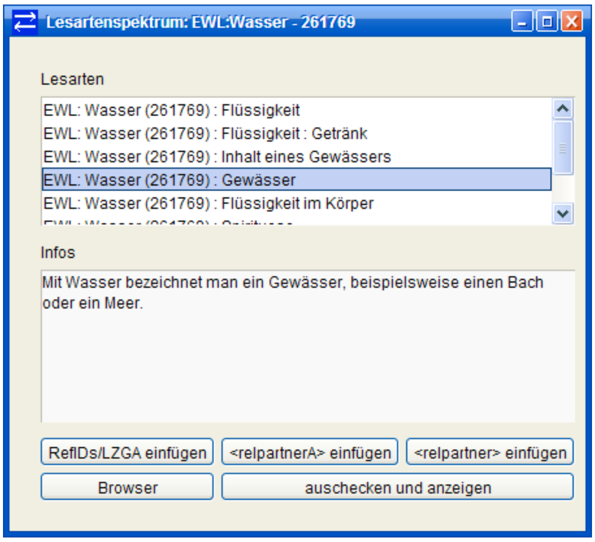
3. Core Functionality

3.1 Cross-reference handling in *vernetziko*

vernetziko has primarily been developed as a software tool for the automated insertion, correction and checking of cross-references in an extensible set of XML-based electronic dictionaries. Cross-references in an *elexiko* ‘source’ article document – typically more than 20 – relate an ‘address’, i.e. a specific XML element of this document, to another address that usually belongs to another entry, possibly in a different dictionary. In this manner, cross-references are stored in a strictly ‘local’ and non-redundant way.

Most of the functionality of *vernetziko* is designed to overcome practical issues with this pragmatic approach, particularly with regard to referential integrity:

- Manually checking the consistency and validity of all outgoing cross-references encoded in an *elexiko* article would require far too much effort. *vernetziko* cross-checks all references in the presently edited document with the database and computes appropriate status information, automatically updating its displays when the document is modified in the XML editor.
- As said above, the target of a cross-reference is specified using ID strings, viz. the values of id attributes of the targeted article and XML element. In some cases, two nested elements – for a sense and its targeted subsense – must be specified in this way. When a new cross-reference is created, manually inserting such ID values is clumsy and error-prone. With *vernetziko*, lexicographers only have to specify a lemma and then select one of its (sub-)senses from a list to let the program fill in or correct all missing details of the desired reference, cf. Fig. 1.
- Incoming cross-references for a given dictionary article can only be found through complex database queries. *vernetziko* automatically performs all necessary queries and then enumerates and checks the status of all existing incoming references for the presently edited document.
- Bidirectional cross-references (e.g., links between synonymous senses of two lexemes that are required to be symmetrical in *elexiko*) are represented as two independent references in two separate XML articles. *vernetziko* matches the lists of outgoing and incoming references for the presently edited article in order to determine whether obligatory bidirectionality is already accounted for.
- Where an incoming cross-reference to the presently edited article is not yet complete or invalid, *vernetziko* can help to update the source document of the cross-reference in a few simple steps.

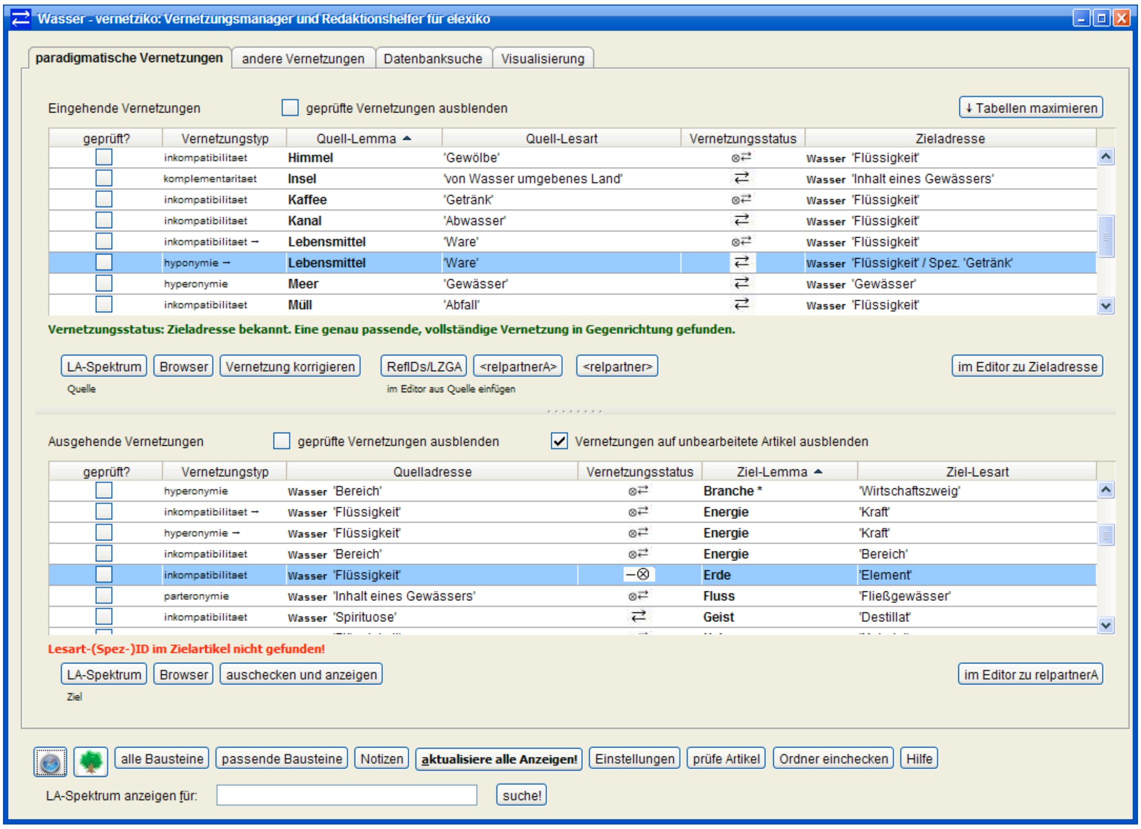


EWL: Wasser (261769) : Flüssigkeit
EWL: Wasser (261769) : Flüssigkeit : Getränk
EWL: Wasser (261769) : Inhalt eines Gewässers
EWL: Wasser (261769) : Gewässer
EWL: Wasser (261769) : Flüssigkeit im Körper

Mit Wasser bezeichnet man ein Gewässer, beispielsweise einen Bach oder ein Meer.

Figure 1: Selecting a word sense

Incoming and outgoing cross-references are listed in tabular form, cf. Fig. 2. References concerning sense-relations (synonymy, hyponymy etc.) are listed separately from all other kinds of references. The tables provide standard sorting and filtering functionality. *vernetziko* also offers a tree view of outgoing word sense references that displays the relevant parts of the XML structure (Fig. 3). Both these tables and the tree view can be used interactively for fast navigation, sorting and reference insertion.



Eingehende Vernetzungen □ geprüfte Vernetzungen ausblenden

geprüft?	Vernetzungstyp	Quell-Lemma ▲	Quell-Lesart	Vernetzungsstatus	Zieladresse
□	inkompatibilitaet	**Himmel**	'Gewölbe'	⊗⇄	Wasser 'Flüssigkeit'
□	komplementaritaet	**Insel**	'von Wasser umgebenes Land'	⇄	Wasser 'Inhalt eines Gewässers'
□	inkompatibilitaet	**Kaffee**	'Getränk'	⊗⇄	Wasser 'Flüssigkeit'
□	inkompatibilitaet	**Kanal**	'Abwasser'	⇄	Wasser 'Flüssigkeit'
□	inkompatibilitaet →	**Lebensmittel**	'Ware'	⊗⇄	Wasser 'Flüssigkeit'
□	hyponymie →	**Lebensmittel**	'Ware'	⇄	Wasser 'Flüssigkeit' / Spez. 'Getränk'
□	hyperonymie	**Meer**	'Gewässer'	⇄	Wasser 'Gewässer'
□	inkompatibilitaet	**Müll**	'Abfall'	⇄	Wasser 'Flüssigkeit'

Vernetzungsstatus: Zieladresse bekannt. Eine genau passende, vollständige Vernetzung in Gegenrichtung gefunden.

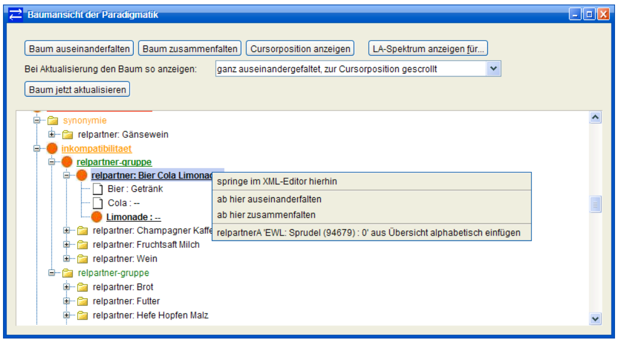
Ausgehende Vernetzungen □ geprüfte Vernetzungen ausblenden ☑ Vernetzungen auf unbearbeitete Artikel ausblenden

geprüft?	Vernetzungstyp	Quelladresse	Vernetzungsstatus	Ziel-Lemma ▲	Ziel-Lesart
□	hyperonymie	Wasser 'Bereich'	⊗⇄	**Branche ***	'Wirtschaftszweig'
□	inkompatibilitaet →	Wasser 'Flüssigkeit'	⊗⇄	**Energie**	'Kraft'
□	hyperonymie →	Wasser 'Flüssigkeit'	⊗⇄	**Energie**	'Kraft'
□	inkompatibilitaet	Wasser 'Bereich'	⊗⇄	**Energie**	'Bereich'
□	inkompatibilitaet	Wasser 'Flüssigkeit'	–⊗	**Erde**	'Element'
□	parteronymie	Wasser 'Inhalt eines Gewässers'	⊗⇄	**Fluss**	'Fließgewässer'
□	inkompatibilitaet	Wasser 'Spirituose'	⇄	**Geist**	'Destillat'

Lesart-(Spez-)ID im Zielartikel nicht gefunden!

LA-Spektrum anzeigen für:

Figure 2: Tabular view of incoming (upper table) and outgoing (lower table) sense relations of *Wasser* ('water')


Baumansicht der Paradigmatik
Baum auseinanderfalten
Baum zusammenfalten
Cursorposition anzeigen
LA-Spektrum anzeigen für...
Baum jetzt aktualisieren
synonymie
relpartner: Gänsewein
inkompatibilitaet
relpartner-gruppe
relpartner: Bier Cola Limonad
Bier : Getränk
Cola : --
Limonade : --
relpartner: Champagner Kaffe
relpartner: Fruchtsaft Milch
relpartner: Wein
relpartner-gruppe
relpartner: Brot
relpartner: Futter
relpartner: Hefe Hopfen Malz
springe im XML-Editor hierhin
ab hier auseinanderfalten
ab hier zusammenfalten
relpartnerA 'EWL: Sprudel (94679) : 0' aus Übersicht alphabetisch einfügen

Bei Aktualisierung den Baum so anzeigen: ganz auseinandergefaltet, zur Cursorposition gescrollt

Figure 3: Using the tree view for sense relations

3.2 Cross-reference status

For the working lexicographer, the most relevant information in the tabular presentation is the *status* of the individual cross-references. The status is symbolized by various arrow icons that inform the user on the extent to which different requirements are met. In particular, cross-references should be *complete* and *well-formed*; more important, they must be *valid*, pointing to a target address that really exists in the lexicographic database, even if this address happens to be a preliminary reference to a still unedited article. Compulsory symmetry and transitivity in certain reference types such as synonymy can be an additional consistency requirement.

A cross-reference may fail to be valid or consistent in many different ways. The status icons are based on a systematic typology of possible cross-reference statuses that is exhaustive but still perspicuous and practical from the lexicographer's point of view.

In order to simplify the exposition of this typology, some terminology will be introduced first. A *unidirectional cross-reference*, or reference for short, is a labeled ordered pair consisting of a source and a target *address*. The label is the *relation type* encoded by the cross-reference, e.g. 'is a synonym of', 'is morphologically derived from'. An address is an identifiable subpart of a *resource*. Besides dictionary entries, examples for possible resources include files, Internet URLs and other digitally represented structured text documents. In a dictionary entry, sections pertaining to specific word senses are examples of addresses. If a dictionary entry is encoded as an XML document, any XML element within that document is a potential address, as long as it is systematically identifiable by an XPath expression. In many resources, different address types must be distinguished, such as word senses vs. sections on grammar in a dictionary. When no reference to a subpart of a resource is possible or necessary, this will be modeled as an address type with only one trivial address per resource. – Note that source and target address may belong to the same resource.

Simplifying somewhat, *vernetziko* distinguishes between the following statuses of unidirectional references:

a. The target resource does not exist or its specification is either formally inadmissible or factually inconsistent.
b. The specification of the target resource is incomplete.
c. The target resource is correctly specified, but the target address within that resource does not exist or its specification is either formally inadmissible or factually inconsistent.
d. The target resource is correctly specified, but the target address within that resource is not fully specified, possibly because the target resource is an as yet unedited entry.
e. The target address is correctly and fully specified.

If at least the target resource has been specified correctly in two different cross-references and there are no inconsistencies or other errors in both references (i.e., only cases d. and e. apply), these two references form a *possible bidirectional cross-reference* and are thus *possible reverse cross-references* to each other if and only if their relation types match (e.g. hyponymy vs. hyperonymy) and the target address of each reference is either equal to the source address of the other reference or contains this source address as a subpart.

For a given reference R this leads to the following panoply of possibilities regarding reverse references:

f. There is no possible reverse cross-reference for R, although the relation type of R admits of such references.
g. There is no possible reverse cross-reference for R, although this is considered compulsory (e.g. in case of synonymy, at least for *elexiko*).
h. R and exactly one of the potential reverse cross-references both have status e. above (target address correctly and fully specified). This is the case of a 'perfect' bidirectional reference.
i. There is more than one possible reverse reference for R, but none of these cases meets the requirements of h. above.
j. There is exactly one potential reverse reference, but at least one of the two references is not fully specified (in the sense of d. above).

In order to establish the status of cross-references, *vernetziko* uses Oracle's XML-enabled full text search capabilities to obtain all incoming cross-references, then reads in the XML data of all entries referencing and referenced by the presently edited one, parses all XML documents using a StAX parser and finally tries to match all cross-references with addresses in the respective entries and with possible reverse references. The user can start this process manually; a background task checking periodically for relevant changes in the currently edited XML document updates status information every five seconds.

3.3 Implementational aspects: Handling the interplay with the XML Editor

A fair amount of typical editing functions must be present in the XML editor's API, such as navigating the caret to arbitrary XML elements, inserting, deleting and modifying XML elements, opening and closing XML documents etc. As stated above, a Java interface represents all methods used to call editor functionality from within *vernetziko*. The editor-specific API calls themselves are encapsulated in a single class that comforms to this interface. For the *elexiko* project, two implementations of the interface have been developed so far, viz. for Corel XMetaL 3.1 and for the <oXygen/> XML editor (version 13). Any editor suitable for this kind of modular setup must either be usable as an application server to other standalone programs (for instance, through a COM mechanism in MS Windows operating systems; this is the case with XMetaL) or expose its API via some sort of plugin architecture (this is the technique chosen for <oXygen/>). These two scenarios have rather different technical implications, however; changing from one of them to the other is not a trivial task. In the first case, *vernetziko* is a standalone desktop application, in the second, it is provided as a bunch of plugin classes.

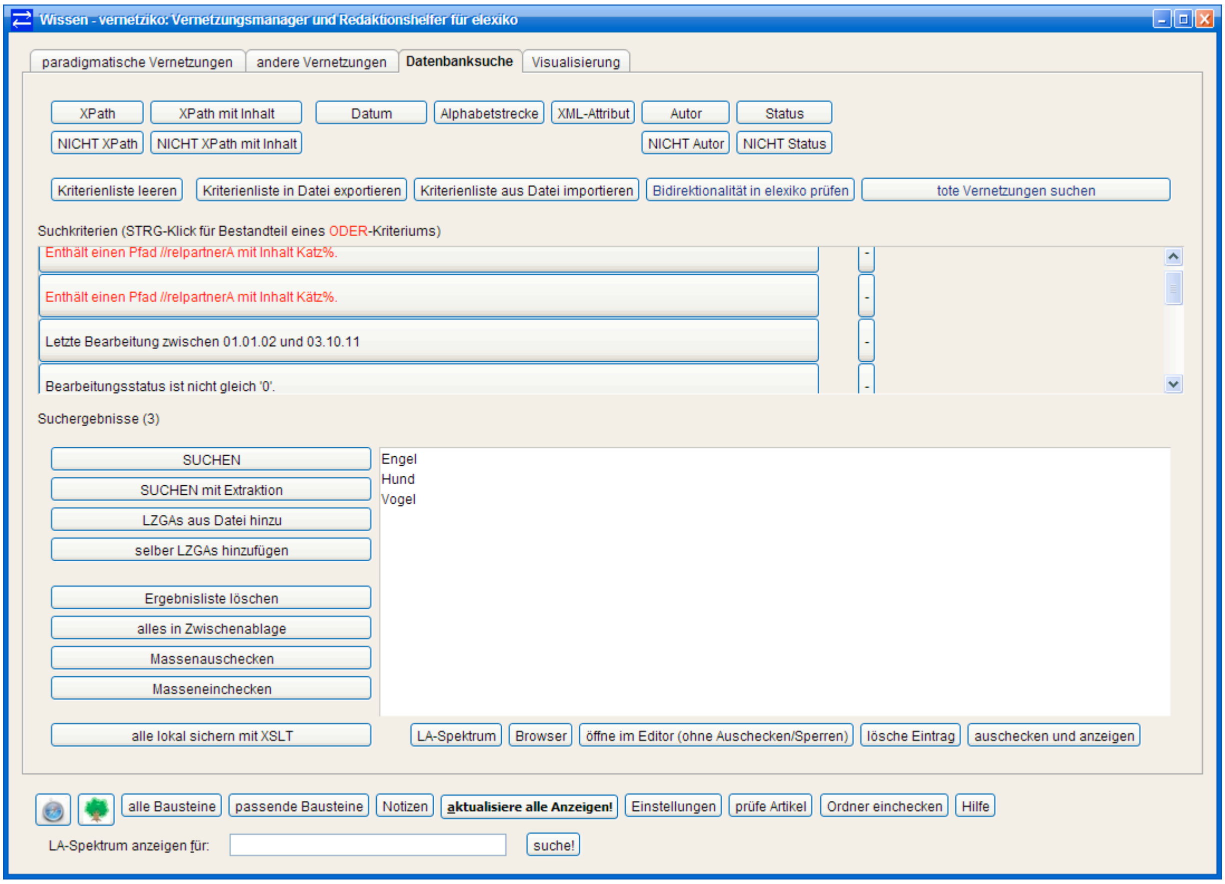
The most difficult aspect of a modular approach to remote-controlling the XML editor is that different editors use different, mostly proprietary, APIs to describe the structure of XML documents. Naturally, all of these APIs bear a certain similarity to, e.g., the Java DOM API. Since the editor-specific API classes representing XML nodes, elements, documents and attributes must be processed in many ways by *vernetziko*, it is necessary to devise editor-independent interfaces that represent the needed functionality of node/element/attribute/document classes. The editor-specific XML objects are then referenced in wrapper classes implementing these interfaces. This way, we obtain an editor-independent DOM-like representation of the editor's XML nodes; throughout *vernetziko*'s code, only the wrapper classes are used.

4. Further assistive management tools

4.1 Features of the user interface

vernetziko features a number of additional tools that help to speed up and simplify the editing process:

- Article-specific notes and XML snippets can easily be stored, retrieved and inserted into the edited document.
- An advanced database search tool allows complex Boolean combinations of search criteria including metadata and XPath expressions, cf. Fig. 4.
- Administrators may perform operations on large sets



Suchkriterien (STRG-Klick für Bestandteil eines ODER-Kriteriums)

Enthält einen Pfad //relpartnerA mit Inhalt Katz%.

Enthält einen Pfad //relpartnerA mit Inhalt Kätz%.

Letzte Bearbeitung zwischen 01.01.02 und 03.10.11

Bearbeitungsstatus ist nicht gleich '0'.

Suchergebnisse (3)

Engel
Hund
Vogel

LA-Spektrum anzeigen für:

Figure 4: Extended database search and scanning options

of entries (alter user access rights, check in or out) that can be defined by search criteria or specified manually or from lemma list text files.

- Cross-references concerning sense relations such as synonymy and hyponymy can be visualized graphically. The visualization program traverses arbitrarily long chains of incoming or outgoing cross-references and can recursively construct graphs with very large numbers of nodes (word senses).

vernetziko not only helps to secure consistency of cross-references in individual dictionary entries, but also provides tools for scanning the entire lexicographic database of *elexiko* for problematic cross-references, viz.

- inconsistent references, in particular 'dead' references pointing to inexistent entries or word senses;
- unidirectional references for which a required reverse reference does not exist yet.

The results of these scans are output as UTF-8 text files.

4.2 General considerations on database design

There are several reasons why the seemingly obvious strategy of storing the cross-reference structure of a dictionary in a separate relational data structure in the database is not always feasible. For instance, the exact position of cross-references within the source element in an article's XML representation might vary depending on lexicographical considerations, which would necessitate the use of 'pointers' from within the XML document to the external link table. In such cases, a separate cross-reference table introduces new sources of possible inconsistencies and considerably complicates the editing process for dictionary entries since two database tables must be modified concurrently and kept in synch. It can be argued that an automatically updated relational 'cache' table that simply duplicates basic cross-reference data (addresses and reference type) is the right solution to meet performance requirements in these cases (cf. Joffe et al., 2003; Meyer & Müller-Spitzer, 2010). For the time being, even this solution is not used in the *elexiko* project since the database XML query technology is still fast enough to cope with real-time requirements.

The approach taken for *elexiko* therefore employs a maximally lean and redundancy-free database design and shifts the administrative burden to the external software tool *vernetziko*.

5. Managing Cross-References in a Dictionary Portal: Software Engineering Considerations

elexiko forms part of OWID, a web portal of German electronic dictionaries (Müller-Spitzer, 2010b). A tool such as *vernetziko* should be easily adaptable to the integration of new lexicographic resources into the portal, in particular with respect to managing cross-references between different dictionaries of potentially heterogeneous structure, i.e. with widely differing DTDs/schemas.

The design of such management software has to address many challenges, if maximum generality, extensibility and reusability of software components are to be combined with a maximally perspicuous and parsimonious approach. These challenges include the following points:

- In an Internet portal, new online dictionaries may be added at any time. Entries in different dictionaries may be structured in various ways – conforming to widely differing DTDs/XML schemas – and contain disparate types of addresses.
- A specific type of address (e.g., word senses in a dictionary) may be encoded differently in different dictionaries. For example, the structure of the XPath associated with a word sense in an XML-based entry might vary according to the dictionary.
- The very same address type might be referred to in different ways depending on the referring resource, using, e.g., different XML attribute names.
- Sometimes the position where a reference is encoded in a document is relevant to the identification of the reference, sometimes not.
- Address types may differ to a great extent in the kind of informational structure associated with them; compare references to word senses with references to web URLs or citations.
- The programmer should be able to add new types of addresses or references in a modular way, if possible without touching existing classes.
- Different resources, address types, and reference types require different operations in a management tool. There is no set of common methods for all address or reference types pertaining to a certain resource. One and the same address type might even have to be treated differently, depending on the resource it occurs in. The implementation of methods that take references as input can depend on the resource and address type of both source and target entry.

Thus, from a software engineering perspective, *vernetziko* has to cope with a variant of what is often called the *expression problem*: New dictionaries and cross-reference types may require the addition of both new classes representing types of resources/addresses/references and new operations on objects of such classes. For *vernetziko*, a very simple solution based on parameterized types will be presented here. The solution is not strictly type-safe in that it uses type checks and subsequent casts, but given the lack of self types, multimethods, mixins etc. in Java, any completely type-safe solution produces an enormous overhead in static languages, cf. (Torgersen, 2004). In the pragmatic approach taken for *vernetziko*, there is one and only one class that is responsible for dispatching all method calls concerning resources, addresses and references. After adding new classes of any of these entity types, only the dispatcher class needs to be modified accordingly; type checks and casts are performed only in this class.

5.1 Domain entity classes

Entries. Since the different portal dictionaries are no suitable candidates for domain entities – no elementary operations are performed on dictionaries as a whole –, the notion of an entry belonging to a specific dictionary (or, more generally, that of a resource) is taken as the point of departure for the domain class model. All entry classes such as *Dictionary1Entry*, *Dictionary2Entry*, … derive from an abstract class *Entry* and store information that identifies the particular individual resource.

Addresses. Different address types are represented by subclasses (*WordSenseAddress*, *GrammarAddress*, …) of an abstract *Address* class that contain a reference to the *Entry* object the address object 'belongs' to. Different address types will require widely differing sets of fields for the information associated with them. One and the same address type may appear in entries of different resources; for instance, two dictionaries may each have dedicated sections for different word senses within every entry. On the other hand, a distinction between word senses in Dictionary1 and Dictionary2 is still needed, since they might have slightly differing formal representations, such as differing names of the relevant XML elements or attributes. Therefore, we parameterize the static address types on the type of the *Entry* field. In Java notation, the same sort of address, e.g., word senses, is reflected by different static types, e.g. *SenseAddress <Dict1Entry>* and *SenseAddress <Dict2Entry>*, according to the resource its entry belongs to.

References can be dealt with accordingly. In many scenarios, a single *Reference* class will suffice whose fields are references to the source and the target *Address* objects. Depending on the context, further fields will be used to represent classificatory or status-related information about a reference. Here, we parameterize on the types of both the source and the target address. The static type of a specific reference from a word sense in a dictionary entry to a paper in a specific volume of a linguistic journal may then look as follows in Java: *Reference <SenseAddress <DictionaryEntry>, PaperAddress <JournalEntry>>* (where *JournalEntry* objects model journal volumes).

5.2 Dispatcher class

Although objects of classes *AddressX <Dict1Entry>* and *AddressX <Dict2Entry>* share the same internal class makeup – representing the same sort of address in two different resources and therefore both being of type *AddressX<? extends Entry>* –, they must possibly be handled differently, requiring, e.g. different code for navigating in the editor to the corresponding element. On the other hand, code duplication has to be avoided in the case where certain (but possibly not all!) methods pertaining to these both types can in fact be implemented identically.

In addition, not every address type is 'compatible' with a given resource (images don't have word senses); additionally, most combinations of a source and a target address type do not amount to a valid reference type. Many operations may only be relevant for a small subset of, say, address types (consider the task of printing information about an address). These many 'holes' in the matrices of actually existing type combinations and actually permitted parameterized types per operation cannot be accounted for in advance by the type system or some sort of inheritance hierarchy.

In typical scenarios, most methods don't change the state of entry, address and reference objects, the latter rather being used like 'passive' information containers. In addition, new functionality operating on addresses or references might be added at any time to the management application, which increases the danger of bloated and ever growing interfaces with empty implementations for many subclasses.

All considerations mentioned above point to a solution where domain entity objects are treated as mere data containers with minimal public interfaces. All public methods of the domain entity classes relay to the special dispatcher class mentioned above. As an example, a method call like *myAddress.moveXMLEditorCaretHere()* would be relayed by calling a static method, *Dispatcher. moveXMLEditorCaretHere(myAddress)*. The static method *moveXMLEditorCaretHere(Address<?> anAddress)* of the dispatcher class then type-checks the input parameter *anAddress* and, after a corresponding cast, calls the appropriate method of some service class in a type-safe manner. Note that though the Java compiler erases type information in generics, the parameter type can be obtained at runtime by getter methods: In our example, *myAddress* holds a reference to the resource (i.e. dictionary entry) it belongs to; the runtime type of this resource object is identical to the parameter type of *myAddress*.

6. Concluding Remarks

The software tool *vernetziko* adds advanced cross-reference management facilities and various helper tools to an already existing dictionary database system and editing environment. This is possible due to a modular software design that encapsulates the access to both other components of the IT environment, such as the XML editor, and the internal structural makeup of the lexicographic data involved. Support for new dictionary resources and new types of cross-references within and between dictionaries can easily be added in a plugin-like fashion. While most of the functionality provided by *vernetziko* is part and parcel of many commercial dictionary writing systems, the main advantage of the approach taken with *vernetziko* is that the software can be adapted to interconnect a wide variety of off-the-shelf components (database management systems and editors) and allows tailor-made access to and administration of almost arbitrary XML resources and legacy dictionary data, thus providing the 'glue' for a tailor-made lexicographical workbench without vendor lock-in – ideally suited to large-scale projects and to the management of cross-references between multiple dictionaries.

7. References

Gamma, E., Helm, R., Johnson, R.E., & Vlissides, J. (1995). *Design Patterns. Elements of Reusable Object-Oriented Software*. Amsterdam: Addison-Wesley Longman.

Haß, U. (ed.) (2005). *Grundfragen der Elektronischen Lexikographie: Elexiko - Das Online-informationssystem zum deutschen Wortschatz*. Berlin, New York: de Gruyter.

Joffe, D., de Schryver, G.-M. & Prinsloo, D.S. (2003). Computational features of the dictionary application "TshwaneLex". *Southern African Linguistics and Applied Language Studies* 21(4), pp. 239-250.

Klosa, A. (ed.) (2011). *elexiko - Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs*. Tübingen: Gunter Narr.

Lemnitzer, L., Romary, L. & Witt, A. (to appear). Representing Human and Machine Dictionaries in Markup Languages. In R.H. Gouws, U. Heid, W. Schweickhard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography*. Berlin/New York: de Gruyter.

Martin, R.C. (1997). Acyclic Visitor. In R.C. Martin, D. Riehle & F. Buschmann (eds.) *Pattern languages of program design 3*. Boston, MA: Addison-Wesley Longman, pp. 93-103.

Meyer, P., Müller-Spitzer, C. (2010). Consistency of Sense Relations in a Lexicographic Context. Workshop on Semantic Relations, *International Conference on Language Resources and Evaluation (LREC) 2010*, May 18, Malta.

Müller-Spitzer, C. (2007). Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. *Hermes* 38, pp. 137-171.

Müller-Spitzer, C. (2010a). The Consistency of Sense-Related Items in Dictionaries. Current Status, Proposals for Modelling and Potential Applications in Lexicographic Practice. In P. Storjohann (ed.) *Lexical-Semantic Relations. Theoretical and Practical Perspectives*. Lingvisticæ Investigationes Supplementa. Amsterdam/New York: Benjamins, pp. 145-162.

Müller-Spitzer, C. (2010b). OWID – A dictionary net for corpus-based lexicography of contemporary German. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress. Leeuwarden, 6-10 July 2010*. Fryske Akademy: Leeuwarden, pp. 445-452.

Müller-Spitzer, C., Schneider, R. (2009). Ein XML-basiertes Datenbanksystem für digitale Wörterbücher – Ein Werkstattbericht aus dem Institut für Deutsche Sprache. *it – Information Technology* 51(4), pp. 197-206.

Torgersen, M. (2004). The Expression Problem Revisited. Four New Solutions Using Generics. In M. Odersky (ed.) *ECOOP 2004 - Object-Oriented Programming: 18th European Conference, Oslo, Norway, June 14-18, 2004. Proceedings* (Lecture Notes in Computer Science 3086). Berlin: Springer, pp. 123-146.

The "Online Bibliography of Electronic Lexicography" (OBELEX)

Christine Möhrs, Antje Töpel
Institut für Deutsche Sprache
R 5, 6-13, D-68161 Mannheim
E-mail: moehrs@ids-mannheim.de, toepel@ids-mannheim.de

Abstract

Digital or electronic lexicography has gained in importance in the last few years. This can be seen in the increasing number of online dictionaries and publications focusing on this field. OBELEX (http://www.owid.de) – one of the bibliographic projects of the Institute for German Language in Mannheim – takes this development into account and makes both online dictionaries and research contributions available in a bibliographical database searchable by different criteria. The idea for OBELEX originated in the context of the dictionary portal OWID, which incorporates several dictionaries from the Institute for German Language (http://www.owid.de). OBELEX has been available online free of charge since December 2008. As of 2011, OBELEX includes two search options: a search for research literature and (as a completely new feature) a search for online dictionaries, a service which is unique in the world.

Keywords: bibliography; database; online dictionaries

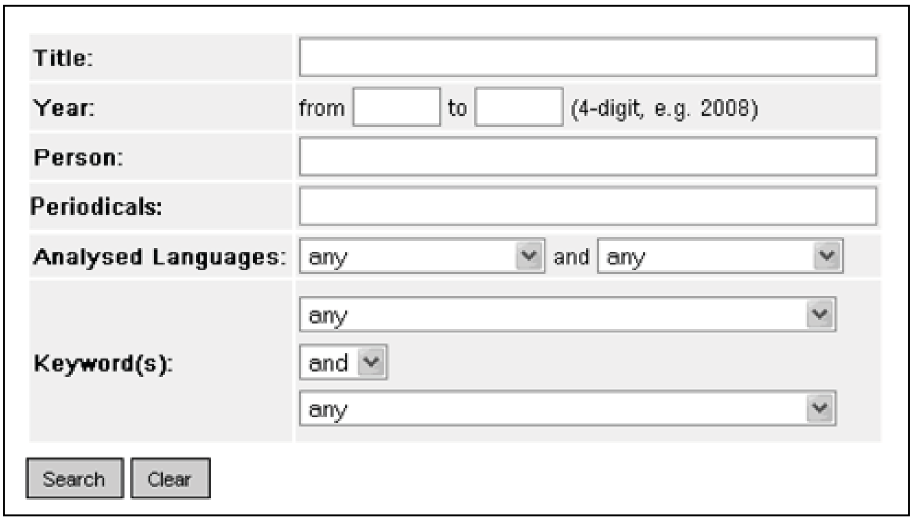
1. Introduction

Digital or electronic lexicography has gained in importance in the last few years. This can be seen in the increasing number of online dictionaries and publications focusing on this field. OBELEX (http://www.owid.de) – one of the bibliographic projects of the Institute for German Language in Mannheim – takes this development into account and makes both online dictionaries and research contributions available in a bibliographical database searchable by different criteria. The idea for OBELEX originated in the context of the dictionary portal OWID, which incorporates several dictionaries from the Institute for German Language (http://www.owid.de). OBELEX has been available online free of charge since December 2008. As of 2011, OBELEX includes two search options: a search for research literature and (as a completely new feature) a search for online dictionaries, a service which is unique in the world.

2. Database on research literature

The database on research literature contains articles, monographs, anthologies and reviews published since 2000 with respect to electronic lexicography, as well as some older relevant works (current size: approx. 1000 entries). Each bibliographic entry gives information on title, year, person, periodical, analysed languages or keyword(s). Since all publications are associated with our keyword list, a thematic search is possible.

In addition to the systematically evaluated sources (see below), other relevant research literature is included in OBELEX, such as monographs from the field of electronic lexicography and articles from journals that are not systematically evaluated. Reviews are also included because they often present interesting metalexicographic aspects concerning the critical evaluation of electronic dictionaries and are quite often not easily accessible. As far as possible, abstracts are given, especially for articles from conference proceedings. In the future, OBELEX will be extended systematically.



Title:	
Year:	from to (4-digit, e.g. 2008)
Person:	
Periodicals:	
Analysed Languages:	any and any
Keyword(s):	any and any

Figure 1: Search form for research literature

2.1 Evaluated Sources

The systematically evaluated literature (with a focus on electronic lexicography) in OBELEX includes the following sources:

- Dictionaries: Journal of the Dictionary Society of North America (DSNA).
- Handbücher zur Sprach- und Kommunikationswissenschaft (HSK): Hausmann, Franz Josef / Reichmann, Oskar / Wiegand, Herbert Ernst / Zgusta, Ladislav (ed.) (1989-1991): Wörterbücher. Dictionaries. Dictionnaires: Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie international de lexicographie (HSK 5.1). Berlin / New York: de Gruyter (= Handbücher zur Sprach- und Kommunikationswissenschaft. Handbooks of Linguistics and Communication Science. Manuels de linguistique et des sciences de communication 5.1-5.3).
- Hermes: Journal of language and communication studies. Institution: Aarhus School of Business.
- International Journal of Lexicography (Oxford Journals).
- Lexicographica. Internationales Jahrbuch für Lexikographie. International annual for lexicography. Revue internationale de lexicographie. Institutions: Dictionary Society of North America (DSNA) and the European Association for Lexicography (EURALEX).
- Lexicographica: series maior. Institutions: Dictionary Society of North America (DSNA) and the European Association for Lexicography (EURALEX).
- Lexikos: Annual Journal of the African Association for Lexicography (AFRILEX).
- Conference Proceedings of European Association for Lexicography (EURALEX); 2000 (Stuttgart), 2002 (Copenhagen), 2004 (Lorient), 2006 (Turin), 2008 (Barcelona) and 2010 (Leeuwarden).
- Conference Proceedings of the 9th-12th "International Symposium on Lexicography" at the University of Copenhagen.

2.2 Other bibliographic projects

OBELEX supplements other bibliographic projects in a useful way: firstly, the printed "Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung" by Herbert Ernst Wiegand (Wiegand, 2006/2007), and secondly the "Bibliography of Lexicography" by R.R.K. Hartmann (Hartmann, 2007), and lastly the "International Bibliography of Lexicography" of Euralex (cf. also DeCesaris/Bernal, 2006). OBELEX differs from all these bibliographic projects:

- The bibliography by Wiegand certainly has the broadest approach, but it does not focus on electronic lexicography. Furthermore, searching for a specific publication in this bibliography is not easy, since the forthcoming register volume has not yet been published (cf. Dziemianko, 2008). Besides this, the main focus is on dictionary research within the field of German Studies. OBELEX, however, has no such restriction.
- The Euralex bibliography will include all publications from Euralex conference proceedings. However, other periodicals or journals will not be included systematically. Thus, if one does not want to restrict a bibliographical search to publications in the Euralex proceedings, the use of OBELEX is advisable.
- The Hartmann bibliography is comprehensive and international. However, for anthologies, only the title of the book as a whole is listed, not the individual articles. Thus, searching for specific articles or reviews is not possible in this bibliography, while OBELEX lists each article separately.

2.3 Search options

As mentioned above, there are different search options in OBELEX, which are explained below.

Search by title: The title search is a real full-text search, so results are obtained by inserting a search string in the title box.

Search by publication year: The search for publications in OBELEX can also be delimited by publication year. For instance, it is possible to search for all titles on the subject of electronic lexicography published from 2005 to 2010. For this particular search, "2005" should be entered in the "from"-field and "2010" in the "to"-field.

Search by person or periodicals: This incremental search option offers the possibility of typing in the first letters of an author's name (such as "be") or of a periodical's name (such as "int"). Then all appropriate people (e.g. "Bernal, Elisenda") or periodicals (e.g. "International Journal of Lexicography") included in OBELEX appear and can be selected from a list.

Search by keywords or analysed language: Two of the most important functions of OBELEX are the options to search by keyword and by analysed language. These fields allow a thematic search. For example, in combination with the chosen language it is possible to search for all bibliographic entries from the field of "online lexicography" that deal with online dictionaries in "Slovenian".

3. Database on online dictionaries

As an entirely new feature, OBELEX offers a search for online dictionaries. Firstly, this service will help orientate users in the growing market of Internet dictionaries, helping them to find a dictionary for certain languages or with special characteristics. Secondly, lexicographers and metalexicographers can search for all online dictionaries with a specified set of features. To date (July 2011), about 19,000 dictionaries are listed in the database of online dictionaries. This huge number results mainly from itemizing all language pairs in dictionary portals, such as Dicts.info or Sanakirja.org.

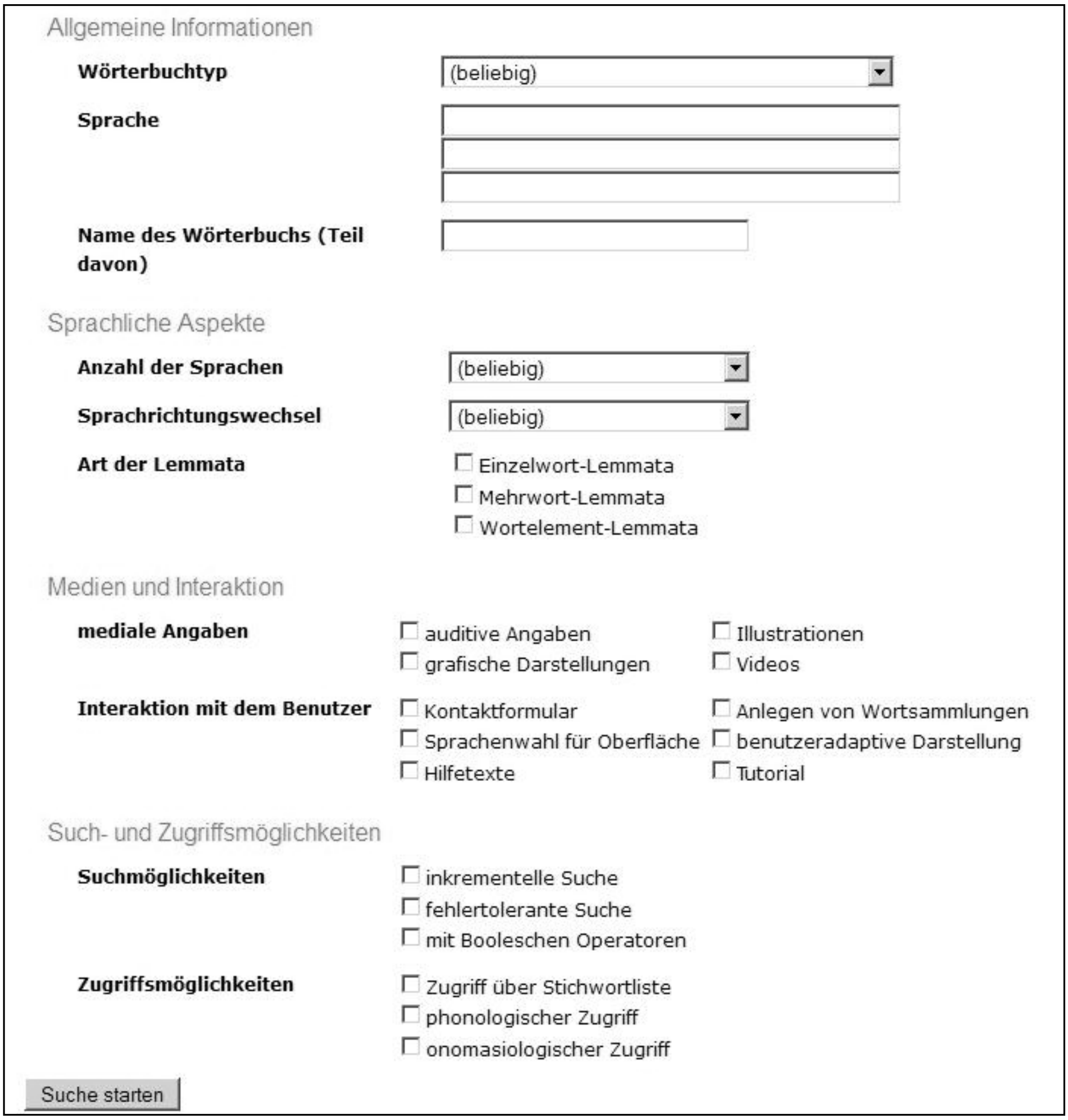
The database contains different kinds of information on each dictionary, for example name, language and search options offered to the user, as well as more specific information. The result pages present direct hyperlinks to the dictionaries covered (cf. also section 3.2).

In future, in addition to routine maintenance, further extension of the underlying database and new search options (e.g. for language families or groups) are planned.

3.1 Evaluated Sources

The database of online dictionaries has been compiled by sifting through link lists (such as Linguist List or LinseLinks), bibliographies and metalexicographic literature on online dictionaries. Of particular interest have been the proceedings of relevant conferences on lexicography, such as the EURALEX-conference, and metalexicographic journals with special sections on the presentation of existing or new online dictionaries (cf. also section 2.1 for a list of corresponding conferences and journals).

Many of these online dictionaries and most of the metalexicographic literature also mention other online dictionaries, for instance in the form of link lists. Including these dictionaries in the database produces a snowball effect, resulting in a growing number of listed dictionaries.



Allgemeine Informationen

Wörterbuchtyp	(beliebig)	
Sprache		
Name des Wörterbuchs (Teil davon)		

Sprachliche Aspekte

Anzahl der Sprachen	(beliebig)	
Sprachrichtungswechsel	(beliebig)	
Art der Lemmata	□ Einzelwort-Lemmata □ Mehrwort-Lemmata □ Wortelement-Lemmata	

Medien und Interaktion

mediale Angaben	□ auditive Angaben □ grafische Darstellungen	□ Illustrationen □ Videos
Interaktion mit dem Benutzer	□ Kontaktformular □ Sprachenwahl für Oberfläche □ Hilfetexte	□ Anlegen von Wortsammlungen □ benutzeradaptive Darstellung □ Tutorial

Such- und Zugriffsmöglichkeiten

Suchmöglichkeiten	□ inkrementelle Suche □ fehlertolerante Suche □ mit Booleschen Operatoren	
Zugriffsmöglichkeiten	□ Zugriff über Stichwortliste □ phonologischer Zugriff □ onomasiologischer Zugriff	

Figure 2: Search form for online dictionaries

3.2 Search options

As can be seen in Figure 2 above, the database on online dictionaries is searchable for various kinds of information by means of a very detailed search form that is explained in this section. Similar criteria are grouped together under corresponding headlines.

General information (*Allgemeine Informationen*): In the top section of the search form, the user can specify general information on the dictionaries. These are probably the most important and most widely used search criteria. The search by type of dictionary (*Wörterbuchtyp*) consists of choosing a dictionary type (e.g. learner's dictionary or dictionary of synonyms) from the drop-down list. Perhaps the most central criterion of a dictionary is its object language(s) (*Sprache*). These can be typed into an autocompleting text box (incremental search). The third and last option under this headline is the search by name of dictionary (*Name des Wörterbuchs (Teil davon)*), a full-text search for the complete dictionary name or just a part of it.

Linguistic aspects (*Sprachliche Aspekte*): The second section of the search form covers linguistic aspects of the online dictionaries. This includes number of languages (*Anzahl der Sprachen*) (monolingual, multilingual) as well as language direction (*Sprachrichtungswechsel*), both realised via drop-down lists. There is also a search by different types of headwords (*Art der Lemmata*) (affixes, single-word entries, multiword items), which can be selected by corresponding check boxes.

Media and interactivity (*Medien und Interaktion*): In this section of the search form, the focus is on how the dictionary makes use of multimedia and interactivity. 'Use of multimedia elements' (*mediale Angaben*) refers to audio files, illustrations, graphics and videos. 'Interaction with the user' (*Interaktion mit dem Benutzer*) relates to the existence of contact forms, vocabulary trainers, help texts and tutorials, but also to the option of choosing a user interface language or of having user-adaptive views. All items in these groups can be selected by check boxes.

Search and access (*Such- und Zugriffsmöglichkeiten*): The last part of the search form deals with search options (*Suchmöglichkeiten*) and the access structure (*Zugriffsmöglichkeiten*) of the online dictionaries. There are three check boxes for various types of search (incremental, fault-tolerant, Boolean operators) and for ways of accessing the dictionary (via lemma list, phonological access, onomasiological access).

The user may select any possible combination of the search criteria outlined above. However, the user interface automatically rules out logically impossible requests. For instance, if the user has already selected two languages, the option 'monolingual dictionary' is grayed out.

Having started a search, a second page presents a shortlist of all hits with hyperlinks to the online dictionaries, their name and their language(s). The next page displays the full title containing all the information on a particular online dictionary.

For instance, if you are looking for a Slovenian-German online dictionary with audio files or for a French learner's dictionary, you can now easily find them by using the OBELEX search form on online dictionaries.

4. Conclusion

The two systematic applications of OBELEX (research literature and online dictionaries) take the growing importance of digital or electronic lexicography into account. With OBELEX, we hope to provide an extensive service for researchers, lecturers and students who specialise in digital lexicography and research on online dictionaries.
The participants of the eLex2011 conference are therefore the users we have in mind while compiling OBELEX.

5. References

DeCesaris, J., Bernal, E. (2006). Lexicography as Reported: the EURALEX Conference Proceedings Bibliography (1983-2004). In E. Corino, C. Marello, & C. Onesti (eds) *Proceedings of the Twelfth EURALEX International Congress, Torino, Italia, September 6th - 9th, 2006*. Alessandria: Edizioni dell'Orso, pp. 1241-1247.

Dicts.info (2003ff.). Accessed at: http://dicts.info/.

Dziemianko, A. (2008). Review: Wiegand, Herbert Ernst. Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung: Mit Berücksichtigung anglistischer, nordistischer, romanistischer, slavistischer und weiterer metalexikographischer Forschungen. *International Journal of Lexicography,* 21(3), pp. 359-360.

Euralex: International Bibliography of Lexicography. Accessed at: http://euralex.pbworks.com.

Hartmann, R.R.K. (2007). Bibliography of Lexicography. Accessed at: http://euralex.pbworks.com.

LinseLinks. Die Linkdatenbank der Linse. Accessed at: http://links.linse.uni-due.de/.

OWID – Online-Wortschatz-Informationssystem Deutsch (2008ff.), edited by Institut für Deutsche Sprache, Mannheim. Accessed at: http://www.owid.de.

Sanakirja.org (2005ff.), edited by Jonne Jyrylä. Accessed at: http://www.sanakirja.org/.

The Linguist List. International Linguistics Community Online. Accessed at: http://linguistlist.org/langres/index.cfm.

Wiegand, H.E. (2006/2007). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung: mit Berücksichtigung anglistischer, nordistischer, romanistischer, slavistischer und weiterer metalexikographischer Forschungen*. Berlin/New York: de Gruyter.

What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project

Carolin Müller-Spitzer, Alexander Koplenig, Antje Töpel
Institute for German Language (IDS), Mannheim

E-mail: mueller-spitzer@ids-mannheim.de, koplenig@ids-mannheim.de, toepel@ids-mannheim.de

Abstract

This paper presents empirical findings from two online surveys on the use of online dictionaries, in which more than 1,000 participants took part. The aim of these studies was to clarify general questions of online dictionary use (e.g. which electronic devices are used for online dictionaries or different types of usage situations) and to identify different demands regarding the use of online dictionaries. We will present some important results of this ongoing research project by focusing on the latter. Our analyses show that neither knowledge of the participants' (scientific or academic) background, nor the language version of the online survey (German vs. English) allow any significant conclusions to be drawn about the participant's individual user demands. Subgroup analyses only reveal noteworthy differences when the groups are clustered statistically. Taken together, our findings shed light on the general lexicographical request both for the development of a user-adaptive interface and the incorporation of multimedia elements to make online dictionaries more user-friendly and innovative.

Keywords: dictionary use; empirical research; online dictionary

1. Introduction

Research into the use of online dictionaries is still quite a new field. Although some 250 to 300 studies have been carried out to date, the current state of knowledge still needs to be improved (Wiegand, 1998; Loucky, 2005; Welker, 2008; Engelberg & Lemnitzer, 2008; Tarp, 2009). Most studies are methodologically limited to the analysis of log files (e.g., de Schryver & Joffe, 2004; Bergenholtz & Johnson, 2005). While log file studies are able to provide reliable data about requested lemmas and related types of information, this method is not well suited to gaining insights into actual user demands. Take for instance the following hypothetical but plausible situation: Alex does not know the spelling of a particular word. To solve this problem, he visits an online dictionary. However, when trying to find the search window, he stumbles across various types of innovative buttons, hyperlinks and other distracting features. Instead of further using this online dictionary, he decides to switch to a well known search engine, because he prefers websites that enable him to easily find the information he needs. In this example, there would not be any data to log (except for an unspecified and discontinued visit to the website). In contrast, the market for online dictionaries is expanding both for academic lexicography and for commercial lexicography, with sales figures for printed reference works in continual decline. This has led to a demand for reliable empirical information on how online dictionaries are actually being used and how they could be made more user-friendly. As the example above indicates, relying completely on log file data can lead to biased conclusions in this context.

The remainder of this paper is structured as follows. In section 2, we will give a short overview of our project. Section 3 presents some of the hypotheses to be tested regarding online dictionary users' demands, while section 4 explains the methodological procedure. Section 5 describes some basic results. Finally, this study concludes with a short discussion of the implications of our findings (section 6) and briefly outlines our future work (section 7).

2. Project background

The project "User-adaptive access and cross-references in elexiko (BZVelexiko)" (www.using-dictionaries.info) aims to make a substantial contribution to closing this research gap. BZVelexiko is an externally funded joint research project at the Institute for German Language in Mannheim. For a period of three years, a group of researchers from different academic backgrounds (lexicographers, linguists, social scientists) is undertaking several extensive studies on the use of online dictionaries, using established methods of empirical social research. The first two studies focused on online dictionaries in general; subsequent studies in our project are restricted to monolingual German online dictionaries such as *elexiko* or the dictionary portal OWID (www.owid.de).

3. Demands on online dictionaries

Providing reliable empirical data that can be used to answer the question of how users rate different aspects of online dictionaries is an important issue for practical lexicography, because it can be used as the basis of various decisions that have to be made in this context. Is it more important to use financial and human resources to extend the corpus and improve its accessibility for the user community, or to focus on keeping the dictionary entries up to date? Which is more user-friendly, a fast user interface or a customizable user interface? Do different user groups have different preferences? For example, one of our hypotheses was that, compared to non-linguists, linguists would have a stronger preference for the entries to be linked to the relevant corpus, because this documents the scientific basis of the given information.

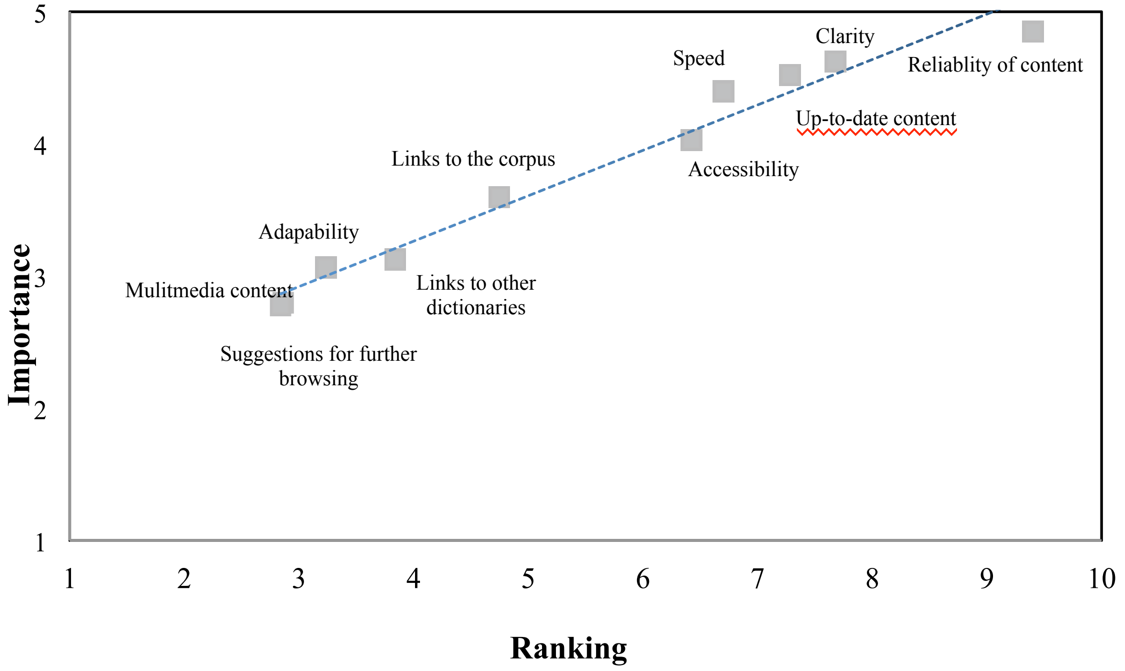


Figure 1. Correlations between means of ranks and means of importance regarding the use of an online dictionary. *Note*. Means of ranks are on 10-point scales and means of importance are on 5-point scales; both with higher values indicating higher levels of benefit

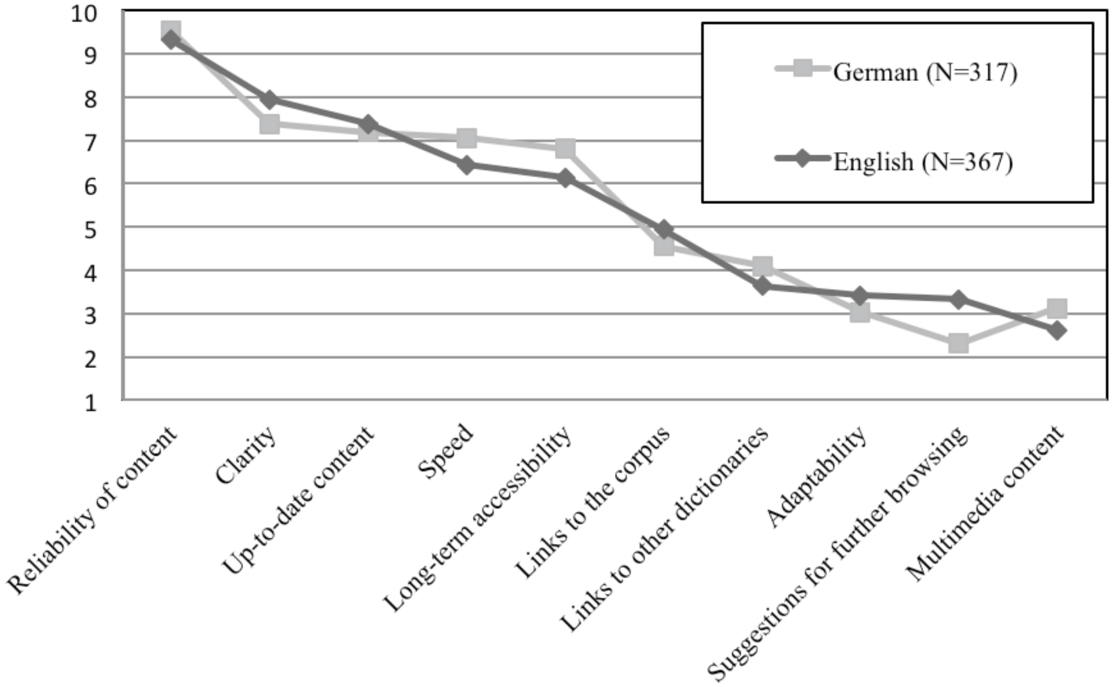


Figure 2. Means of Rankings as a Function of Language Version. *Note*. Means (Fig. 2, Fig. 3 and Fig. 4) are on 10-point scales with higher values indicating higher levels of importance regarding the use of an online dictionary.

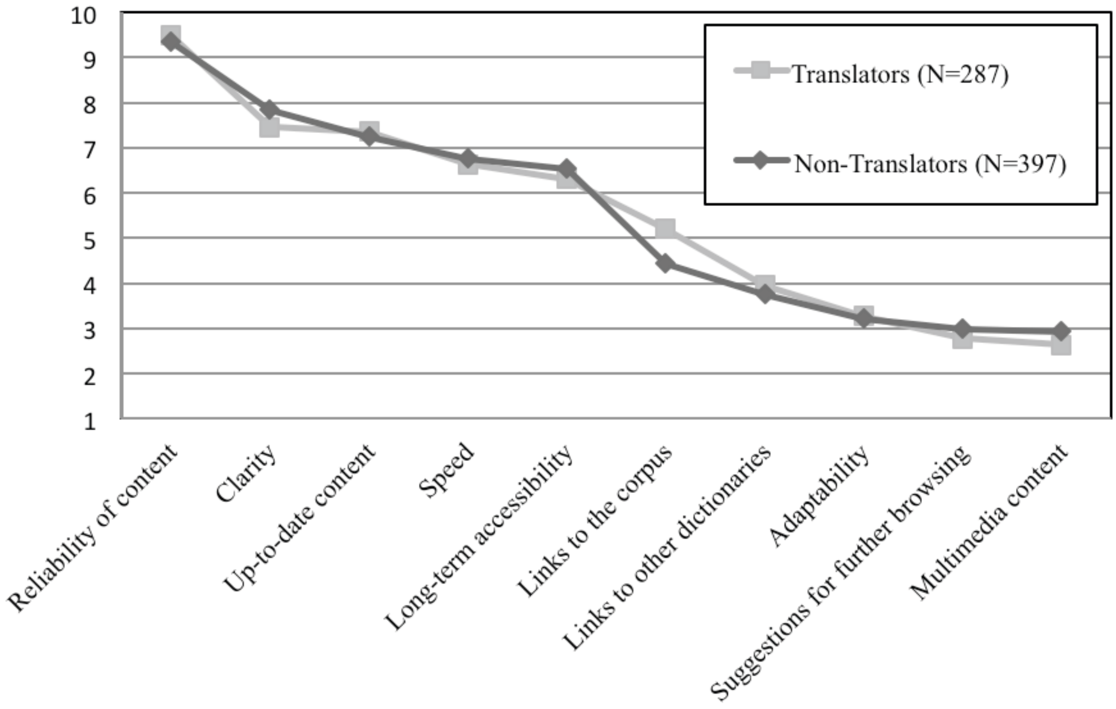


Figure 3. Means of Rankings as a Function of Professional Background (*Translators* vs. *Non-Translators*).

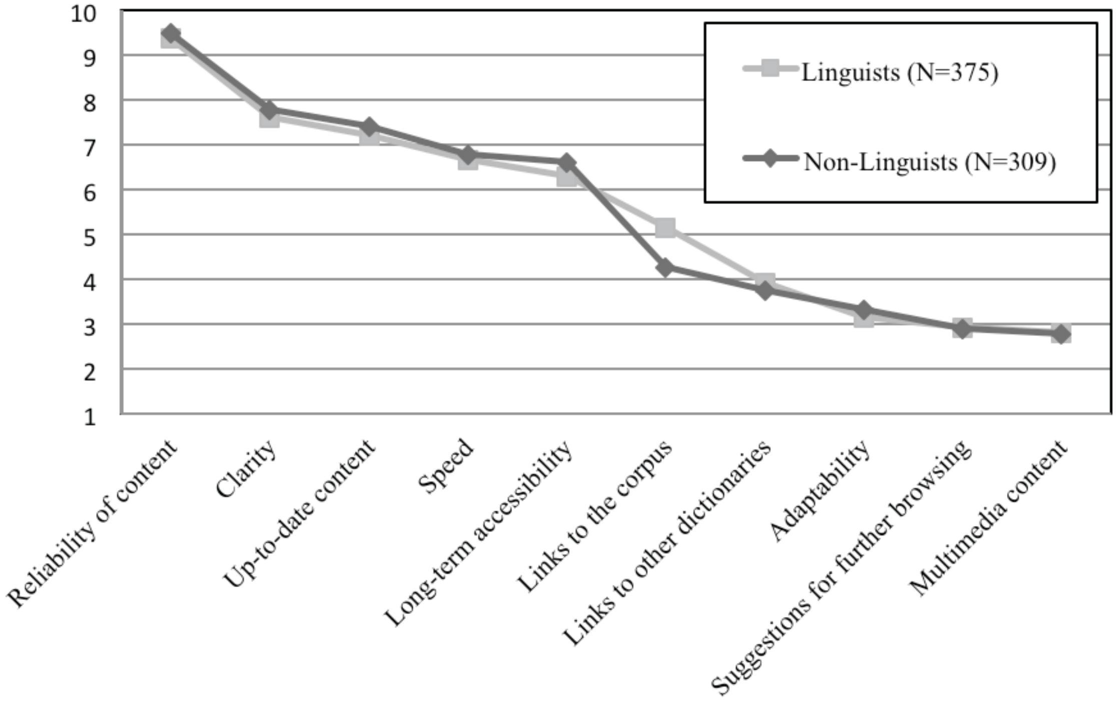


Figure 4. Means of Rankings as a Function of Academic Background (*Linguists* vs. *Non-Linguists*).

Another hypothesis was that we expected translators to rate, on average, a user interface that is adaptable to be more important for an online dictionary than non-translators, since professional translators rely heavily on dictionaries in their daily work. An adaptable user interface could enhance their individual productivity.

4. Method

To identify different user demands, we conducted two online surveys in English and German in 2010. A total of 1,074 respondents participated. Among other questions, respondents in the first survey (N = 684) were asked to rate ten aspects of usability on 5-point Likert scales (1 = not important at all, 5 = very important) regarding the use of an online dictionary (in the questionnaire, all criteria were explained fully): *Adaptability, Clarity, Links to other dictionaries, Links to the corpus, Long-term accessibility, Multimedia content, Reliability of content, Speed, Suggestions for further browsing, Up-to-date content.*

After this, participants were asked to create a personal ranking according to importance. The most important criterion was placed in tenth position, whereas the least important criterion was placed in first position. Furthermore, participants could choose in which language they wanted to complete the questionnaire (English/German) and were asked whether they work as a linguist and/or as a translator (yes/no) in order to analyze whether different users groups have different demands.

5. Results

5.1 Correlation Analysis

Analysis of (Spearman's rank) correlation revealed a significant association between importance and ranking; r = 0.39 [0.20; 0.56]; p < .01. These results indicate that the individual ranking can be used as a reliable indicator of users' demands as intended (cf. fig. 1).

5.2 Subgroup analyses

As mentioned in Section 3, another objective of the study was to assess whether the size of this difference depends on further variables, especially the participants' background (linguistic vs. non-linguistic; translator vs. non-translators) and the language version of the online survey chosen by the participants (German vs. English). Surprisingly, there are no noteworthy rating differences – on average – between different groups, as a visual inspection clearly demonstrates (cf. fig. 2, fig. 3, and fig. 4).

Statistical analyses of variance (not reported here) reveal that some of the differences in average ratings across subgroups are significant. However, this is mainly due to the high number of participants. In fact the F-Value (1, 682) as a test for statistical significance ranges from 0.20 to 59.11 with 8.08 on average, yielding highly significant differences (p < .001) in only 8 out of 30 cases.

Another way of framing these findings is to state that the relative ranking orders represented by the shapes of the curves correspond in each figure except for fig. 2, where a small difference between the two criteria rated on average as least important and second least important occurs. In other words, these results indicate that knowledge of the participant's background allows hardly any conclusions to be drawn about the participant's individual ranking.

5.3 Cluster Analysis

In order to better interpret these results, we conducted a cluster analysis to see how users might group together regarding their individual ranking. A two-cluster solution was identified. Means, standard deviations, and N of each cluster are presented in Table 1.

	Cluster 1 (N = 206)		Cluster 2 (N = 478)	
	M	*SD*	*M*	*SD*
Criterion				
Reliability of content	9.09	1.79	9.54	0.91
Clarity	6.96	1.98	7.97	1.35
Up-to-date content	6.89	2.28	7.45	1.50
Speed	5.52	2.56	7.21	1.47
Long-term accessibility	5.43	2.47	6.86	1.86
Links to the corpus	7.01	1.93	3.77	1.60
Links to other dictionaries	4.72	2.11	3.46	1.47
Adaptability	3.59	2.04	3.08	1.73
Suggestions for further browsing	3.35	2.19	2.64	1.55
Multimedia content	2.43	1.75	3.02	1.89

Table 1. Means and Standard Deviations of Rankings as a Function of the Cluster Analysis

Analyses of variance with the cluster as independent variable and the respective criterion as a response variable yielded highly significant differences (p < .001) for every criterion (10 out of 10 cases) with F (1, 682) ranging from 11.22 to 520.30 (93.08 on average). Most strikingly, only preceded by "Reliability of content", respondents in Cluster 1 rate the criterion "Links to the corpus" on average as the second most important aspect of a good online dictionary (M = 7.01, SD = 1.93), whereas this criterion only plays a minor role for respondents in Cluster 2 (M = 3.77, SD = 1.60), F(1, 682) = 520.30, p < .000 (cf. fig. 5). Taken together, the findings reported here suggest that our initial hypothesis that different groups have different demands was too simple.

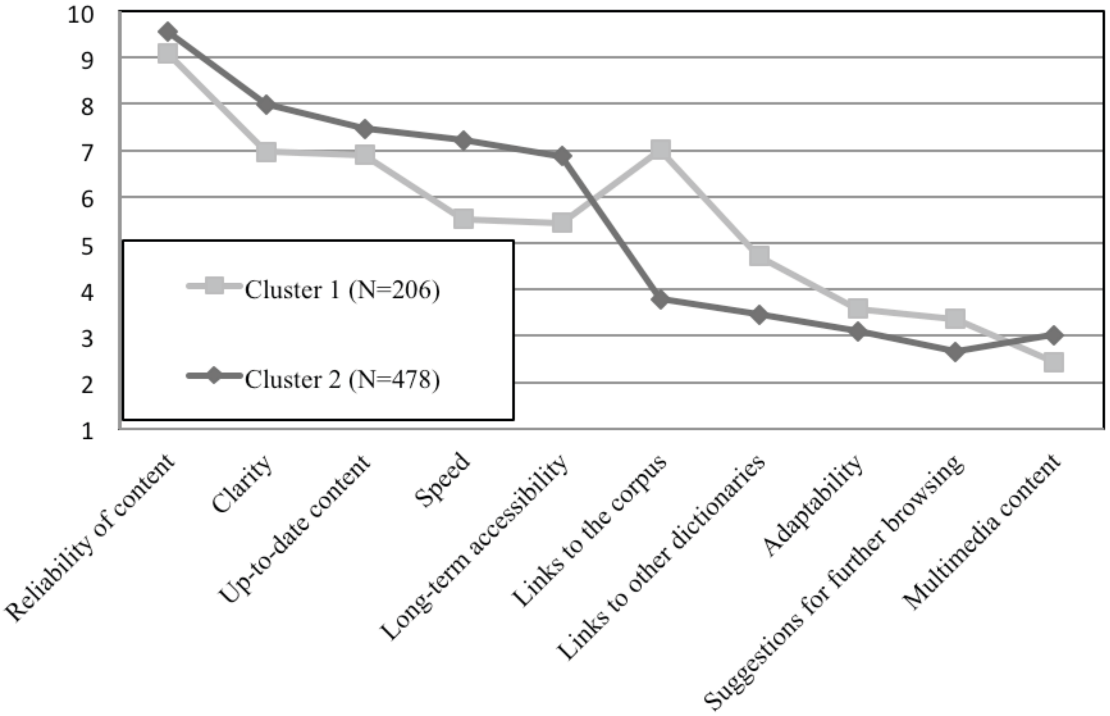


Figure 5. Means of Rankings as a Function of the Cluster Analysis

In Müller-Spitzer/Koplenig (*manuscript in preparation*), we argue that different background variables seem to interact with each other. By using a binary logistic regression model, we show that the probability of belonging to one of the two clusters (as an indicator for sharing similar individual demands regarding the use of an online dictionary) depends on academic background and on professional background and on the language version chosen. Our model indicates, for example, that the probability of belonging to the first cluster (N=206) for subjects in the English language version who work as translators and who have a linguistic academic background is 0.42 (0.95 confidence interval: 0.33 - 0.55), compared to a likelihood of only 0.13 for subjects in the German language version who do not work as translators and who do not have a linguistic background (0.95 confidence interval: 0.08 - 0.21).

6. Discussion

In our study, the classical criteria of reference books (e.g. reliability, clarity) were both ranked and rated highest, whereas the unique characteristics of online dictionaries (e.g. multimedia, adaptability) were rated and ranked as (partly) unimportant.

This result conflicts with the general lexicographical request both for the development of a user-adaptive interface and the incorporation of multimedia elements to make online dictionaries more user-friendly and innovative (e.g., de Schryver, 2003; Müller-Spitzer, 2008; Verlinde & Binon, 2010 present evidence challenging that view).

As is the case for printed dictionaries, our results indicate that online dictionaries are initially being used as a reference work providing reliable and accurate information. The unique characteristics of online dictionaries (e.g. multimedia, adaptability) only seem to play a minor role.

Nevertheless, this does not mean that the development of innovative features of online dictionaries is pointless. As we show elsewhere in detail (Koplenig, 2011; Müller-Spitzer & Koplenig, in preparation), users tend to appreciate good ideas, such as a user-adaptive interface, but they are just not used to online dictionaries incorporating those features. As a result, they have no basis on which to judge the usefulness of those features. Thus, in order to make an online dictionary more user-friendly by implementing innovative features, it is essential that users are also shown the potential benefits of those features.

7. Future Research

The results presented in this paper are still at a preliminary stage. Nevertheless, we believe that they show that both practical lexicography and theoretical lexicology can benefit from this research agenda by shedding some light on an important aspect of dictionary usage in a way that would not be possible through the use of log file analyses alone.

As a next step, to further enhance our understanding of online dictionary usage, we plan to incorporate the insights gained from an eye-tracking study that we have conducted.

8. Acknowledgements

This project is financed by the Gottfried Wilhelm Leibniz Scientific Community ("Joint Initiative for Research and Innovation").

9. References

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

de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2), pp. 143-199.

de Schryver, G.-M., Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams, S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, Lorient, France, July 6th–10th*. Lorient: Université de Bretagne Sud, pp. 187-196.

Engelberg, S., Lemnitzer, L. (2008). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg Verlag.

Koplenig, A. (2011). Understanding How Users Evaluate Innovative Features of Online Dictionaries – An Experimental Approach (Poster). Presented at the *eLexicography in the 21st century: new applications for new users (eLEX2011)*, organized by Trojina, Institute for Applied Slovene Studies, Bled, November 10-12, 2011.

Loucky, J.P. (2005). Combining the Benefits of Electronic and Online Dictionaries with CALL Web Sites to Produce Effective and Enjoyable Vocabulary and Language Learning Lessons. *Computer Assisted Language Learning*, 18(5), pp. 389-416.

Müller-Spitzer, C. (2008). Research on Dictionary Use and the Development of User-Adapted Views. In A. Storrer, A. Geyken, A. Siebert & K.-M. Würzner (eds.) *Text Resources and Lexical Knowledge Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008*. Berlin: de Gruyter, pp. 223-238.

Müller-Spitzer, C., Koplenig, A. (in preparation). *Demands on Online Dictionaries: An Exploration of Group Differences and its Lexicographical Consequences (working title)*.

Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Frankfurt a.M./Berlin/Bern/Bruxelles/NewYork/Oxford/Wien: Peter Lang, pp. 17-32.

Verlinde, S., Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Ljouwert: Afûk, pp. 1144-1151.

Welker, H.A. (2008). Sobre o Uso de Dicionários. *Anais do 8^{o} Encontro do CELSUL*, pp. 1-17.

Wiegand, H.E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Berlin, New York: de Gruyter.

The usage of field labels in English-Spanish bilingual e-dictionaries from the perspective of translators

María Teresa Ortego
University of Valladolid
Campus Duques de Soria s/n, 42004 Soria (Spain)
E-mail: tortego@lesp.uva.es

Abstract

Translation is the vehicle to spread progress among societies that do not share the same language. However, the translation of specialized vocabulary is a problem that translators have to face. The first sources they look to when they do not know an equivalent are general bilingual e-dictionaries, according to previous research. In order to distinguish specialized vocabulary from general vocabulary, dictionaries use different mechanisms and the most prevalent one is field labelling. Our aim is to study how field labels are used to tag specialized vocabulary, so we analyse these field labels in the macrostructure and microstructure of a selection of general bilingual e-dictionaries from a translator's perspective. In the macrostructure of each dictionary, we look for references to the use and selection of field labels, we search for a list of field labels, then we check whether all of these are included in entries, and we extract the most representative fields, having counted the number of entries tagged with field labels. In the microstructure, we find where dictionaries place field labels in the structure of the entry, the typology used, and we test whether dictionaries label the same units by analysing five randomly selected units. Finally, we show the analysis results found in each dictionary, we compare them and we draw conclusions.

Keywords: bilingual e-dictionary; field; field label; specialized vocabulary; translation

1. Introduction

The priority for scientists is to spread progress among societies that do not share the same language, and the way to achieve this is through translation. However, the translation of specialized vocabulary is a problem that translators have to face. The first sources they look to when they do not know an equivalent are general bilingual e-dictionaries, according to previous research (Meyer, 1988; Roberts, 1990; Mackintosh, 1998; Varantola, 1998; Atkins & Varantola, 1998; Corpas et al., 2001). Although it can be thought that bilingual e-dictionaries do not contain specialized vocabulary because they are referred to as "general", they actually include a representative selection of lexical units belonging to different linguistic levels and subsets of language (Haensch, 1997) that a middle class user knows by his or her culture and the influence of the media.

However, translators are users who have some features which differ from other e-dictionary users: they master the source language and the target language, and they are trained in dictionary search. While they are not trained in the field for which they translate, they have the skills in order to learn quickly about the topic of translation.

Having justified the inclusion of specialized lexical units in bilingual e-dictionaries and described the profile of translators as users of general bilingual e-dictionaries, we focus on the most used mechanism to distinguish specialized vocabulary from general vocabulary, that is, field labels. Field labels are very helpful for translators because they show the field to which lexical units belong, especially in lemmas with polysemous meanings, helping translators to choose the correct meaning and hence, the correct equivalent for the context.

Our aim is to study how field labels are used to tag specialized vocabulary in the main English-Spanish general e-dictionaries, and we try to answer the following questions:

1. Do dictionaries explain the use and selection of field labels?
2. How many field labels are used in each dictionary?
3. How many fields are represented? Which are the fields with the highest number of tagged entries?
4. Where are field labels located in each entry of the dictionary?
5. Do dictionaries use the same typology of field labelling?
6. Do dictionaries tag the same specialized vocabulary?

2. Methods

To answer these questions and determinate how field labels are used to tag specialized vocabulary, first of all, we select two general bilingual e-dictionaries. Then, we explain how we are going to analyse information in the macrostructure and microstructure of the selected general bilingual e-dictionaries.

2.1 The selection of general bilingual e-dictionaries

We select two general bilingual e-dictionaries according to the following parameters extracted on previous research about analysis and assessment of dictionaries (Mary Haas, 1964; in Landau, 2001; Cabré & Gelpí Arroyo, 1996; Roberts, 1997; Landau, 2001; Santamaría Pérez, 2003; Gelpí Arroyo, 2003; Atkins & Rundell, 2008): to be an e-dictionary; to mention translators among its users; to be an unabridged dictionary (Roberts: 1997); to distinguish between English and Spanish varieties; to have an intuitive structure so that learning how to use it does not take too much time for users; to be reliable, that is, to be based on corpus during its compilation; and, to be accessible, so that the user can find the dictionary in bookshops and libraries.

The dictionaries chosen which fulfil these criteria are:

-GALLIMBERTI, B. & RUSSELL, R. (eds.): *Gran diccionario Oxford: Español-Inglés, Inglés-Español.* 4th edition. Oxford: Oxford University Press, 2008; from now on *GDO*.

-SCRIBEN, R. et al. (dirs.): *Collins Universal* Español-*Inglés, English-Spanish.*9th edition. Barcelona: Random House Mondadori; Glasgow: Harper Collins Publishers, 2009, from now on *CU*.

Those dictionaries are offline e-dictionaries for PC:

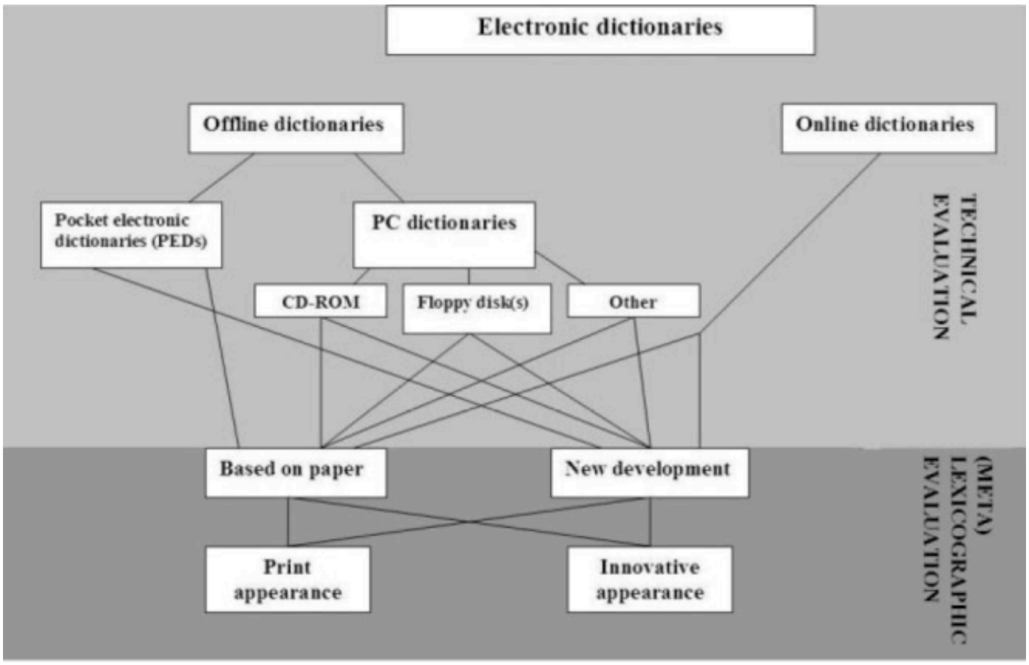


Figure 1: Classification of electronic dictionaries by Lehr (1996, translated in de Schryver, 2003)

CU and *GDO* use an easy interface, they include translators among the range of users to which the dictionary is aimed, and they are divided into two sections, one for understanding (English-Spanish) and the other for production (Spanish-English). In terms of their sizes, *GDO* includes almost 60,000 entries in each section. However, while we have not found any reference to the precise number of entries in *CU*, it is mentioned that more than 750,000 references and translations are included. Both dictionaries tag entries with usage labels to show the place where the unit is used.

	Spanish	English
CU	27	11
GDO	24	6

Table 1: Number of labels used to distinguish dialects.

Both dictionaries use different sizes of letters and different colours to distinguish different types of information, and so are well structured for new users.

Moreover, during the compilation of *CU* and *GDO*, lexicographers based their work on previous editions and on the use of corpora. *CU* is based on the results obtained in *Bank of English*[1] and *Banco de Español,* while *GDO* is based on *Oxford English Corpus*[2] and *Oxford Reading Programme*.

The last parameter, accessibility, is clearly satisfied - according to a study from Corpas Pastor et al. (2001), the

[1] Bank of English is the actual name for COBUILD, compiled by the University of Birmingham. Now it is composed by 550 million words of everyday English from different oral and written texts.

[2] Oxford English Corpus is composed by texts collected from 2000 until now. In 2010, it contained more than two billion words from different sources.

selected dictionaries are the most used dictionaries by translation students from the University of Malaga (Spain).

Having chosen and justified the e-dictionaries in which we are going to analyse field labels, the next step in our research is to describe how we are going to carry out the analysis.

2.2 The analysis of macrostructure and microstructure

In order to answer the proposed questions, we divide the analysis in to two parts. In the first part, we focus on the macrostructure in order to answer questions 1-3. We look for any references to field labelling in the Help section of the e-dictionaries. Then, we look for a list with the number of field labels used in each dictionary. Once we have the list, we test if each field label is used in each section of the selected dictionaries and we compare the results between sections and between dictionaries. Then, we try to find a list of domains and verify if the number of domains used in each section matches up with the list of field labels. Next, we find the domains with the highest number of tagged entries by using the searching options available in each dictionary.

The second part of the analysis is to answer questions 4-6 by looking at the microstructure of the dictionary. We observe how each dictionary uses field labels within the entry: place, typology, abbreviations, etc.

Then, we select five words at random belonging to specialized vocabulary from the most represented fields. We test if they are tagged in each section of the selected dictionaries following the principle of reversibility[3] (Svensen, 2009).

Finally, we compare the results obtained in *CU* and *GDO*.

3. Results

In this section we describe the results obtained from the search in the selected dictionaries in order to find answers to the proposed questions.

3.1 Do dictionaries explain the use and selection of field labels?

As we have mentioned before, references to the use and selection of field labels should be described in the Help section. *CU* explains that field labels are used when the meaning of a word is technical. It also offers a list with every kind of abbreviation and label. However, *GDO* is less explicit than *CU* and does not offer advice on the use of field labels. In spite of the lack of data about field labels, it offers a list with the different labels used. From both lists, we have extracted only field labels.

3.2 How many field labels are used in each dictionary?

From the total list of labels used in each of the selected e-dictionaries, we chose only those which refer to fields. In *CU*, we found 99 abbreviations related to a field which are used to tag specialized vocabulary. Then, we tested if all of them appeared in both sections of the dictionary and the result is that 24 of them were duplicated.

	English-Spanish	Spanish-English
Architecture	Archit	Arquit
Biology	Bio	Biol
Commerce	Comm	Com
Sewing	Sew	Cost
School	Scol	Escol
Pharmacy	Pharm	Farm
Railways	Rail	Ferro
Philosophy	Philos	Fil
Physics	Phys	Fís
Physiology	Physiol	Fisiol
Photography	Phot	Fot
Computers	Comput	Inform
Mathematics	Math	Mat
Mechanics	Mech	Mec
Meteorology	Meteo	Met
Mythology	Myth	Mit
Music	Mus	Mús
Nautical	Naut	Náut
Optics	Opt	Ópt
Psychology	Psych	Psic
Chemistry	Chem	Quím
Theatre	Theat	Teat
Technical	Tech	Téc
Typography	Typ	Tip

Table 2: Fields with two labels in *CU.*

For example, to tag a specialized lexical unit from Computers, *CU* uses the label (Comput) in the English-Spanish section and the label (Inform) in the Spanish-English section.

GDO offers a list of abbreviations from which we extracted field labels. However, the microstructure of the dictionary does not use any abbreviations. Instead, it uses the name of the domain in English for the English-Spanish section (Ex: Medicine) and in Spanish for the Spanish-English section (Ex: Medicina). In all, *GDO* uses 188 field labels, 94 labels in each section.

3.3 How many fields are represented? Which are the fields with the highest number of tagged entries?

Although one might think that the number of field labels and fields represented in each section of the dictionary would match up, we found that the number of fields

[3] The principle of reversibility means that the equivalent must be included in the other section of the dictionary. For example, if we look up *exit* in English-Spanish section, the equivalent *salida* should be included as an entry in Spanish-English section and the equivalent proposed should be *exit*.

represented in *CU* rises to 75 whereas the number of labels is 94.

We found that if all the fields tag any entry in both sections of *CU,* the result is that the following fields are not included:

ENGLISH-SPANISH		SPANISH ENGLISH	
FIELD	LABEL	FIELD	LABEL
Biology	Bio	Stock Exchange	St Ex
Science	Sci	Science	Sci
Sport	Dep	Ecology	Ecol
Bullfighting	Taur	Skiing	Ski
		Government	Govt
		Industry	Ind
		Radio	Rad

Table 3: Fields which are not represented by sections.

So according to the figures, *CU* includes 71 fields in the English-Spanish section and 66 fields in the Spanish-English section.

In *GDO,* only 89 fields are used to tag entries in the English-Spanish section: Arms, Entertainment, Printing and Publishing, Bullfighting and Wine are not included. The Spanish-English section only contains 81 fields. We were unable to find lexical units tagged with Anthropology, Post, Railways, Nuclear Physics, Printing, Publishing, Civil Engineering, Electric Engineering, Chemist Engineering, Mechanics, Occultism, Labour Relations and Tourism.

The second part of the question is to find the most represented fields. This question is complicated because the search engine in *CU* does not accept brackets in the search options nor recognizes the difference between upper and lowercase letters. So, we had to count the lexical units tagged with field labels individually.

ENGLISH-SPANISH		SPANISH-ENGLISH	
FIELD	ENTRIES	FIELD	ENTRIES
Medicine	588	Medicine	861
Military	571	Military	688
Computers	465	Commerce	621
Commerce	416	Law	595
Law	406	Sport	568
Automobiles	405	Politics	565
Music	362	Religion	537
Politics	355	Nautics	535
Nautics	332	Technical	519
Economy	320	Botany	469

Table 4: The most representative fields by sections in *CU.*

The same procedure was carried out in *GD*O. However, it was easier than in *CU* because the search engine accepts brackets and capital letters.

In the following table we offer a synthesis of the most represented fields in *GDO*:

	ENGLISH-SPANISH		SPANISH-ENGLISH	
	Field	Entries	Field	Entries
1	Sport	534	Medicine	417
2	Law	445	Law	394
3	Computing	437	Sport	387
4	Medicine	424	Zoology	343
5	Military	411	Cookery	336
6	Music	367	History	331
7	Linguistics	329	Religion	302
8	Cookery	320	Military	280
9	Religion	301	Music	271
10	Finance	281	Botany Computing	265

Table 5: The most represented fields by sections in *GDO*.

3.4 Where are field labels located in each entry of the dictionary?

The location of field labels depends on the meaning of the lexical units they label. For monosemous lexical units, field labels are placed after the spelling, pronunciation and grammar category in both dictionaries. For polysemous lexical units, field labels are placed after the number of letters which indicates which sense is specialized and before the equivalent.

3.5 Do dictionaries use the same typology of field labelling?

CU uses abbreviations between brackets, in italics and a blue colour. *GDO* uses the name of the field in English in the English-Spanish section and in Spanish in the Spanish-English section, between brackets and the first letter in uppercase. It also uses a blue colour.

3.6 Do dictionaries tag the same specialized vocabulary?

From the list of the most represented fields, we have randomly selected five lexical units:

FIELD	LEXICAL UNIT
Medicine	pacemaker
Military	tank
Sport	defender
Law	bailiff
Computers	flash memory

Table 6: Randomly selected units.

We searched those lexical units in the English-Spanish section of each selected e-dictionary and then we tested if the selected dictionaries followed the principle of reversibility. The results of searching lexical units in English-Spanish sections and the equivalents proposed as entries in the Spanish-English section of both dictionaries are shown in the following tables:

ENGLISH-SPANISH		SPANISH-ENGLISH	
Lemma	Equivalent	Lemma	Equivalent
Pacemaker	(Med) marcapasos	Marcapasos	Pacemaker (Sin arca)
Tank	(Mil) tanque, carro (de combate)	Tanque	(Mil) tank
		carro	(Mil) tank
defender	(Sport) defensa	Defensa	(Dep) la defensa (= jugadores) the defence, the defense (EEUU)
bailiff	(Jur) alguacil	Alguacil	(Jur) bailiff, constable
Flash memory	Memoria flash	Memoria flash	

Table 7: Results in *CU.*

ENGLISH-SPANISH		SPANISH-ENGLISH	
pacemaker	(Medicine) marcapasos	Marcapasos	Pacemaker (sin marca)
tank	(Military) tanque, carro de combate	Tanque	(Armas) (carro) tank
		Carro de combate	Tank (sin marca)
defender	(Sport) defensa	defensa	(Deporte) (conjunto) defense* Defensa (jugador) defender
Bailiff	(Law) (in UK) alguacil (in US) funcionario que custodia al acusado en un juzgado	Alguacil	(oficial) bailiff (sin marca)
Flash memory	(Computing) memoria flash	Memoria flash	Flash memory

Table 8: Results in *GDO.*

From the previous table, we observe that only *tank-tanque* and *defender-defensa* in *CU* and *GDO* are tagged with field labels in both sections. Moreover, only *bailiff-alguacil* are tagged with field labels in *GDO*.
Then, we have found two phenomena which affect other units. Firstly, units are described with field labels in one section and without field labels in the other section. For example in the entry of *pacemaker*, the equivalent *marcapasos* is tagged with (Med) and (Medicine) in both dictionaries, whereas the entry *marcapasos* in the Spanish-English section doesn't have any field labels. Secondly, some equivalents in English-Spanish are not included as entries in the nomenclature of the Spanish-English section. For example, the equivalent of *flash memory, memoria flash*, is not included in the nomenclature of the Spanish-English section.

4. Conclusions

In conclusion, we find that labelling specialized vocabulary in general bilingual e-dictionaries is not systematic.

In the macrostructure, *GDO* does not describe the procedure followed by compilers to select and tag specialized vocabulary and to insert it into the nomenclature. Furthermore, neither of the selected dictionaries offers a list composed by field labels used. The lists consulted include all type of labels, so translators and other users find hard to get used to labels, especially when they have to use two or more dictionaries.

The number of field labels used in the selected dictionaries does not match up. This phenomenon also occurs with the fields listed in the dictionaries. In addition, the number of field labels does not fit between sections in the same dictionary.

So we reach to the conclusion that dictionaries from the same size and category do not include the same proportion of specialized vocabulary, and the results reached by a translator will depend on the dictionary used to aid translation.

In the microstructure, the place that field labels take in the entry is the same in both dictionaries, although the typography changes. *CU* prefers the use of abbreviations while *GDO* uses the name of the field between brackets. It would be useful to normalize the typography of field labels and the fields used in dictionaries of same typology in order to make searches easier for translators.
Moreover, labelling of specialized vocabulary is not systematic. If we look at the results of the comparison of five randomly selected lexical units, we observe that they are labelled in one section of the dictionary, but the equivalents are not labelled in the other section or even equivalents are not included in the nomenclature.

To sum up, we will continue with some more studies into specialized lexical units in the framework of our PhD project. Meanwhile, we would like to ask compilers and editors to normalize the use of field labels in general bilingual e-dictionaries, to be more systematic in the labelling of equivalents in other sections of the dictionary, and to be more careful in field labelling in order to save users time and to save translators time in their searching tasks. In addition, dictionaries should be more systematic and mark the same units equally in both sections.

5. Acknowledgements

I would like to thank the financial funds from University of Valladolid (Spain) to assist to e-LEX2011.

6. References

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

Atkins, B.T.S., Varantola, K. (1998). Monitoring dictionary use. In B.T.S. Atkins (ed.) *Using dictionaries: Studies of dictionary use by language learners and translators*. Tübingen: Max Niemeyer, pp. 83-122.

Cabré, M.T. & Gelpí Arroyo, C. (1996). La lexicographie bilingue catalane contemporaine: analyse et évaluation. In H. Béjoint, P. Thoiron (eds.), *Les Dictionnaires bilingues*. Louvain-la Neuve: Duculot, pp. 213-230.

Corpas Pastor, G., Leiva Rojo, J. & Varela Salinas, M.J. (2001). El papel del diccionario en la formación de traductores e intérpretes: análisis de necesidades y encuestas de uso. In M.C. Alaya Castro (ed.) *Diccionarios y enseñanza*. Alcalá de Henares: Universidad de Alcalá, pp. 239-273.

De Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2), pp. 143-199.

Gelpí Arroyo, C. (2003). Mesures d'avaluació lexicogràfica de diccionaris bilingües. PhD thesis. Barcelona: Universitat Pompeu Fabra.

Haensch, G. (1997). *Los diccionarios del español en el umbral del siglo XXI: Problemas actuales de la lexicografía.* Salamanca: Universidad de Salamanca.

Landau, S.I. (2001). *Dictionaries: The art and craft of lexicography*. Cambridge: Cambridge University Press.

Mackintosh, K. (1998). An empirical study of dictionary use in L2-L1 translation. In B.T.S. Atkins (ed.) *Using dictionaries: Studies of dictionary use by language learners and translators*. Tübingen: Max Niemeyer, pp. 123-149.

Meyer, I. (1988). The general bilingual dictionary as a working tool in" thème". *Meta,* 33(3), pp. 368-376.

Roberts, R. (1990). Translation and the bilingual dictionary. *Meta,* 35(1), pp. 74-81.

Roberts, R. (1997). Using dictionaries efficiently. *38th Annual Conference of the American Translators Association,* San Francisco, California.

Santamaría Pérez, M.I. (2003). *La fraseología española en el diccionario bilingüe español-catalán: Aplicaciones y contrastes*. Alicante: Universidad de Alicante.

Svensen, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge: Cambridge University Press.

Varantola, K. (1998). Translators and their use of dictionaries: User needs and user habits. In B.T.S. Atkins (ed.) *Using dictionaries: Studies of dictionary use by language learners and translators*. Tübingen: Max Niemeyer, pp. 179-192.

Interactive, dynamic electronic dictionaries for text production

D.J. Prinsloo*, Ulrich Heid, Theo Bothma***, Gertrud Faaß****
* Department of African Languages and *** Department of Information Science,
University of Pretoria, Pretoria 0002, South Africa
** Department of Information Science and Natural Language Processing,
Hildesheim University, Marienburger Platz 22, 31141 Hildesheim, Germany and
Department of African Languages, University of Pretoria
E-mail: danie.prinsloo@up.ac.za, heid@uni-hildesheim.de, theo.bothma@up.ac.za, gertrud.faass@uni-hildesheim.de

Abstract

An interactive, dynamic electronic dictionary aimed at text production should guide the user in innovative ways, especially in respect of difficult, complicated or confusing issues. This paper proposes a design for bilingual dictionaries intended to guide users in text production; we focus on complex phenomena of the interaction between lexis and grammar. It will be argued that a dictionary aimed at guiding the user in lexical selection should implement a type of "decision algorithm". In addition, it should flag incorrect solutions and should warn against possible wrong generalisations of (foreign) language learners. Our proposals will be illustrated with examples from several languages, as the design principles are generally applicable. The copulative construction which is regarded as the most complicated grammatical structure in Northern Sotho will be analyzed in more detail and presented as a case in point.

Keywords: bilingual electronic dictionaries; user guidance; text production; dictionary design

1. Introduction

The electronic era was met with great enthusiasm and expectations. Early publications on electronic dictionaries were all about the potential of the new medium and the expected revolution it would bring along, thereby antiquating the paper dictionary in a decade or two. De Schryver (2009), however, rightfully expresses disappointment in respect of the pace of development of electronic dictionaries. More exciting was the introduction of what could be called "true electronic features" such as pop-up boxes, alternative access routes to the data, audible pronunciation and sophisticated search features. Some electronic dictionaries also solve problems in respect of lemmatisation, which cannot be resolved in paper dictionaries. Electronic dictionaries of today, however, could enter a more advanced dimension in fulfilling more sophisticated needs of the users, e.g. if access to data were not only based on a single lemma. Rundell (2009:9) refers to "game changing" developments that have "expanded the scope of what dictionaries can do and (in some respects) changed our view of what dictionaries are for". De Schryver (2009) calls in this context for an adaptive and intelligent dictionary (aiLEX) that will be able to "study and understand its user" and consequently to "present itself to that user". In most cases what is currently offered in dictionaries claiming that they give guidance in text production is in fact still on the level of text reception, and they generally give an overload of information. An interactive and dynamic electronic dictionary aimed at text production should guide the user in innovative ways, especially in respect of difficult, complicated or confusing issues. The underlying lexicographic concepts remain the same. What is at stake here are improvements in the article structure and access possibilities of electronic dictionaries.

2. Phenomena and proposals for their presentation

This paper proposes a design for bilingual dictionaries intended to guide users in text production; we focus on complex phenomena of the interaction between lexis and grammar. Our proposals can be illustrated with examples from several languages, as the design principles are generally applicable. The complex morpho-syntactic phenomena of the South African Bantu languages do particularly require a design of the proposed kind. Adaptivity to individual users (in De Schryver's (2009) sense) is not the main focus of this paper. We assume fixed user profiles for novice and expert users, and task profiles of text production and text reception. Nevertheless, our design allows for more flexibility beyond this simplistic parameterization.

Lexical selection in text production can be seen as a decision process. Grammar rules, semantics and communicative intentions, as well as (idiosyncratic, lexicalised) exceptions are among the parameters that influence the choice. Very often these rules are so complex and/or comprehensive that the average user of a dictionary or a grammar text does not (immediately) understand the rules that are being explained, or is simply overwhelmed with the amount of information presented. It is proposed that a dictionary tool is needed to simplify the decision process for the user and/or reduce the amount of information presented to the user to exactly what is needed to address the user's information need. A dictionary aimed at guiding the user in lexical selection should therefore implement a type of "decision algorithm". In addition, it should flag incorrect solutions and should warn against possible wrong generalisations of (foreign) language learners. As it stands in current sources on, for example, Northern Sotho copulatives in dictionaries and grammar books, the guidance given could be regarded as cognitive aids. Our aim is to address the complexity by moving from the cognitive to text production by means of a selection process. This then also constitutes the rational for linking the dictionary with corpus data.

As a first prerequisite, this type of interactive, dynamic electronic dictionary should guide the user to the production of correct text. Prinsloo (2002) states the role of the lexicographer in this regard as a *mediator* between a complicated linguistic issue on the one hand and the dictionary user on the other, cf. also Tarp's (2011) idea of dictionaries as tools.

Text production support can be at different levels of complexity, for example:

- A simple decision algorithm (decision tree) based on one or two variables, illustrated by means of example sentences with limited additional explanation (available on demand).
- A situation where the grammatical rules are highly complex and follow a complex decision algorithm based on multiple variables, for example, "if *a* then *b* or *c*; if *b* then *d*, but if *c* then *e*, etc.".

Examples of the two levels of complexity will be described below. The first two examples reflect a very simple situation and the third a highly complex one. There are obviously multiple levels of complexity, and the above two reflect the extremes – all such support situations can be plotted on a continuum of complexity, each with its unique type of solution. Each decision tree (with its accompanying explanatory text and number of examples) depends on the nature of the data and the nature of the complexity of the problem.

An example from text understanding is homographic forms with different grammatical functions or meanings.

A case in point is Afrikaans *sy* which can be a personal pronoun (cf. (1)) or a possessive (cf. (2)). The decision algorithm is based on the context: the user verifies the presence of verbal governors (then *sy* is a feminine personal pronoun) or adjacent nominals (then *sy* is always and only a masculine possessive determiner).

(1) Sy het die boeke gekoop
She has the books bought
(She bought the books)
(2) Sy boek
His book

In the above case a simple decision algorithm and a few example sentences followed by a brief explanation should be sufficient to help the user to select the correct interpretation in a text understanding situation, or the correct equivalent in translation from Afrikaans.

Possessive determiners are also a major problem in beginners' text production, e.g. for English speakers learning a Romance language (our examples are in French): while English has different forms depending on the natural gender of the possessor (cf. *his* (masc.) vs. *her* (fem.)), French possessives agree with the grammatical gender of the possessed object, but don't mark the natural gender of the possessor, cf. (3).

(3) son livre (masc.) ("his/her book")
sa famille (fem.) ("his/her family")
ses livres/familles (plural)
("his/her books/families")

The decision algorithm for the selection of possessives thus has to ask for other parameters (number, gender) in French than in English or Afrikaans. Text production support for French possessives therefore requires a different decision algorithm than the above Afrikaans example, but should also be accompanied by a brief grammatical explanation and examples.

As a third example consider the user who wishes to express the basic copulative concepts *is*, *am* and *are* in Northern Sotho (Sepedi), a Bantu language spoken in South Africa. This is a very complex grammatical problem and therefore requires a more complex decision algorithm with multiple variables for text production support. In this case the decision algorithm for the selection of copulatives entails distinguishing between an *identifying* vs. a *descriptive* vs. an *associative* relation existing between the subject and its complement as in (4):

(4)

is
[identifying. copulative], ke lengwalo **(**it is a letter)

[descriptive. copulative], mosadi o bohlale
(the woman is clever)

[associative copulative], Satsope o na le Sara
(Satsope is with Sara)

Learners of Northern Sotho who want to use copulatives in speech or text production have at best to do intensive study of the copulatives from dictionaries and grammar books. Dictionaries typically provide basic and sometimes inadequate information. Grammar books such as Poulos and Louwrens (1994), on the other hand, provide an overdose (37 pages) of grammatical information, in a desperate effort to cover all the relevant and possible copulatives. Such details may be useful in a cognitive situation where the user would like to learn everything about the copulative, but they are hardly useful in a text production situation where the user simply wants guidance on which form to use. Such information overload could easily lead to "information death" (cf. Bergenholtz & Bothma, 2011). Compare the following extract from their summary of the identifying copulative:

The identifying copulative

The indicative series *The present tense Principal Identifying* pos. lst and 2nd persons: ***SC - CB*** Classes: ***CP - CB*** neg. 1st and 2nd persons: ***ga - SC - CB*** Classes: ***ga - se - CB*** *Participial* pos. 1st and 2nd
person: ***SC - le - CB*** Classes: ***CP - le - CB*** neg. lst and 2nd person: ***SC - se - CB*** Classes: ***CP - se - CB***

The Lemmatization of Copulatives in Northern Sotho 27

The future tense Principal pos. 1st and 2nd person: ***SC - tlô/tla - ba + CB*** Classes: ***CP - tlô/tla - ba +***
CB neg. 1st and 2nd person: ***SC - ka - se - bê + CB SC*** Classes: ***CP - ka - se - bê + CB*** *Participial* pos.
1st and 2nd person: ***SC - tlô/tla - ba + CB*** Classes: ***CP - tlô/tla - ba + CB*** neg. 1st and 2nd person:
SC - ka - se - bê + CB Classes: ***CP - ka - se - bê + CB*** *The past tense Principal* pos. 1st and 2nd person:
SC - bilê + CB Classes: ***CP - bilê + CB*** neg. 1st and 2nd person: ***ga - se - SC - be + CB ga - se - SC2 -***
a - ba + CB ga - SC2 - a - ba + CB Classes: ***ga - se - CP - bê + CB ga - se - SC2 - a - ba + CB1 ga -***
SC2 - a - ba - CB *Participial* pos. lst and 2nd person: ***SC - bilê + CB*** Classes: ***CP - bilê + CB*** neg. lst
and 2nd person: ***SC - sa - ba + CB*** Classes: ***CP - sa - ba + CB***

The potential *Principal and participial* lst and 2nd person: pos. ***SC - ka - ba + C*** neg. ***SC - ka - se -***
bê + CB Classes: pos. ***CP - ka - ba + CB*** neg. ***CP - ka - sê***
- bê + CB

The subjunctive 1st and 2nd person: pos. ***SC - bê + CB*** neg. ***SC - se - bê + CB*** Classes: pos. ***CP - bê +***
CB neg. ***CP - se - bê + CB*** Note also the compound negative ***SC/CP - se - kê + SC2 - a - ba + CB***

The consecutive lst and 2nd person: pos. ***SC2 - a - ba + CB*** neg. ***SC2 - a - se - bê + CB*** Classes: pos.
SC2 - a - ba + CB neg. ***SC2 - a - se - bê + CB*** Note also the compound negative ***SC2 - a - se - ke +***
SC2 - a - ba + CB

The habitual 1st and 2nd person: pos. ***SC - be + CB*** neg. ***SC - se - be + CB - be + CB*** Classes pos.
CP - be + CB neg. ***CP - se - be + CB***

The infinitive pos. ***go - ba + CB*** neg. ***go - se - bê + CB***

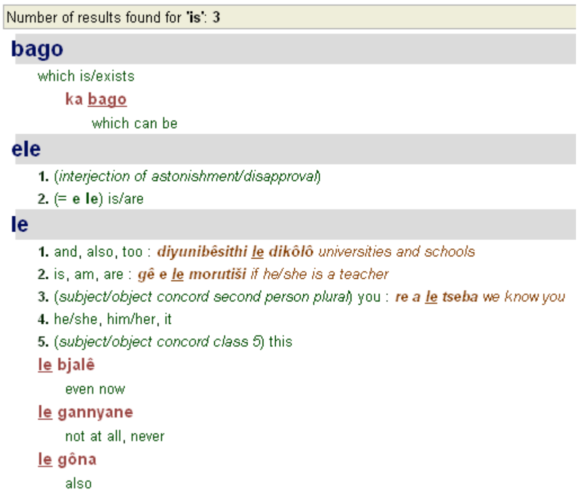
The imperative pos. ***e - ba - ng + CB*** or ***ba - a - ng + CB*** neg. ***se - bê - ng + CB***

(Poulos and Louwrens1994:320)

Dictionaries, and especially electronic dictionaries, fail to give even basic receptive guidance or to treat the three main copulative relations in (4). Consider the article for the lemma **is** in the *Sesotho sa Leboa (Northern Sotho) - English Dictionary* (2003) in Figure 1.

In this example two of the three copulative categories, i.e., the identifying and associative copulatives, have not been treated, not to mention giving proper receptive or productive guidance. Paper dictionaries for Northern Sotho reflect the same deficiencies.

In the e-environment it is, however, possible to provide the user with the required guidance on which form is the correct one for a given situation, and to provide exactly the amount of information that is needed for each of the possible choices. In such a case a decision tree will reduce the amount of information considerably and the user can, at any stage, decide that his/her information need has been met and return to his/her primary task, namely to write a text.



Number of results found for **'is'**: 3

bago

which is/exists

ka bago

which can be

ele

1. (*interjection of astonishment/disapproval*)

2. (= **e le**) is/are

le

1. and, also, too : ***diyunibêsithi le dikôlô*** *universities and schools*

2. is, am, are : ***gê e le morutiši*** *if he/she is a teacher*

3. (*subject/object concord second person plural*) you : ***re a le tseba*** *we know you*

4. he/she, him/her, it

5. (*subject/object concord class 5*) this

le bjalê

even now

le gannyane

not at all, never

le gôna

also

Figure 1: The lemma **is** in the *Sesotho sa Leboa (Northern Sotho) - English Dictionary* (2003)

For example, when the user wants to write “the woman is clever” in Northern Sotho he/she should be guided to *mosadi o bohlale* and guarded from the typical error **mosadi ke bohlale*. The user can then be guided to subsequent levels of decisions, e.g. concerning person and noun class of the subject, tenses and moods, as well as a number of lexicalised exceptions, cf. Appendix 1.

The phenomena sketched above may usefully be presented to the user in terms of subsequent choices, e.g. by means of check boxes, radio buttons, etc. The visual appearance of the interface should make clear that the selections are the result of a decision process involving several steps. Instead of complex tables giving all options, a path through sub-tables should be shown, but together with links to synoptic tables which indeed allow the user to see the full picture if he/she wishes to. For a set of function words of the same category, the basic decision tree is constant. Users will only follow different paths through this tree, depending on their actual needs.

The internal representation of the data should be adapted to the particularized decision-tree-like access to the data. For this, not only synoptic tables of function words, but also a representation of the selection rules is needed, e.g. by means of linked templates.

A number of interface solutions should be considered:

- Just solve the problem, suggest the correct solution and give a visual presentation and link to ‘read more’ sections such as FAQs or outer texts.

- Supply a link to *read more* information where distinctions on a cognitive level are made.
- Supply a link to guidance on the basis of e.g. *frequently made errors*.
- Give good, typical examples of use throughout.

All envisaged actions should be based upon a grammatical description of the construction to be tackled e.g. pronouns in Afrikaans, English, French or the copulative construction in Northern Sotho. One could argue that these issues have been sufficiently described in standard grammars of these languages. However, one should not assume that the format of these descriptions is such that they are ready to use for our purposes. A reorganization of the data will be necessary.

The process to produce such a dictionary article requires at least three sequential steps, building on one another:

- Step 1 would be to acquire comprehensive and accurate data for the set of rules etc. to be described. This includes the grammatical rules as well as pertinent examples, common errors, etc.
- In Step 2 the lexicographer in collaboration with a database expert needs to reorganise the data so that it will be possible for a programmer to implement a decision tree. This requires at least two processes:
    - The logic of the decision process needs to be worked out very carefully, i.e., what is the logical sequence of the decisions, how much information is required to make and/or support the decisions, when are what type of examples needed, when are links to outer texts required, etc.
    - The data need to be marked up in such a way that each of the data elements defined in the analysis of a specific complex problem can be identified at the required level of granularity. This implies that the database should make provision for such extensions, either by using an extensible XML schema or additional tables and fields in a relational database (depending on the original design of the system), (cf. Bothma (2011)).
- In Step 3 the programmer takes the flow diagram of the decision tree together with all the explanations, examples and linked data, and implements this. The programmer should also design a "user-friendly" interface that is intuitive for the average user and supports him/her to follow the correct trail through the decision tree for the given information need.

3. Exemplification: complex cases of copulative selection

In a text production situation a user can consult the dictionary as an external source to obtain the required information. However, it is also possible that the support the user requires be integrated into a word processor the user is using to construct his/her text. In such a case the user may require feedback on his/her own text production efforts based on his/her grammatical knowledge without specifically consulting the dictionary. In such a case the e-dictionary could be integrated into the word processor as a grammar checker, similar to the features currently available in popular word processing software.

Let us depart from a most common error scenario in Northern Sotho, for example, the user typing **lesogana ke bohlale*. Learners usually know that *ke* means 'it is' and no distinction is made between *he is*, *she is*, *they are* and *it is* in Northern Sotho: all convert to *it is*, e.g. *(monna) ke morutisi '* he is (it is) a teacher'. As a second example consider **monna o morutiši* instead of *monna ke morutiši* 'The man is a teacher'. Learners are accustomed to using the subject concord *o* with class 1 nouns in sentence construction and it is the correct form in two out of the 3 copulative relations (descriptive and associative copulatives: so attempting to use it also in the identifying copulative is a common error).

The student types **lesogana ke bohlale* in a word processor linked to the electronic dictionary and all three words are or only the *ke* is flagged as incorrect. A quick solution is offered by means of a suggestion box, in this case offering three possibilities namely *le*, *e le* 'is/am/are' and *e lego* 'who/what is/am/are'. The user who has basic knowledge of the modal system will know which one to select. Most users, however, would need further guidance and this is offered by a decision process guiding him/her through the three possible moods (Indicative *le*, Situative *e le* or Relative *e lego*) of the decision tree for the descriptive copulative with sub-decisions. The process for **monna o morutisi* is similar, i.e. a decision process guiding him/her through the three possible moods (Indicative *ke*, Situative *e le* or Relative *e lego*) of the decision tree for the identifying copulative respectively, with sub-decisions.

3.1 Different levels of user guidance

Figure 2 provides a schematic illustration of a pop-up guidance screen sequence for **lesogana ke bohlale.*

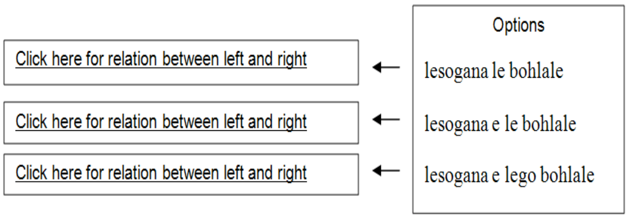
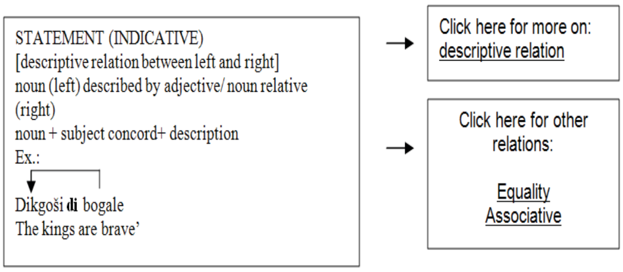

Options
Click here for relation between left and right
lesogana le bohlale
Click here for relation between left and right
lesogana e le bohlale
Click here for relation between left and right
lesogana e lego bohlale

Figure 2: Dictionary feedback for **lesogana ke bohlale*

If more guidance in respect of the descriptive relations in the Indicative, Situative and Relative is required, the user can click the buttons in Figure 2 to display the information given in Figures 3, 4 and 5.

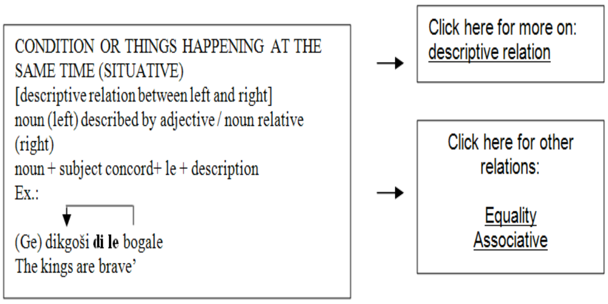


STATEMENT (INDICATIVE)
[descriptive relation between left and right]
noun (left) described by adjective/ noun relative (right)
noun + subject concord+ description
Ex.:
Dikgoši **di** bogale
The kings are brave'

→ Click here for more on: descriptive relation

→ Click here for other relations:
Equality
Associative

Figure 3: Pop-up 2a: Information boxes for *lesogana le bohlale* in Level 1

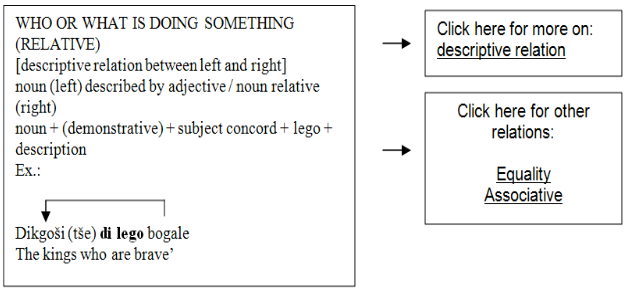


CONDITION OR THINGS HAPPENING AT THE SAME TIME (SITUATIVE)
[descriptive relation between left and right]
noun (left) described by adjective / noun relative (right)
noun + subject concord+ le + description
Ex.:
(Ge) dikgoši **di le** bogale
The kings are brave'

→ Click here for more on: descriptive relation

→ Click here for other relations:
Equality
Associative

Figure 4: Pop-up 2b: Information boxes for *lesogana e le bohlale* in Level 1



WHO OR WHAT IS DOING SOMETHING (RELATIVE)
[descriptive relation between left and right]
noun (left) described by adjective / noun relative (right)
noun + (demonstrative) + subject concord + lego + description
Ex.:
Dikgoši (tše) **di lego** bogale
The kings who are brave'

→ Click here for more on: descriptive relation

→ Click here for other relations:
Equality
Associative

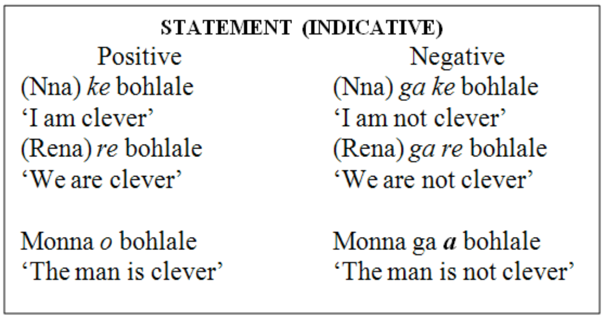
Figure 5: Pop-up 2c: Information boxes for *lesogana e lego bohlale* in Level 1

In each case, the panel given in the left part of the mock-up provides the information needed for text production. Users with more (cognitive) needs can access a fuller picture via the buttons on the right hand side.

3.2 From text production guidance to full grammatical guidance

Pop-up boxes giving more information and typical examples of descriptive relations can be provided on a third level for the Indicative, Situative and Relative. See, for example, additional information for the Indicative in Figure 6.

A second scenario is where comprehensive guidance is required, e.g. when the user wants to know how to say *is* in Northern Sotho. In this case a combination of decision processes is required. These processes are enriched with information from corpora and processed corpus data linked with the dictionary.



STATEMENT (INDICATIVE)

Positive	Negative
(Nna) *ke* bohlale 'I am clever'	(Nna) *ga ke* bohlale 'I am not clever'
(Rena) *re* bohlale 'We are clever'	(Rena) *ga re* bohlale 'We are not clever'
Monna *o* bohlale 'The man is clever'	Monna ga ***a*** bohlale 'The man is not clever'

Figure 6: Pop-up 3a: Information boxes for *descriptive relation* in level 2

4. Conclusion

The project described above is driven by two underlying motivations, namely the urge to compile electronic dictionaries that can do better than current ones through maximal utilization of advanced modern technologies and the need for intelligent and dynamic dictionaries guiding the user in new innovative ways. We believe that step-by-step guidance, mainly through sequences of choices, the provision of additional relevant information on request as well as protection against incorrect conclusions, are the cornerstones of the design of such intelligent dictionaries.

5. References

Bergenholtz, H., Bothma, T.J.D. (2011). Needs-adapted data presentation in e-information tools. *Lexikos*, in press.

Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. 2011. London & New York: Continuum, pp. 71-102.

De Schryver, G-M. (2009). State-of-the-Art Software to Support Intelligent Lexicography. In R. Zhu (ed.) (2009) *Proceedings of the International Seminar on Kangxi Dictionary & Lexicology*. Beijing: Beijing Normal University, pp. 565–580. Also: http://www.hcxf.cn/read.asp?id=570.

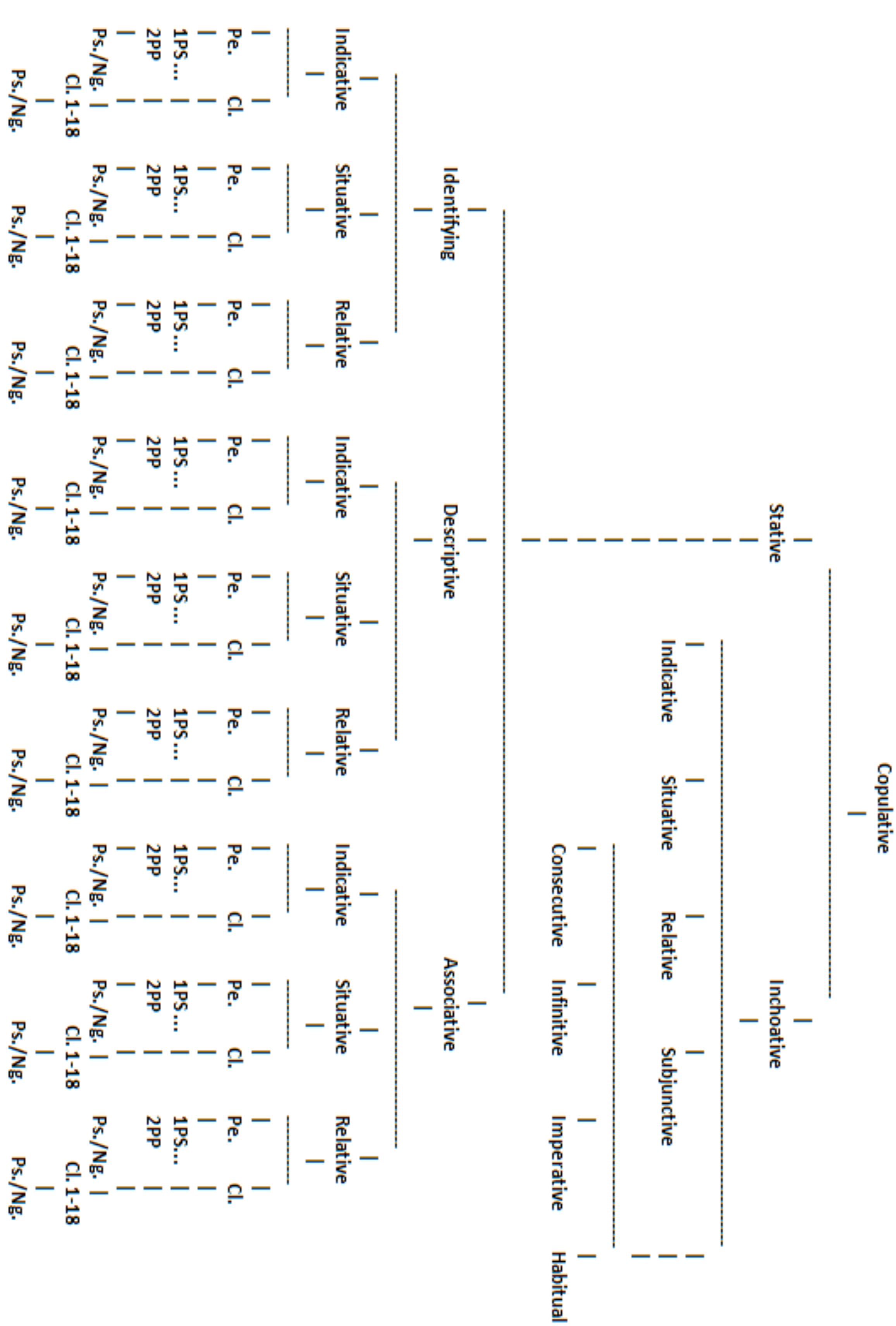
Prinsloo, D.J. (2002). The Lemmatization of Copulatives in Northern Sotho. *Lexikos*, 12, pp. 21-43.

Rundell, M. (2009). The Road to Automated Lexicography: First Banish the Drudgery then the Drudges? In S. Granger, M. Paquot (eds.) *eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, pp. 9-10.

Sesotho sa Leboa (Northern Sotho) - English Dictionary (2003) http://africanlanguages.com/sdp/.

Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In P.A. Fuertes-Olivera, H. Bergenholtz (eds.) e-*Lexicography: The Internet, Digital Initiatives and Lexicography*. 2011. London & New York: Continuum, pp. 55-70.

Appendix 1: The Copulative in Northern Sotho



(Pe. = Person; Cl.=Noun classes; PS=Person Singular; PP=Person Plural; Ps.=Positive; Ng.=Negative)

The Spanish Learner's Dictionary *DAELE* on the Panorama of the Spanish E-lexicography

Irene Renau, Paz Battaner
Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)
C/Roc Boronat, 138, 08018 Barcelona (Spain)
E-mail: irene.renau@upf.edu, paz.battaner@upf.edu

Abstract

This paper presents a prototype of an Internet-based Spanish dictionary for foreign learners, the *Diccionario de aprendizaje del español como lengua extranjera* (*DAELE*, http://www.*DAELE*.eu), focusing on the possibilities that the new electronic format offers, as well as the details of the editing of the dictionary and the properties of the user interface. *DAELE* is, first of all, *a dictionary,* and the adjective *electronic* comes after paying attention to all the classic lexicographical tasks: in a context—that of Spanish lexicography—where paper format and traditional approaches are most frequently found, *DAELE* has been created from scratch. Firstly, we start off by offering a short retrospective of Spanish and European e-lexicography; secondly, we explain the characteristics of the dictionary as a learners' tool, developed under the influence of Sinclair's conception of lexicography (put in practice in *COBUILD*, Sinclair and Hanks, 1987). Thirdly and finally, we adhere to De Schryver's account of lexicographers' dreams (2003), and we contribute with our suggestions to the future—a future that still seems far away and some dreams that seem to be already ageing.

Keywords: corpus-driven lexicography; monolingual lexicography; online dictionary projects; Spanish learners' dictionaries

1. Introduction

This paper [1] describes the main features of the *Diccionario de aprendizaje de español como lengua extranjera (DAELE),* a prototype of a learners' web-based dictionary created from scratch in terms of both contents and format. On the following pages, we will try to show that these two aspects (content and format) are strongly interconnected, as the format the Internet provides seems to be the most suitable medium for one of the most salient characteristics of the dictionary: focusing at the same time both on production and reception. Thus, this web interface allows for an exhaustive exploitation of crucial information, such as grammar, extensive exemplification and other complementary aspects (see section 3). It seems, in the case of lexicography, that technological innovations pave the way for the materialization of consolidated ideas about how a good learners' dictionary should be. *DAELE* is currently in progress and it is too soon to properly call it *a dictionary,* but it is at least *a dictionary project* or *prototype,* with only a group of about 350 verbs published on the Internet. In spite of its short and modest story, the fact that from its inception it was never meant to be available in printed version makes it, even in the present days, a very uncommon and innovative project both in the Spanish and European context.

Our objective on the following pages will be to describe this prototype and examine it within the European context, and, specifically, in the context of Spanish e-lexicography. The paper has been divided into the following sections: firstly, we begin offering a short retrospective of Spanish and European e-lexicography; secondly, we explain the main features of *DAELE* in general, as a Spanish learners' dictionary and describe the web interface, with slight comparisons with other current Spanish web dictionaries; and finally we explain our future plans related to improvements in the interface and the compilation process.

2. A Short Retrospective of Spanish and European E-lexicography

With the popularisation of the Internet, a wide number of lexicographic studies paid attention to this new revolutionary format, whose invention has been considered one of the most important technical changes in history related to the transmission of knowledge (Harnad, 1991). Very early, lexicographical theory wondered about how an online dictionary could be and what advantages the new medium had. As this does not seem to be the place to make a wide state-of-the-art consideration, we consider as referential—among other works—the studies of Atkins (1996), De Schryver (2003), De Schryver and Joffe (2004), Verlinde, Leroyer and Binon (2010) and Rundell and Kilgarrif (2011). In Dzemianko (2010) and Chen (2010), comparisons are made between paper and electronic dictionaries, with quite optimistic results praising the latter.

Apart from numerous metalexicographical studies, practical lexicography is gradually incorporating and adapting new resources to the Internet realm. Cerquiglini (in Pruvost, 2000:118, cited in De Schryver, 2003:143-144) describes this progressive integration of dictionaries to the electronic age in three stages: first

[1] This article has been funded with the following projects: "Agrupación semántica y relaciones lexicológicas en el diccionario", dir.^{a}: J. DeCesaris (HUM2009-07588/FILO), from the Spanish Ministry of Education and Science, and *Diccionario de aprendizaje del español como lengua extranjera*, project of Fundación Comillas. English version revised by Javier Gimeno.

stage, lexicographers helped themselves with computers; second stage, they put the paper versions onto CD-ROMs, the web or other media which are currently obsolete (such as diskettes); third stage, the next step has to do with creating electronic dictionaries from scratch, as brand-new, independent products. These three stages are conceived by Cerquiglini as consecutive, but in fact, nowadays, they are superimposed. To focus on the Spanish tradition, most current dictionaries are distributed also in digital version, a few of them (and the most renowned ones belong in this group) available in both paper and digital—CD-ROM / Internet—versions, and no Spanish dictionary (as far as we are concerned) has been created directly for the Internet. Thus, when we talk about "online" or "web" dictionaries, one first important distinction must be made (according to the tree stages listed above) between those in paper versions slightly modified to fit the digital format and those genuinely thought of as electronic products. From this perspective, *DAELE* is closer to European learners' dictionaries such as DAFLES (Selva, Verlinde and Binon, 2002; http://ilt.kuleuven.be/blf) or ELDIT (Abel and Weber, 2000; http://dev.eurac.edu:8081/MakeEldit1/Eldit.html) than web versions – for Spanish as well as for other languages – which start from previous materials published in paper. In our opinion, it is not pertinent to pay much attention to a comparison between *DAELE* and other online Spanish dictionaries because the conception is too different. We will offer, however, some short remarks on some of these dictionaries in the next section.

2.1 New Tools, Traditional Approaches

As already mentioned, dictionaries do not yet seem to have accomplished a fully adaptation to the Internet age. Probably due to commercial reasons, the majority of the online lexicographic products are of inferior quality than their CD-ROM counterparts (when there is a CD-ROM version). If we focus on Spanish monolingual dictionaries for the general public, in our opinion, the following ones must be considered: *Diccionario de la lengua española (DRAE), Clave: diccionario de la lengua española (Clave), Diccionario Salamanca de la lengua española (Salamanca)* and the aggregators *Diccionarios.com* (which offers the *Diccionario de uso del español de América y España, DUEAE*) and *Wordereference.* All of them are (or contain) versions of the traditional paper-format dictionaries, and very few changes have been made. They can be summarised as follows:

- A basic use of hyperlinks.
- Some search facilities.
- Direct links to verb conjugation.
- Meanings offered in separated paragraphs.
- Use of colour.
- Exploitation of web 2.0 resources.

Despite the easiness to implement some of these features, not all the dictionaries mentioned before offer them to users. Hyperlinks or search facilities are, in our opinion, underexploited resources. Some webs do not include hyperlinks at all in the microstructure. In other cases, there are hyperlinks only to synonyms, when, ideally, all words should be linked to their head entry (as Wordreference and Diccionarios.com do). In relation to search facilities, in many cases they do not exist or are insufficient (with the exception of Wordreference). In most of these web sites, the use of forums or other web 2.0 resources is not implemented either, with the exception of — again — Wordreference and Diccionarios.com. In sum, in spite of the space and other resources that can be used on the web, all of them are exact copies of paper versions, and, thus, they very often lack examples and other complementary information such as lexical combinatory or morphological relations. Hyperlinking is an underexploited resource, as for linking pages of the same web as for linking the web to other webs. As a consequence of all these aspects, this group of widely used dictionaries are not as good as they could be, their quality being inferior to that of the CD-ROM versions. In addition to that, they do not make the most of a medium with a great potential.

2.2 Experimental Online Lexical Databases in Spanish

In a different context, there are other projects (in progress and not oriented to a general public) which represent new ways of offering lexicographic—or, more generally, lexical—resources on the web for the Spanish language. Two databases are being developed for verbal analysis: SenSem (Fernández-Montraveta, Vázquez and Castellón, 2006, http://grial.uab.es/synset/synset2.php) and Adesse (García-Miguel and Albertuz, 2005, García-Miguel, González Domínguez and Vaamonde, 2010; http://adesse.uvigo.es/ADESSE/Inicio). Both are devoted to syntactic and semantic description of Spanish verbs. SenSem lexical database analyses 250 Spanish verbs and offers lexicographical definitions, semantic roles of verb arguments, argument structure as well as other types of information. For each verb, SenSem contains a rather low number of example sentences per verb (100 approximately), but still the database is very complete if the most frequent Spanish verbs are considered. Every verb is linked to its definition and sentences are semantically and syntactically annotated. The Adesse database also offers a semantic-syntactic approach, adding analysis of lexical combinatory of verbs and other differences.

We should not finish this section withouth mentioning some other well-known lexical resources (which will not be described here): WordNet (Fellbaum 1998) and FrameNet (Fillmore, Wooters and Baker, 2001, Fillmore, Johnson and Petruck, 2003), each of which having a corresponding Spanish version, Spanish WordNet, currently in its 3.0 version (Fernández-Montraveta, Vázquez and Fellbaum, 2008, http://grial.uab.es/synset/synset2.php), and Spanish

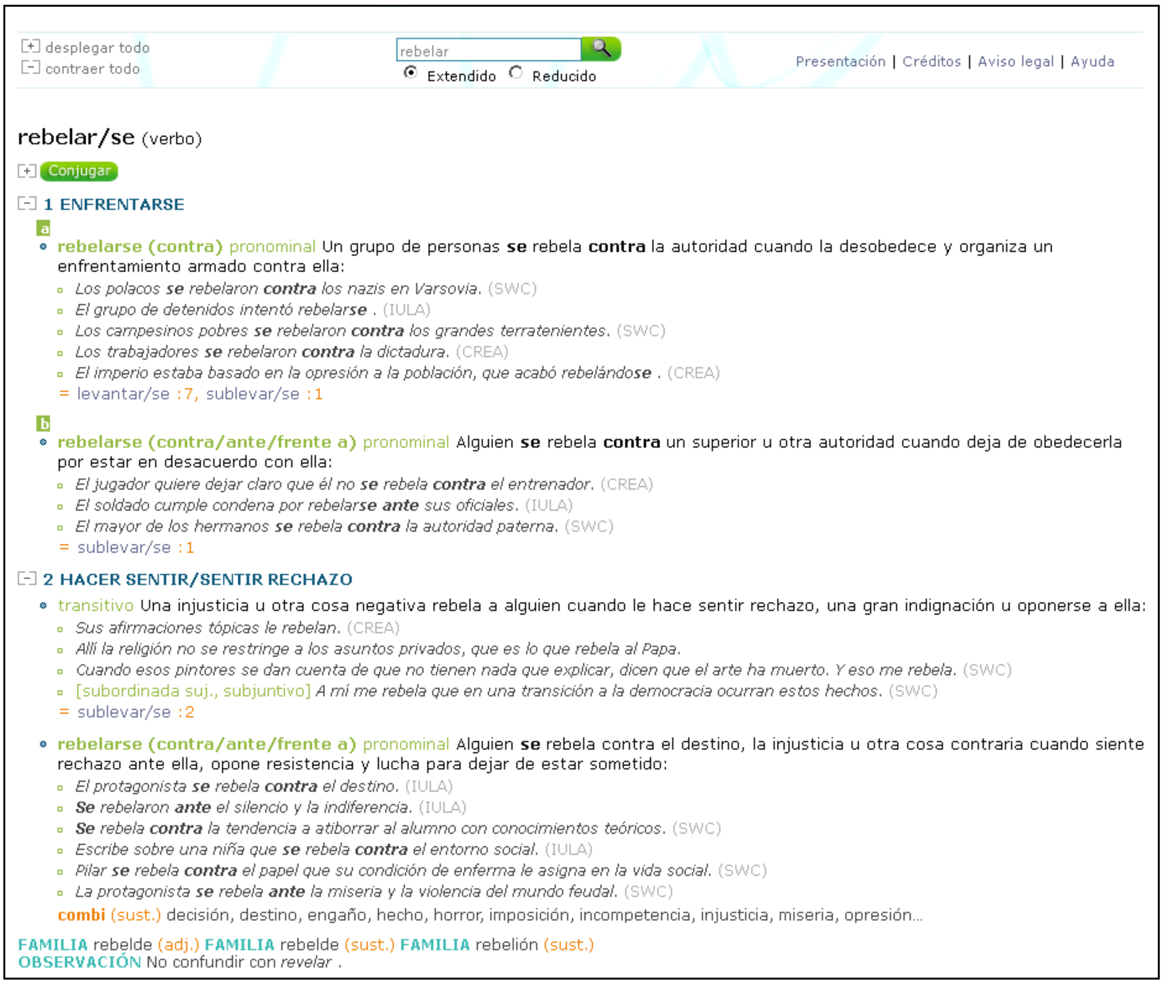
FrameNet (Subirats and Petruck, 2003; http://gemini.uab.es/SFN). These projects, although not strictly "lexicographical", represent the exploitation of electronic lexical resources for the Spanish language probably in a more satisfying way than those web dictionaries shown in section 2.1. One of the differences—described below—with *DAELE*'s approach is that these projects are addressed to a specialised public, whereas *DAELE* targets the general public, particularly foreign-language learners, thus a discussion on pedagogical aspects is necessary regarding the conception of the tool.

3. *DAELE*, an Online Spanish Learners' Dictionary Project

As it has already been pointed out in the introduction, *DAELE* seeks to be a completely new dictionary and intends to take advantage of as many web utilities as possible, but an important clarification must be made: the *DAELE* project has started from scratch in electronic format without disregarding any of the classic lexicographic tasks. The work, thus, has a double face: on the one hand, an innovative (considering today's Spanish dictionaries in any format) dictionary is offered, and, on the other hand, the strictly new contents mentioned above are designed and offered directly for the Internet version. Other online dictionaries have shown many good ideas, yet they just reflect the contents that were created for another medium.

3.1 General Features

In previous papers (Battaner and Renau, 2008; Renau and Battaner, 2010, DeCesaris and Bernal, 2006, Mahecha and DeCesaris, forthcoming, among others), several features of the dictionary strictly related to the lexicographic conception have already been explained. Basically, we adopt the approaches of both Sinclair (1991; 2004) and Hanks (2004, forthcoming), who claim that meanings are associated with the context in which a word appears.



rebelar/se (verbo)

[-] **1 ENFRENTARSE**

a

- **rebelarse (contra)** pronominal Un grupo de personas **se** rebela **contra** la autoridad cuando la desobedece y organiza un enfrentamiento armado contra ella:
  - *Los polacos **se** rebelaron **contra** los nazis en Varsovia.* (SWC)
  - *El grupo de detenidos intentó rebelar**se** .* (IULA)
  - *Los campesinos pobres **se** rebelaron **contra** los grandes terratenientes.* (SWC)
  - *Los trabajadores **se** rebelaron **contra** la dictadura.* (CREA)
  - *El imperio estaba basado en la opresión a la población, que acabó rebelándo**se** .* (CREA)

  = levantar/se :7, sublevar/se :1

b

- **rebelarse (contra/ante/frente a)** pronominal Alguien **se** rebela **contra** un superior u otra autoridad cuando deja de obedecerla por estar en desacuerdo con ella:
  - *El jugador quiere dejar claro que él no **se** rebela **contra** el entrenador.* (CREA)
  - *El soldado cumple condena por rebelar**se ante** sus oficiales.* (IULA)
  - *El mayor de los hermanos **se** rebela **contra** la autoridad paterna.* (SWC)

  = sublevar/se :1

[-] **2 HACER SENTIR/SENTIR RECHAZO**

- transitivo Una injusticia u otra cosa negativa rebela a alguien cuando le hace sentir rechazo, una gran indignación u oponerse a ella:
  - *Sus afirmaciones tópicas le rebelan.* (CREA)
  - *Allí la religión no se restringe a los asuntos privados, que es lo que rebela al Papa.*
  - *Cuando esos pintores se dan cuenta de que no tienen nada que explicar, dicen que el arte ha muerto. Y eso me rebela.* (SWC)
  - [subordinada suj., subjuntivo] *A mí me rebela que en una transición a la democracia ocurran estos hechos.* (SWC)

  = sublevar/se :2

- **rebelarse (contra/ante/frente a)** pronominal Alguien **se** rebela contra el destino, la injusticia u otra cosa contraria cuando siente rechazo ante ella, opone resistencia y lucha para dejar de estar sometido:
  - *El protagonista **se** rebela **contra** el destino.* (IULA)
  - ***Se** rebelaron **ante** el silencio y la indiferencia.* (IULA)
  - ***Se** rebela **contra** la tendencia a atiborrar al alumno con conocimientos teóricos.* (SWC)
  - *Escribe sobre una niña que **se** rebela **contra** el entorno social.* (IULA)
  - *Pilar **se** rebela **contra** el papel que su condición de enferma le asigna en la vida social.* (SWC)
  - *La protagonista **se** rebela **ante** la miseria y la violencia del mundo feudal.* (SWC)

  combi (sust.) decisión, destino, engaño, hecho, horror, imposición, incompetencia, injusticia, miseria, opresión...

FAMILIA rebelde (adj.) **FAMILIA** rebelde (sust.) **FAMILIA** rebelión (sust.)
OBSERVACIÓN No confundir con *revelar* .

Figure 1. The entry *rebelar/se* in DAELE.

Consequently, in order to know what a word means, it is necessary to analyse its context, in terms of syntax but also with regard to combinatory and semantic types of arguments. Sinclairian theories were put into practice in the *COBUILD* project (Sinclair and Hanks, 1987), and this is the type of dictionary we took as reference to create *DAELE*. In addition to that, Hanks' Corpus Pattern Analysis (CPA) is the procedure that we use for the analysis of the data (Renau and Alonso, forthcoming, show how this analysis process is made).

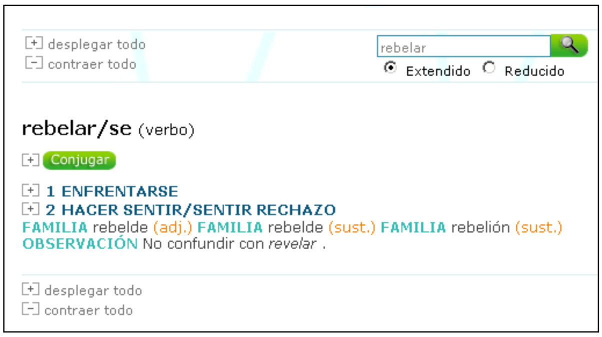
DAELE's microstructure of verb entries is organised following semantic criteria: the entry is divided into big general sense groups (meanings), generally subdivided into specific meanings related to patterns of usage. Each of these meanings is labelled with a hypernym or another general word or phrase which helps the user to easily find the definition they are looking up (Battaner, 2010). Every meaning has at least one "sub-meaning", which in fact corresponds to what we usually know as "meaning": the definition with some syntactic characteristics (such as the grammatical structure or syntactic aspects), exemplified by one or more slightly modified corpus sentences and other common informations. Apart from the numerous grammatical aspects included within each entry, word family (basically the morphologically-related words) and also lists of collocates (words that often combine with the verb, usually nouns in direct object position) are considered innovative lexicographical data.

As an example, see figure 1, corresponding to the entry *rebelar/se* ('to rebel'). This entry is organised in two general meanings: 'to face an authority' (meaning 1, ENFRENTARSE) and 'to display contempt for others' (meaning 2, HACER SENTIR/SENTIR RECHAZO). The two general meanings are represented by full sentence definitions and exemplified with sentences taken from various corpora. Synonyms (such as "levantar/se" or "sublevar/se") are also offered, and some examples also include notes (like the ones in green—colour which signals a clause in subject position, "subordinada suj., subjuntivo"). There are also notes for lexical combinations (in orange, in the last meaning), word families (FAMILIA) and general notes about the whole entry (OBSERVACIÓN; in this case, there is a spelling note about a usual confusion between *rebelar/se* and *revelar*, 'to reveal'). Finally, bold type is used for prepositions and other important words in definitions and examples.

3.2 Electronic features

As we explained above, in Renau and Alonso (forthcoming) there is a description of the process of corpus analysis and how some computational use to lexicographical work are implemented. In this section, we will only pay attention to aspects related to the electronic interface. We must say in advance that *DAELE*, as product designed for the Internet, is still far from realizing the lexicographers' dreams mentioned by De Schryver (2003). Nevertheless, if we take his review as a reference, we must emphasize that *DAELE* has exploited the following advantages of the web format:

- As there are *no space constrictions,* we can offer a large number of examples for every meaning (currently, about 6 examples per meaning). In contrast, having unlimited space leads us to the problem of how to organise the information (Atkins and Rundell, 2008:20-24; Bernal and Renau, 2010).
- It is possible to *quickly update the dictionary,* so that it is easy to make corrections, additions, suppressions or other changes in the database (TshwaneLex, Joffe and De Schryver, 2004).
- It is a *free, open resource,* which can be consulted simultaneously by hundreds of people at no cost. This is crucial for the numerous schools teaching Spanish as a foreign language in the world.
- *It can be used anywhere* (at home, in the classroom, etc.) as long as users have access to the Internet. In general, it is free of abbreviations, so the metalanguage is easier to understand.
- It has *customising options* (see the explanation below) which allow users to organise the data, facilitating access to the information.



[+] desplegar todo
[−] contraer todo

rebelar
● Extendido ○ Reducido

rebelar/se (verbo)

[+] Conjugar

[+] 1 ENFRENTARSE
[+] 2 HACER SENTIR/SENTIR RECHAZO
FAMILIA rebelde (adj.) FAMILIA rebelde (sust.) FAMILIA rebelión (sust.)
OBSERVACIÓN No confundir con *revelar* .

[+] desplegar todo
[−] contraer todo

Figure 2. The initial menu for the entry *rebelar/se* in DAELE.

With regard to this last aspect, when a user looks up a verb, what appears is just the menu with the most general semantic labels, along with some other general information such as the morphological family or notes on the whole entry (figure 2). After reading this initial information, the user can open other sections by using the [+] buttons. If users click on "desplegar todo" ('display all') the whole entry opens. The rest of the buttons, at the beginning of every part, only open the corresponding section. Once the learners access to the selected information, they can use the two versions of the dictionary: the extended one ("Extendido") or the concise one ("Reducido"). The concise version offers the same information with only two examples per meaning and without other complementary data (e.g. notes, morphological family or combinatory).

4. Concluding Remarks: Going Beyond the Traditional Concept of Dictionary

In the previous sections, we have expounded the main characteristics of *DAELE* as we set the dictionary within the context of the current Spanish lexicography available on the net. These are the aspects to be implemented in the future:

- Addition of the pronunciation at head words.
- Improvement of the way in which examples are shown, in order to make the information search process easier. It is necessary to improve the two options which are currently available, that is, "extended" and "concise", in order to make them more dynamic and useful.
- Improvement of the looking-up process, making possible, for instance, searching not only head words, but also in the whole entry (definitions and examples).
- Similarly, users could be given the possibility of using a corpus to be able to retrieve more usage examples. The ideal scenario would be that in which every meaning would be connected to its correspondent corpus concordances, annotated following CPA criteria. As Hanks (2010) points out,

  > "Electronic dictionaries of the future will be much in demand— for computational, pedagogical, and other applications–if they can be used as resources for mapping word meaning systematically onto word use".

- The use of full sentence definitions makes possible to use colours for distinguishing the parts of the sentence (e.g. subject, verb and direct object). The same colours could also be used in some of the examples. This could help to carry out the syntactic analysis providing a better understanding of the examples.

Current technology makes possible to easily implement all of these improvements, so most of them will be available in 2012. Nevertheless, it is soon to implement some of them (for example, a complete use of hyperlinks or the use of a corpus) because only some few entries have already been completed. With the work which already has been done and what will be accomplished in the near future, we expect to make our modest contribution to the model for the "dictionary of the 21st century", bringing the future a little bit closer.

Finally, Spanish is one of the most spoken languages in the world, and it is currently one of the most learnt as a foreign language. It is also one of the most widely used on the Internet[2]. Thus, in this context, better Spanish lexicographic tools are required on the web, and DAELE could be one of the references to create them.

5. References

Abel, A., Weber, V. (2000). ELDIT – A Prototype of an Innovative Dictionary. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds.). *Proceedings of the Ninth Euralex International Congress, EURALEX 2000.* Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, pp. 807–818.

Atkins, S. (1996). Bilingual Dictionaries: Past, Present and Future. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström & C. Röjder Papmehl (eds.) *Euralex'96 Proceedings.* Göteborg (Sweden): Göteborg University, pp. 515-546.

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

Battaner, P. (2010). El uso de las etiquetas semánticas en los artículos lexicográficos de verbos en el *DAELE*. *Quaderns de Filologia. Estudis Lingüístics,* 15, pp. 139-158.

Battaner, P., Renau, I. (2008). Sobre las construcciones pronominales y su tratamiento en algunos diccionarios monolingües de cuatro lenguas románicas. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, pp. 495-504.

Bernal, E., Renau, I. (2010). ¿Lo que necesitan es lo que encuentran? Reflexiones a propósito de la representación de los verbos en los diccionarios de aprendizaje de español. In A. Dykstra, T. Schoonheim (2010). *Proceedings of the XIV Euralex International Congress.* Ljouwert (Netherlands): Fryske Akademy, pp. 484-496.

Chen, Y. (2010). Dictionary Use and EFL Learning. A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries. *International Journal of Lexicography,* 23(3), pp. 275-306.

[*Clave*] Maldonado, C. (ed.). (1996). *Clave. Diccionario de uso del español actual.* Madrid: SM. Web version: http://clave.librosvivos.net.

DeCesaris, J., Bernal, E. (2006). Consideraciones previas a la representación de las formas nominales en el *Diccionario de aprendizaje del español como lengua extranjera (DAELE).* In J. DeCesaris, E. Bernal (eds.). *Palabra por palabra: estudios ofrecidos a Paz Battaner.* Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, pp. 83-92.

De Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic Dictionary Age. *International Journal of Lexicography,* 16(2), pp. 143-199.

De Schryver, G.-M., Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams, S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress. Euralex 2004.* Lorient: Université de Bretagne, pp. 187-196.

[*Diccionarios.com*] http://www.diccionarios.com.

[2] According to Moreno and Otero (2007), there are approximately 400.000.000 native speakers of Spanish (currently one of the most spoken languages in the world) and approximately 23.000.000 speakers with no native competence. It is the third or fourth most spoken language in the world depending on sources. It is also one of the most used on the Internet (Rojo and Sánchez, 2010:103-107). In all cases, the number is increasing every year.

[*DRAE*] Real Academia Española (2001). *Diccionario de la lengua española.* Espasa: Madrid. Web version: http://buscon.rae.es/draeI.

[*DUEAE*] Battaner, P. (ed.) (2001). *Diccionario de uso del español de América y España.* Barcelona: Spes Editorial.

Dziemianko, A. (2010). Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography,* 23(3), pp. 257-273.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge: MIT Press.

Fernández-Montraveta, A., Vázquez, G. & Castellón, I. (2006). *SenSem: a Databank for Spanish Verbs. In Proceedings of the X Ibero-American Workshop on Artificial Intelligence, IBERAMIA.* Ribeirão Preto, Brasil.

Fernández-Montraveta, A., Vázquez, G. & Fellbaum, C. (2008). The Spanish Version of WordNet 3.0. In A. Storrer, A. Siebert A & Geyken (eds.) *Text Resources and Lexical Knowledge.* Berlin: Mouton de Gruyter, pp. 175-182.

Fillmore, C.J., Wooters C. & Baker C.F. (2001). Building a large lexical databank which provides deep semantics. *Proceedings of the Pacific Asian Conference on Language, Information and Computation,* Hong Kong.

Fillmore, C.J., Johnson, C.R. and Petruck, M.R.L. (2003). Background to FrameNet. *International Journal of Lexicography,* 16(3), pp. 235-250.

García-Miguel, J.M., Albertuz, F.J. (2005). Verbs, Semantic Classes and Semantic Roles in the ADESSE project. In K. Erk, A. Melinger, S. Schulte im Walde (eds.) *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.* Saarbrücken, February 28 - March 1 2005.

García-Miguel, J.M., González Domínguez, F. & Vaamonde, G. (2010). ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC).* Valletta (Malta), May 17-23.

Hanks, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.

Hanks, P. (2010). Elliptical Arguments: a problem in relating meaning to use. In S. Granger, M. Paquot (eds.) *E-Lexicography in the 21st century: New challenges, new applications. Proceedings of ELEX2009*. Cahiers du CENTAL. Louvain-la-Neuve : Presses Universitaires de Louvain, pp. 109-124.

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

Harnad, S. (1991). Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge. *Computer Systems Review,* 2(1), pp. 39-53.

Joffe, D., De Schryver, J.-M. (2004). TshwaneLex – Professional off-the-shelf lexicography software. In *Third International Workshop on Dictionary Writing Systems.* Brno: Faculty of Informatics, Masaryk University.

Mahecha, V., DeCesaris, J. (2011). Representing nouns in the *Diccionario de aprendizaje del español como lengua extranjera (DAELE).* In I. Kosem, K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New Applications for New Users (Proceedings of eLEX2011).* Bled, 10-12 November 2011. Ljubljana, Slovenia: Trojina.

Moreno, F., Otero, J. (2007). *Demografía de la lengua española.* Madrid: Fundación Telefónica - Instituto Complutense de Estudios Internacionales.

Renau, I, Alonso, A. (forthcoming). Using Corpus Pattern Analysis for the Spanish Learner's Dictionary *DAELE* (Diccionario de aprendizaje del español como lengua extranjera). *Proceedings of the 2011 Corpus Linguistics Conference,* Birmingham, July 20-22, 2011.

Rojo, G., Sánchez, M. (2010). *El español en la red.* Barcelona-Madrid: Ariel-Fundación Telefónica.

Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end?. In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds.) *A Taste for Corpora. A tribute to Professor Sylviane Granger.* Amsterdam: John Benjamins.

[*Salamanca*] Gutiérrez Cuadrado, J. (ed.) (1996). *Diccionario Salamanca de la lengua española.* Madrid: Santillana. Web version: http://fenix.cnice.mec.es/diccionario.

Selva, T., Verlinde, S., Binon, J. (2002). Le Dafles, un nouveau dictionaire électronique : pour apprenants du français. In A. Braasch, C. Povlsen (eds.) *Proceedings of the Tenth Euralex International Congress, Euralex 2002.* Copenhagen: CST, pp. 199-208.

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

Sinclair, J. (2004). *Trust the text: language, corpus and discourse* (Carter, R., ed.). London: Routledge.

Sinclair, J., Hanks, P. (eds.). (1987). *Collins Cobuild English Language Dictionary.* Glasgow: Harper-Collins.

Subirats, C., Petruck, M.R.L. (2003). Surprise: Spanish FrameNet! Presentation at Workshop on Frame Semantics. Prague (Czech Republic): *International Congress of Linguists.* July 29, 2003.

Verlinde, S., Leroyer, P. & Binon, J. (2010). Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-oriented Multifunctional Leximats. *International Journal of Lexicography,* 23(1), pp. 1-17.

[*Wordreference*] http://www.wordreference.com.

Detecting Structural Irregularity in Electronic Dictionaries Using Language Modeling

Paul Rodrigues, David Zajic, David Doermann, Michael Bloodgood, Peng Ye
University of Maryland
College Park, MD
E-mail: {prr, dmzajic, doermann, meb, pengyu}@umd.edu

Abstract

Dictionaries are often developed using tools that save to Extensible Markup Language (XML)-based standards. These standards often allow high-level repeating elements to represent lexical entries, and utilize descendants of these repeating elements to represent the structure within each lexical entry, in the form of an XML tree. In many cases, dictionaries are published that have errors and inconsistencies that are expensive to find manually. This paper discusses a method for dictionary writers to quickly audit structural regularity across entries in a dictionary by using statistical language modeling. The approach learns the patterns of XML nodes that could occur within an XML tree, and then calculates the probability of each XML tree in the dictionary against these patterns to look for entries that diverge from the norm.

Keywords: anomaly detection; error correction; dictionaries

1. Introduction

Many dictionaries today are developed using tools that save to Extensible Markup Language (XML)-based standards, such as the Lexical Markup Framework (LMF) (Francopoulo et al., 2007), the Lexicon Interchange FormaT (LIFT) (Hosken, 2009), or the Text Encoding Initiative (TEI) (Burnard & Bauman, 2007). Often, these standards allow high-level repeating elements to represent lexical entries and utilize descendants of these repeating elements to represent the structure of each lexical entry, in the form of an XML tree.

This paper presents a method to audit the structural regularity across all the entries in a dictionary, automatically. This approach uses statistical language modeling (LM), a technique commonly used in natural language processing, to learn the linear combinations of XML nodes that could occur within a lexical entry, and then evaluates each of these lexical entries against the learned patterns, looking for entries that diverge from the norm.

Technical users of XML often utilize tools to check the well-formedness of an XML document, or to determine the validity of a document as applied to a particular data schema. These help catch certain types of errors, such as syntax, or data relationship errors.

With many dictionary schemas, however, the structure within entries can vary from entry to entry. This structural permissiveness can allow a dictionary writer to introduce or underspecify ambiguous relationships, or to accidently place a node underneath an incorrect parent node in the entry's XML tree. These kinds of errors may be valid XML and may conform to the data specification, so they will not be caught by traditional XML tools, but they are *semantically* incorrect.

The LM technique described here linearizes the lexicon structure, ignoring the underlying text, converting the opening tags in XML into tokens, and then considering the string of tokens representing a lexical entry to be a sentence. A probabilistic language model is learned from these example sentences, and then that model is evaluated against each lexical entry in the corpus. Nodes that are in unusual positions produce a high perplexity, identifying possible anomaly points.

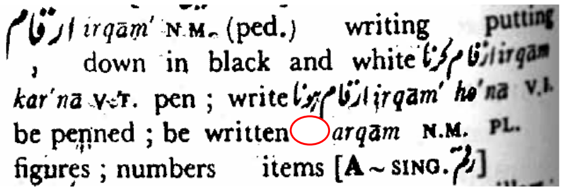
2. XML

Extensible Markup Language (XML), a text format used to store hierarchical data electronically, is often described by a data modeling definition such as a Document Type Definition (DTD), XML Schema (Gao, 2011), or RELAX NG (ISO, 2008). These data modeling definition documents use a regular language to define the data permissible in the XML document. Tools are available to validate, or ensure strict compliance of an XML document, to the data modeling definition. These tools result in a Boolean decision as to whether the data conforms to the specification, and are unable to alert the user to structurally valid, but illogical or rare structures that one may wish to investigate.

3. Structural Errors in Dictionaries

3.1 Dictionary Creation

Dictionaries are often the product of long-term research projects, or large-scale projects created quickly with multiple collaborators. Without strict conformance to a recording standard, entries can drift in style across time or between collaborators. Additionally, dictionaries can be large and complex, leaving them expensive to edit. Whether due to cost or deadlines, dictionaries are published that have errors and inconsistencies.

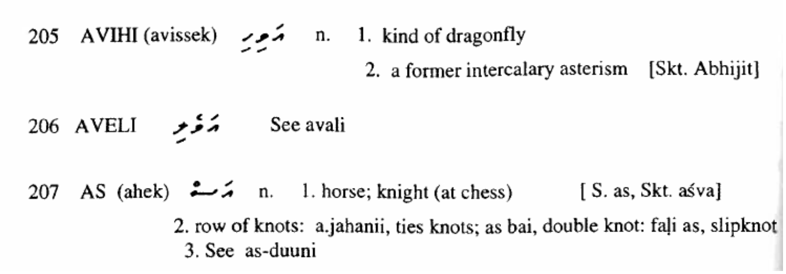


ارقام irqām' N.M. (ped.) writing putting down in black and white ارقام کرنا irqām' kar'na v.t. pen ; write ارقام ہونا irqām' ho'na v.i. be penned ; be written arqām N.M. PL. figures ; numbers items [A ~ SING. رقم]

Figure 1: Missing Orthography. Scan from Qureshi (2003)

3.2 Dictionary Digitization

The process of digitizing dictionaries from a printed book by optical character recognition (OCR), or manually keying in content, can cause additional structural errors to be introduced. Typically in print dictionaries, typefaces, text size, text position, and unreserved symbols are used in combination to indicate the structure of a lexical entry and the scope of linguistic operators (such as English words *and* or *or*). Typographical errors that occur in the original print dictionary, misinterpretation by the OCR system, operator ambiguity, or typist error during the digitization stage can alter the intended structure of the dictionary. These errors can result in incorrect marking of subcomponents within a lexical entry or incorrectly understanding scope within the language examples. In bilingual dictionaries, translations may be forgotten (Figure 1), and languages may be mixed with no delineation (Figure 2).



205 AVIHI (avissek) އަވިހި n. 1. kind of dragonfly
2. a former intercalary asterism [Skt. Abhijit]

206 AVELI އަވެލި See avali

207 AS (ahek) އަސް n. 1. horse; knight (at chess) [S. as, Skt. aśva]
2. row of knots: a.jahanii, ties knots; as bai, double knot: faḷi as, slipknot
3. See as-duuni

Figure 2: No signal between Dhivehi pronunciation and English meanings. Scan from Reynolds (2003)

4. Anomaly Detection Using Language Modeling

While language modeling is not a common approach for structural anomaly detection, it has been employed to detect anomalies in language use. Language models in natural language processing are commonly used to model linear combinations of word tokens or part-of-speech types. Lexicon XML structure is similar to the latter, in that the node names and attributes within the XML are chosen from a small closed class.

Xia & Wong (2006) used language modeling to tag lexemes in Chinese-language Internet chat transcripts as either standard Chinese, or anomalous. The authors noted that chat speak has a dynamic lexicon, and training corpora for supervised systems in this domain can obsolesce.

The authors trained trigram language models on standard Chinese newspaper corpora in order to induce typical values for trigram entropy on words and parts of speech in the standard language. They then learned a language model on a hybrid corpus consisting of newspaper corpora, and a chat transcript corpus. With the typical entropy values known from the newspaper corpus, they evaluated a language sample with the hybrid model. When a trigram had higher entropy than the standard average, they marked that trigram as an anomaly. The authors found words to be a better indicator of anomaly than POS tags, reaching an F-score of 0.85 for words and 0.70 for POS tags in their best conditions.

The authors do not qualify the data that is flagged anomalous. It would be interesting to know if this data is dialectal Chinese, misspelled words, bad grammar, emoticons, or lexemes unique to Chinese Internet chat. The POS tag condition is more comparable to our scenario, as both their POS categories and our structure description language have a small vocabulary.

Jabbari (2010) was interested in detection of anomalous words in the context of the words around them, which has practical applications including real-word spelling error detection and word sense disambiguation. The author examined an approach using bags of words, one using language models, and then a combination of the two. The language modeling approach marked a word as an anomaly if the probability of the word in that context was less than the expected probability of not having that word in the context. The language modeling system received an overall F-score of 0.71. It performed less well than the author's bag of words model, and the combined model. The author did not look at parts of speech, which is more relevant to our task.

5. LM Anomaly Detection for Flattened Structures

In the previous section, we showed how language modeling has been used to detect anomalies in a linear string of tokens. This section explains how to convert the XML tree into a linear string of tokens, and how this is used to build a language model.

The input to the language model is determined by specifying a repeating node in the XML file that contains child trees to be examined. Each of these repeating tree structures is traversed depth-first, and the element names and attributes of children are recorded to a buffer as a tag that identifies that element. In XML, a depth-first search is a linear scan of each node within a tree. At the end of each repeating node the buffer contains a single layer of whitespace-separated tags corresponding to a flattened representation of the tree. We call this a tag sentence. The tag sentences for all the repeating nodes form a corpus of tag sentences.

This corpus can then be used to train a statistical language model. For our experiments, we used the SRI Language Modeling Toolkit (SRILM) (Stolcke, 2002). SRILM includes command line programs and C++ libraries to calculate n-gram statistics for a language, and to measure the perplexity of a text sample to those statistics. SRILM reads and writes to a standard ARPA (Advanced Research Projects Agency) file format for n-gram models. There are other language modeling toolkits available. Typically, the differences between applications are in the speed of the evaluation, the size of the model created, or the statistical smoothing algorithms included for estimating low-occurrence combinations. In our case, speed and model size is not much of an issue, but estimating combinations of low token occurrence is. Since we are training and testing on the same dataset, advanced smoothing algorithms would not add any benefit. SRILM runs with Good-Turing and Katz back-off by default.

6. Evaluation

6.1 Dictionary and Evaluation Data

We perform our evaluations on a bilingual Urdu-English dictionary of 44,237 lexical entries (Qureshi, 2003). This dictionary has been edited by a team of linguists and computer scientists to remove errors using a change-tracking system we refer to as Dictionary Manipulation Language (DML) (Zajic et al., 2011). DML provides a number of benefits for dictionary editing, but the core advantage for this application is that DML can be used to mark every error discovered in the original source dictionary. From the change log, we can create a list of trees we know to be errorful that can be used for evaluating our automatic systems.

6.2 Tree Tiers

The entries in Qureshi (2003) can contain multiple senses, each of which can contain multiple word forms. An entry can also contain word forms directly. These are high-level structures within each entry that can vary significantly. In order to isolate where the errors occur within the entry, we partition some of the structure, performing evaluations on ENTRY, SENSE, and FORM nodes separately. For the ENTRY evaluation, the highest-level SENSE and FORM branches were collapsed into single nodes, with their descendants pruned. For the SENSE trees, descendent SENSE and FORM branches were collapsed. No branches were collapsed in the trees for FORM evaluation. We call the ENTRY, FORM, and SENSE trees *tiers*.

Our three tiers are listed in Table 1, showing the number of occurrences in the dictionary, as well as the number of nodes of that tree tier that had a hand-made correction within the tree.

Tier	Count in Dictionary	Hand-corrected
ENTRY	15,808	7,511
FORM	51,105	9,845
SENSE	88,465	20,037

Table 1: Tree counts and manual correction counts

6.3 n-gram Models

For each of these tiers, content and closing tags were removed, and the trees flattened to form a tag sentence. These three corpora were used to train 2-, 3- and 4-gram language models, without smoothing.

Tree	Unique Tokens	2-gram	3-gram	4-gram
ENTRY	21	178	395	667
FORM	7	25	44	51
SENSE	22	183	384	628

Table 2: Unique Token and n-gram grammar counts at each tree level.

This language model serves to provide prototype trees for comparison, storing which tags can co-occur with which others, and what the likelihood of that co-occurrence will be. Table 2 lists the three tiers, the count of unique tokens (XML descendants) under that tier, and the number of unique n-grams created by the linear combination of those tokens.

6.4 Applying the models

Each tag sentence from the dictionary is then evaluated with this language model, producing statistical measurements for each flattened tree structure: log probability of the sentence (LOGPROB), average perplexity per word (PPW), and average perplexity per word with end tags (PPWET). LOGPROB and PPWET both evaluate trees against n-grams that contain START OF SENTENCE and END OF SENTENCE tags. This helps model differences between tokens that appear initial or final in the tag sentence.

We rank these measurements to force the trees into a decreasing order of anomalousness. For LOGPROB, the trees are sorted in ascending order, and for both PPW and PPWET, the trees are sorted in descending order.

For evaluation, we provide precision at the top R anomalies, where R can be {15, 30, 50, 100, 500, or 1000}. A hit occurs where a tree in the top R of our list has shown up in our errorful tree list. Precision at Rank is defined as the number of hits divided by R.

6.5 4-gram Results

Out of the three n-gram lengths evaluated, 4-grams performed the worst overall. The average of the six precisions at rank scores for each tree tier and each language model measurement were lower than those for both 3- and 2-grams. Several trials in the group did reach the best scores for their Tier-R combination, but these are matched in the 2- and 3-gram models. Results can be seen in Tables 3, 4, and 5.

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.93	.80	.70	.63	.61	.62	.72
FORM	.93	.93	.96	.98	.98	**.99**	.96
SENSE	**.93**	.93	.92	.89	.61	.56	.81

Table 3: Descending PPWET 4-grams

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	**1.0**	.70	.66	.60	.65	.67	.71
FORM	.93	.97	.98	**.99**	**.99**	**.99**	.98
SENSE	**.93**	.93	.92	.90	.56	.54	.80

Table 4: Descending PPW 4-grams

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.87	**.93**	.94	.90	.80	.76	.87
FORM	.80	.90	.92	.95	.98	.78	.89
SENSE	**.93**	.93	.96	.91	.85	.81	.90

Table 5: Ascending LOGPROB 4-grams

6.6 3-gram Results

The 3-gram language model performed well, capturing the best average Tier / R trials for FORM with the PPW measurement. The results can be found in Tables 6, 7, and 8.

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.93	.83	.80	.73	.69	.72	.78
FORM	.93	.93	.96	.98	.98	**.99**	.96
SENSE	**.93**	.90	.92	.91	.68	.69	.84

Table 6: Descending PPWET 3-grams

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.93	.73	.72	.69	.69	.74	.75
FORM	.97	.98	.99	**.99**	**.99**	**.99**	**.99**
SENSE	**.93**	.93	.92	.91	.64	.59	.82

Table 7: Descending PPW 3-grams

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.87	**.93**	.94	**.93**	.86	.84	.90
FORM	.87	.93	.94	.95	.98	.78	.91
SENSE	**.93**	**.97**	**.98**	.94	.91	.87	.93

Table 8: Ascending LOGPROB 3-grams

6.7 2-gram Results

2-gram language models results can be found in Tables 9, 10, and 11. This length n-gram performed the best, with the best average Tier / R trial for ENTRY and for SENSE using the LOGPROB measurement. This measurement, shown in Table 11, has the largest number of Tier/R trials with the highest precision.

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.73	.73	.74	.73	.79	.76	.75
FORM	.93	.97	.98	.98	.98	**.99**	.97
SENSE	**.93**	.83	.74	.81	.74	.78	.81

Table 9: Descending PPWET 2-grams

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.73	.73	.66	.72	.75	.79	.73
FORM	.93	.97	.98	**.99**	**.99**	**.99**	.98
SENSE	**.93**	.93	.92	.83	.71	.76	.85

Table 10: Descending PPW 2-grams

Tier / R	15	30	50	100	500	1000	AVG
ENTRY	.87	**.93**	**.96**	**.93**	**.91**	**.90**	**.92**
FORM	**1.0**	**1.0**	**1.0**	.97	.98	.78	.96
SENSE	**.93**	**.97**	**.98**	**.96**	**.96**	**.91**	**.95**

Table 11: Ascending LOGPROB 2-grams

6.8 Other-grams

Unigram, 5-gram, and 6-gram models were also evaluated according to their LOGPROB. 5- and 6-gram models performed at a lower accuracy for nearly all R and tree levels. Unigram evaluations were inconclusive. Accuracy was slightly higher for some R, but some were far lower.

7. Conclusions

We presented a statistical error detection technique for dictionary structure that uses language modeling to rank anomalous dictionary trees for human review. To create the language model, we split the dictionary into three tiers-ENTRY, FORM, and SENSE, and flatten each to form a tag sentence. We create 2-, 3-, and 4-gram language models based on this flattened structure, and evaluate against the original dictionary using Perplexity Per Word (PPW), Perplexity Per Word with End Tags (PPWET), and log probability (LOGPROB). These measurements were ranked, and we presented Precision-at-Rank for all trials.

We found the highest precision Tier/R trials to be spread across several n-gram length language models, and several language model measurements. In general, we find that the best overall configuration is a 2-gram language model, which ranks the trees by ascending log probability. Averaging our six precisions for this metric, the system reached 92% precision on ENTRY error detection, 95% on SENSE, and 96% on FORM. Evaluating the top 50 anomalies, we reached 96% precision on ENTRY, 98% on SENSE, and 100% on FORM.

8. Comments

Though a large amount of man-hours were dedicated to the eradication of errors in our copy of the dictionary, we can make no assumption that we have found all of the errors present, and some of the trees that have not been marked bad, may indeed be errorful. Evaluation of our system, given this scenario, provides some difficulty. We have a small number of known-bad trees from the original source dictionary. The large remainder of trees is of questionable character, but are probably good. We cannot make large-scale automatic judgments on the questionable trees, but we can make sure the known-bad trees are ranked highly in our system. Actual precision should be considered at least the numbers reported. Unfortunately, without known-good trees, it is difficult to provide reliable recall measurements.

9. Future Work

9.1 Iterative language model improvement

As each error in a dictionary is corrected, the language model created from that dictionary improves. An iterative approach, having a linguist examine a small R, correcting the trees, and then rerunning the model, may be the most efficient use of a linguist's time.

9.2 Bootstrapping a cleaner model

With DML, we can find a small percentage of trees that are guaranteed to have had human review at some level. These trees are more likely to be correct than the completely untouched trees, and a corpus of the trees from the final dictionary could be used to create a higher quality language model to compare against the source dictionary.

9.3 Node-level anomaly detection

Evaluation of a language model on a tag sentence outputs a probability at each word. It would be interesting to show whether the peaks of perplexity correspond to the precise errors corrected in our dictionary.

9.4 Related systems

The language modeling approach is the first in a series of experiments examining anomaly detection on dictionary structure. We have several other frameworks currently under development, and expect approaches that harness structure, instead of flattening structure, will perform with higher accuracy. Additionally, we are planning work on a graphical tool to enable dictionary editors to interact with these anomaly detection systems, and plan to research how these systems can incorporate automatic error correction with assistance of an editor.

10. Acknowledgements

We would like to thank Mona Madgavkar for answering questions about the dictionary we were using for evaluations, as well as the other linguists that have contributed to this dictionary's correction. Additional thanks to C. Anton Rytting, Mike Maxwell, and two anonymous reviewers for their comments on this work.

This material is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. Nothing in this report is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the authors of the product, process, or service that is the subject of this report. No one may use any information contained or based on this report in advertisements or promotional materials related to any company product, process, or service or in support of other commercial purposes.

11. References

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M. & Soria, C. (2007). Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. GLDV (Gesellschaft für linguistische Datenverarbeitung), Tübingen.

Hosken, M. (2009). Lexicon Interchange Format. Version 0.13 DRAFT. Retrieved from http://code.google.com/p/lift-standard/ Sept. 30, 2011.

Burnard, L., Bauman, S. (2007). P5: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative. Retrieved from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ Sept, 30, 2011.

Gao, S., Sperberg-McQueen, C.M., Thompson, H.S. (2011). W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures. W3C Candidate

Recommendation 21 July 2011. Retrieved from http://www.w3.org/TR/xmlschema11-1/ Sept. 30, 2011.

Jabbari, S. (2010). A Statistical Model of Lexical Context. Ph.D Thesis. University of Sheffield.

ISO. (2008). ISO/IEC 19757-2:2008 Information technology -- Document Schema Definition Language (DSDL) -- Part 2: Regular-grammar-based validation -- RELAX NG. ISO.

Qureshi, B.A., Abdul Haq. (2003). *Standard 21st Century Dictionary*. Educational Publishing House, Delhi, India.

Reynolds, C. (2003). *A Maldivian Dictionary*. Curzon. New York: Routledge.

Stolcke, A. (2003). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*. Denver, Colorado. September 2002.

Xia, Y., Wong, K-F. (2006). Anomaly Detecting within Dynamic Chinese Chat Text. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational LinguisticsWorkshop On New Text Wikis And Blogs And Other Dynamic Text Sources*. Trento, Italy. April 4, 2006.

Zajic, D., Maxwell, M., Doermann, D., Rodrigues, P. & Bloodgood, M. (2011). Fixing Errors in Digital Lexicographic Resources Using a Dictionary Manipulation Language. In I. Kosem, K. Kosem (eds.) *eLexicography in the 21st century: New applications for new users, Proceedings of eLEX 2011,* Bled, 10-12 November 2011. Ljubljana, Slovenia: Trojina.

The DANTE database: a User Guide

Michael Rundell, Sue Atkins
Lexicography MasterClass
E-mail: michael.rundell@lexmasterclass.com, sue.atkins@lexmasterclass.com

Abstract

DANTE – the Database of ANalysed Texts of English – is a lexical database which provides a corpus-based description of the core vocabulary of English. It records the semantic, grammatical, combinatorial, and text-type characteristics of over 42,000 single-word lemmas and 23,000 compounds and phrasal verbs, and it also includes over 27,000 idioms and phrases. Every fact recorded in the database is derived from a systematic analysis of a 1.7 billion-word corpus and supported by corpus examples. The complete text of DANTE from M to R is freely available online (at www.webdante.com), and the full database is available through research or commercial licences (http://dante.sketchengine.co.uk/). The website provides basic information about DANTE and a Help function to assist users who wish to search the database. This User Guide is intended to complement the information on the website and explain the rationale of the various components of DANTE's microstructure.

Keywords: lexical database; query-builder; inherent grammar; syntactic context; label; domain; multiword; chunk; collocation; support verb; support preposition; itemiser; pragmatics

1. Introduction

DANTE is a lexical database which provides a fine-grained, corpus-based description of the core vocabulary of English. Every fact recorded in the database is derived from, and explicitly supported by, evidence from a 1.7 billion-word corpus of current English. Almost all of these facts are machine-retrievable.

DANTE – the Database of ANalysed Texts of English – was designed and created for Foras na Gaeilge by the Lexicography MasterClass and an 18-strong team of skilled lexicographers, using the Sketch Engine (www.sketchengine.co.uk/) for corpus-querying, and IDM's Dictionary Production System (DPS: www.idm.fr) for entry-building. The resulting database records the semantic, grammatical, combinatorial, and text-type characteristics of over 42,000 single-word lemmas and 23,000 compounds and phrasal verbs, and includes over 27,000 idioms and phrases, underpinned by over 600,000 sentence examples from the corpus.

Though DANTE's primary function was to provide an 'English framework' for the development of a new English-Irish dictionary (www.focloir.ie/english.asp), it was designed from the start to be a linguistic resource of more general utility. It offers publishers a launchpad for the development or updating of monolingual or bilingual dictionaries, and provides rich data for researchers, language engineers, software developers, and materials writers.

2. The DANTE User Guide

A large section of the DANTE database (the complete text from M to R) is freely available online at www.webdante.com. As well as searching for individual headwords, users can employ a query-builder in 'Advanced Search' mode to create a wide range of specialised searches: to find, for example, every verb that takes a particular construction, or every American English word with the style label 'humorous'. The complete database is available through research or commercial licences (http://dante.sketchengine.co.uk/).

The DANTE User Guide has been prepared to facilitate searches for users of the public website or of the complete database. It has two functions:

- to describe the search parameters available on the DANTE website
- to explain the components of entries in the database, and the editorial policies underlying them.

The website's Help function provides all the information needed to specify a search. But entries in DANTE are often complex; the lexicographers' Style Guide runs to well over 100 pages. So it is important to stress that those entry components available as search options on the website represent only a subset of all possible entry components. One of the objectives of the Guide, therefore, is to explain the content and rationale of every information-type you see when viewing entries returned by a search.

3. The 'lexical unit' in DANTE

One of the key features of DANTE is that each of the main entry components (such as a part of speech label, a grammar code, or a style label) is associated with a particular 'lexical unit'. In DANTE, a 'lexical unit' (or LU) approximates to a 'dictionary sense', and it is the principal 'currency' of the database. More precisely, it is an umbrella term for describing any use of a word that carries its own discrete meaning or function: single-sense headwords, individual senses of polysemous words, idioms, compounds, and phrasal verbs are all lexical units. And (just as with polysemous headwords), if an idiom, compound, or phrasal verb has more than one sense, each counts as a lexical unit.

4. Definitions and examples in DANTE

4.1 Definitions

Since DANTE is a lexical database rather than a dictionary, it does not have conventional definitions. Rather, every LU has a ‘meaning statement’, whose function is not to ‘define’ the item in detail but to provide enough semantic information to enable the user to recognise which meaning area or ‘dictionary sense’ the LU relates to. DANTE’s meaning statements are thus closer to the ‘sense indicators’ used in bilingual dictionaries (Atkins and Rundell, 2008: 503-504) – though they are generally fuller.

4.2 Examples

All the lexical data in DANTE is driven by the corpus, and it is a fundamental design feature that every linguistic fact recorded in the database is illustrated by one or (usually) several corpus examples. Consequently, DANTE includes well over 600,000 example sentences.

Take, for example, the relatively infrequent word *recollection*. In DANTE, *recollection* has two LUs: an uncountable use (‘the act of remembering’), and a countable one (‘a memory’). The second of these has no fewer than 25 examples: the first four exemplify its basic use; the next five illustrate its most frequent adjective collocates (see 8.2.2), and the rest are examples of the various syntactic contexts (see 5.3) in which the noun regularly participates.

The vast majority of examples are taken directly from corpus data, without modification. Occasionally, corpus examples are supplemented by short, formulaic examples, inserted to illustrate the full range of possible contexts for members of a particular semantic set: for example, colour terms include formulaic examples for the noun use like these ones found at *pink*:

- *dressed in pink*
- *wearing pink*
- *available in pink*
- *she likes pink*
- *a shade of pink*

5. Grammar and Syntax in DANTE

This section explains the following search conditions which are available in the ‘Build a search’ drop-down lists on the DANTE website:

- part of speech
- inherent grammar
- syntactic context

5.1 Part of speech (POS)

Table 1 lists the items that appear in the drop-down menu for a ‘part of speech’ search. Most are self-evident, but an explanation is provided in cases where there might be doubt as to how the part of speech is used in DANTE.

Searches on any of the four main parts of speech – adjective, adverb, verb (all types), and noun – can be refined using the ‘inherent grammar’ condition. Inherent grammar is explained in section 5.2.

Part of speech in drop-down list	**Part of speech in DANTE entries**	**Explanation**	**Examples**
adjective	adj		
adverb	adv		
conjunction	conj		
determiner	det	Used for definite and indefinite articles, quantifiers, demonstratives, possessives.	‘a’, ‘both’, ‘either’, ‘my’, ‘some’
interjection	interj		
noun	n		
numeral	num	Used for cardinal and ordinal numbers, and also for numerical uses of lemmas such as *nothing* and *nought*	*She’s five foot nothing; nought point three*
prefix	pref	Used for lemmas that combine freely and can generate closed forms	‘macro-’ as in *macroclimate, macrostructure etc.*
preposition	prep		
pronoun	pron		
suffix	suff	Used for lemmas that combine freely and can generate closed forms	‘-made’ as in *homemade, homemade*; ‘-phone’ as in *francophone etc.*
verb: auxiliary	v_aux	auxiliary verb	There are only three: ‘be’, ‘do’, ‘have’
verb: lexical	v	straightforward lexical verb: the default verb type	‘maintain’, ‘navigate’, ‘operate’, ‘persist’, ‘run’….
verb: modal	v_mod	modal verb	‘may’,’ might’, ‘must’ etc.
verb: phrasal	phr_v	phrasal verb	more details at 3.1.3

Table 1: Parts of speech

5.1.1. Parts of speech not available as search options

DANTE also uses the parts of speech **prp_adj** and **ptp_adj**, and these are not available in the POS search. They refer, respectively, to present participle adjectives and past participle adjectives, and are used only in SUBFORMs (5.1.2), not at headword level. They are applied to adjectival participles which are not sufficiently frequent to qualify for full headword status. Examples include: (prp_adj) 'a *mesmerising* story', 'the *roaring* jet engines'; (ptp_adj) '*maddened* with pain', 'a non-slip *rubberised* surface'.

5.1.2. The SUBFORM field

The SUBFORM field is not available as a search option on the website, but you will sometimes see it in an entry. A SUBFORM is a specific form of the headword which is itself a lexical unit. SUBFORMs include:

- present participle adjectives and past participle adjectives (5.1.1)
- plural forms of nouns with their own distinct meanings and uses (*marbles, dealings*)
- nouns with obligatory 'the' (*the Madonna*)
- capitalized forms of lower case headwords (*King*)
- hyphenated forms (the verb *slam-dunk* at the noun entry *slam dunk*)
- combining forms (*-haired* at *hair, -metre* (as in 'a 1000-metre race') at *metre*).

5.1.3. Phrasal verbs

There is no watertight definition of the category 'phrasal verb'. In DANTE, we recognise three types of phrasal verb:

- verbs with an adverb particle: *get up, point out*
- verbs with a preposition particle: *see through* (someone's plans), *part with* (your money)
- verbs with both types of particle: *make off with, refer back to.*

Phrasal verbs in DANTE have the part-of-speech label **phr_v**. To search on the website for a phrasal verb, select 'verb:phrasal' in the drop-down list of parts of speech. As with other verbs, phrasal verb searches can be further refined using the 'inherent grammar' condition (5.2).

5.2 Inherent grammar

When searching for an adjective, adverb, verb, phrasal verb, or noun, you can refine your search by specifying the item's 'inherent grammar'. For example, the POS label 'adverb' will find *any* type of adverb, but if you add an inherent grammar condition you can narrow your search to find (for example) only adverbs of degree. If using the Advanced search mode on the website, you will see that the drop-down list for 'inherent grammar' is tailored to each of the relevant parts of speech. The codes are explained in Tables 2-6. The following parts of speech have no inherent grammar options in DANTE: conjunction, determiner, interjection, prefix, preposition, pronoun, suffix.

In the database, inherent grammar codes appear in a field called GRAM, but in the entries shown on the website, you will see only the inherent grammar code itself (following the part of speech label), not the field name GRAM.

5.2.1. Inherent grammar: adjectives, adverbs, verbs, phrasal verbs

The available codes are explained, respectively, in Tables 2, 3, 4 and 5. In DANTE policy, phrasal verbs *always* have an inherent grammar code, but for nouns, adjectives, adverbs and 'standard' verbs, inherent grammar codes are not always required.

Code	Explanation	Examples
(no code)	default: an adjective that can occur in both attributive and predicative uses	*small, happy, green*
attr_only	an adjective that is attributive only	*mere* (a *mere* mortal)
comb	combining form: a form of a headword which can combine with other words to produce an adjectival compound. Combining forms appear in the SUBFORM field (above, 2.1.1)	-*conscious* (health-*conscious* consumers), -*maintained* (a poorly-*maintained* building)
pertnm	a pertainym adjective: an adjective that means 'pertaining to X'; pertainyms are attributive only, have no comparative or superlative form, and are never modified	*marital* bliss, *political* acuity, *racial* sensitivity
post_mod	post-modifier adjective	mayor *elect*, heir *apparent*
predic_only	an adjective that is predicative only	*alone, mindful*

Table 2: Inherent grammar: adjectives

Code	Explanation	Examples
(no code)	default: a manner adverb	*accidentally, jauntily, patiently*
deg	a degree adverb	*seriously* ill, *unbelievably* stupid
sent	a sentence adverb (typically sentence-initial, but can appear in any position)	*hopefully,* it won't rain; *personally* I think he's lost it; we could take the train or, *alternatively*, we could drive
view_pt	a viewpoint adverb ('from the X point of view')	*politically* serious, *socially* inept

Table 3: Inherent grammar: adverbs

Code	Explanation	Examples
(no code)	default: a verb whose use is not restricted	*say, walk, accuse*
imper_inf	a verb used only in the imperative or infinitive	something we need to *beware* of*, let* the ceremony begin
impers	an impersonal verb	it *rains* a lot in April*,* it's *snowing*
passive	a verb occurring in only in the passive	it is *rumoured* that...
reciproc	a reciprocal verb	John and Mary *marry,* John *marries* Mary

Table 4: Inherent grammar: verbs

Code	Explanation	Examples
v_adv	a phrasal verb consisting of a verb with an adverbial particle	*pass* the message *on, pass on* the message, the custom *died out*
v_adv_prep	a phrasal verb consisting of a verb with adverbial and prepositional particles	*come up with* an idea
v_prep	a phrasal verb consisting of verb with a prepositional particle	*look at* the screen, *ran through* all his money

Table 5: Inherent grammar: phrasal verbs

5.2.2. Inherent grammar: nouns

Nouns exhibit a wide range of grammatical behaviours. Categorising them is notoriously difficult, and we are not aware of any system that accounts for all possible cases. The approach used in DANTE is a pragmatic one, and DANTE lexicographers used a flowchart to determine which (if any) of ten possible codes should be applied to a noun in a given lexical unit. The inherent grammar codes used for nouns are explained in Table 6, and the lexicographers' 'noun flowchart' is in Table 7.

Code	Explanation	Examples
(no code)	default: a countable noun. Includes concrete objects, countable senses of words with uncount senses, and 'type' and/or 'unit' senses of mass nouns	*cup, dog, idea, teacher, organisation, risk, coffee* ('three *coffees*, please'), *wine* ('the *wines* of New Zealand')
c_u	a noun that is usually countable but has (1) generic-type uncountable uses, or (2) senses which combine the ideas of 'the act or an instance of X'	(1) *dinner* (three *dinners*, *dinner* is at 7), *bus* (three *buses*, go by *bus*); (2) *realignment, rationalisation* (a series of *realignments/rationalisations*; in need of *realignment/rationalisation*)
mass	a mass noun: applied to conceptually mass items, which are typically substances of some kind, including fabrics, foods, liquids, chemical elements and compounds, etc. Names of colours are also coded as 'mass' in DANTE, but (like many mass nouns) they can also have separate senses to cover 'type' or 'unit' uses	*blood, sand, sewage, bedding, pasta, oxygen, titanium, heroin, purple, wine*
pl_0	a noun with unchanged plural form	*sheep, gasworks*
pl_also_0	a noun which has a regular plural form, but which can also have a 'collective' or 'hunting' plural where the form is unchanged	five *herrings,* go fishing for *herring*; equipped with four 20mm *cannons*, we could hear the sound of *cannon*
pl_only	a noun occurring in the plural only	*ablutions, algae, scissors*
proper	a proper noun: typically a name; always capitalised; most operate normally without a definite or indefinite article; some require a definite article; some can be pluralised	*Edinburgh, Christmas, Monopoly, Napoleon, the White House, the Maldives*
u_c	a noun that is usually uncountable but has countable uses in certain predictable contexts:. This code is used mainly for infrequent, non-core lemmas or LUs, when conflating uncountable and countable uses, (typically covering 'the act/process of X' and 'an instance of X' or 'the result of X').	*nominalisation, popularisation*
uncount	an uncountable noun, rarely if ever found in the plural form. Applied to: abstract nouns (typically with definitions that start with any of: 'an act, state, quality, feeling etc'); academic subjects; schools of thought; medical conditions; sports; musical genres, etc. The code 'uncount' is also applied in DANTE to items which are coded 'singular' or 'singular only' in some dictionaries, as in: a *riot* of colour, grab a *bite*, the test was a *breeze*	*anger, maintenance, surrealism, jazz, hockey, weather, geography, bronchitis, asthma*
v_sg	a noun denoting a group of people but taking a singular verb	*government*, *team* (especially in American English)
v_sg_pl	a noun denoting a group of people and taking a singular or plural verb	*government*, *team* (especially in British English)

Table 6: Inherent grammar: nouns

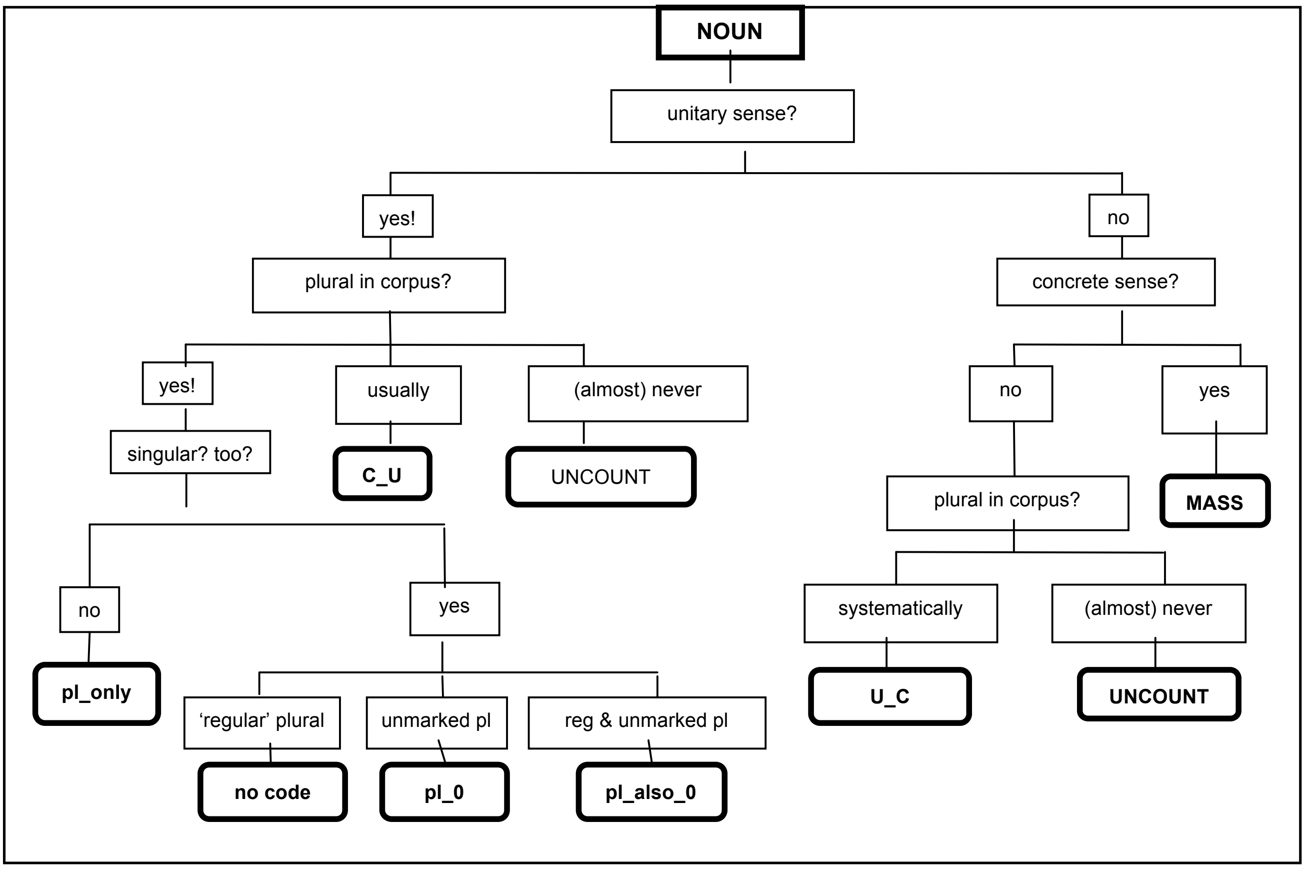


Table 7: Noun flowchart (for determining inherent grammar)

5.3 Syntactic context

When searching for an adjective, noun, or verb, you can refine your search by specifying a particular 'syntactic context'. Syntactic context codes are used for describing those syntax patterns (or constructions) which – according to the evidence in our corpus – are associated with a particular lexical unit. Adding the search condition 'syntactic context' brings up a drop-down list which includes all the codes available for the part of speech you are searching on. Syntactic context codes are optional for adjectives and nouns, but all verbs have at least one. The syntactic context codes are explained in Tables 8 (adjectives), 9 (nouns), and 10 (verbs).

In entries viewed on the website, a syntactic context code is preceded by the word **STRUCTURE** (in red). In entries viewed in the database itself, the code is preceded by: STRADJ (for adjective codes), STRN (for nouns), or STRV (for verbs).

Code	Explanation	Examples
AVP_premod	pre-modified by adverbial	*minimally* cooperative, *significantly* different
PP_X	preposition phrase, with named preposition, e.g. PP_X **at**	amazed *at all this*, happy *with what he had*, delighted *for all of you*
PP_X-NP-Ving	preposition phrase + noun phrase + gerund	aware *of him laughing*
PP_X-Ving	preposition phrase + gerund	tired *of living*, interested *in knowing about it*
PP_X-cl_wh	preposition phrase + wh-clause	curious *about where I can find that information*, absorbed *in what she was reading*
PP_for-Vinf_to	copula + for + 'to' infinitive	is it possible *for you to do it*

Code	Explanation	Examples
Vinf_to	‘to’ infinitive	happy *to know*, eager *to do it*
Ving	gerund	busy *repairing her bicycle*
it-PP_X-Vinf_to	‘it’ preposition phrase + ‘to’ infinitive	*it* was necessary *for her to go*
it-Vinf_to	‘it’ + copula + ‘to’ infinitive	*it is/seems etc.* imperative *to do…*
it-that_0	‘it’ + copula + that-clause with or without explicit ‘that’	*it is/seems etc.* clear *(that)…*
it-wh	‘it’ + copula + wh-clause	*it is/seems etc.* clear *why/how etc. …*
that_0	indicative that-clause with or without explicit ‘that’	I’m sure *(that) you will understand*
that_0_subj	subjunctive that-clause with or without explicit ‘that’	they were insistent *(that) he join in*
wh	wh-clause	curious *where he was*, curious *what to expect*

Table 8: Syntactic context: adjectives

Code	Explanation	Examples
AJ_pert	pre-modified by pertainym adjective	*educational* institution, *chemical* reaction
AVP_post_mod	post-modified by adverb	the journey *home*
N_mod	pre-modified by a noun	*sea* view
N_premod	pre-modifying a noun	rose *petal*
PP_X	preposition phrase, with named preposition, e.g. PP_X **at**	a look *at the screen,* a letter *from home,* an alliance *between the two parties*
PP_X-NP-Ving	preposition phrase + noun phrase + gerund	the thought *of him going*
PP_X-PP_X	preposition phrase x 2	an argument *with John about money*
PP_X-Ving	preposition phrase + gerund	the thought *of going*
PP_X-cl_wh	preposition phrase + wh-clause	questions *about what courses are offered*, concerns *about which to support*
PP_for-Vinf_to	for + ‘to’ infinitive	his wish *for them to be there*, her anxiety *for him to get better*
Vinf_to	‘to’ infinitive	his desire *to be present*, her need *to behave well*
if	whether/if clause	the question *whether he would go*
it_constrn	‘it’ + copula + ‘to’ infinitive	*it’s* a mistake *to think about it*, *it’s* fun *to swim in the sea*
that_0	indicative that-clause	the news *(that) he had arrived*
that_0_subj	subjunctive that-clause	their request *(that) he go with them*
wh	wh-clause	the reason *why he left*, the question *when to go*

Table 9: Syntactic context: nouns

Code	Explanation	Examples
AJP	adjective phrase	you seem *sad*, he looks *taller than you*
AVP	adverb phrase	he had aged *badly*, act *responsibly*
NP	noun phrase	I like *him*, I heard *a story*, I dropped *the metal lid*
NP-AJP	noun phrase + adjective phrase	paint *the wall green*, we found *it very dull*
NP-AVP	noun phrase + adverb phrase	allow *them through*, they floated *it downstream*
NP-NP	noun phrase x 2	crown *him king*, show *me your essay*, give *her a book*, I cooked *her a curry.*
NP-PP_X	noun phrase + preposition phrase	change *the colour to white*
NP-PP_X-Ving	noun phrase + preposition phrase + gerund	admire *somebody for doing*
NP-Part	noun phrase + directional or locative particle	push *it away/into the room*, put *it down/under the table*
NP-Part_X	noun phrase + specific particle	push *it in*, lever *it up* gently
NP-V_ptp	noun phrase + past participle	get *your hair cut*, have *the house valued*
NP-Vinf	noun phrase + bare infinitive	make *him leave*, she let *him go*
NP-Vinf_to	noun phrase + ‘to’ infinitive	we want *you to leave, she dared him to do* it
NP-Ving	noun phrase + gerund	she watched *the children playing,* I heard *him leaving*
NP-cl_that_0	noun phrase + that-clause	tell her *(that) he’s here*
NP_refl	reflexive	cross *oneself*
NP_refl-PP_X	reflexive + preposition	abandon *oneself to a life of pleasure*, dedicate *oneself to a cause*
NP_refl-PP_X-Ving	reflexive + preposition + gerund	dedicated *himself to caring* for her
NP_refl-Vinf_to	reflexive + ‘to’ infinitive	he wouldn’t demean *himself to apologise*
NP_refl-Ving	reflexive + gerund	they enjoyed *themselves swimming*
PP_X	prepositional phrase	they looked *at the screen*, that depends *on the situation*, we thought *of you*
PP_X-NP-Ving	prepositional + noun phrase + gerund	we objected *to him going*
PP_X-PP_X	prepositional phrase x 2	argued *with John about money*
PP_X-Vinf_to	prepositional + ‘to’ infinitive	I would prefer *for him to go*, they looked *to him to do* it
PP_X-Ving	prepositional phrase + gerund	don’t insist *on doing* it, I thought *of going*
PP_X-cl_wh	prepositional phrase + wh-clause	he enquired *about which train I was taking*, it depends *on what you mean*

Code	Explanation	Examples
Part	directional or locative particle	Run *away / into the room*, sit *there / in that chair*
Part_X	specific particle	work *away*, chatter *on*
Quo	quote (direct speech)	*'Get out of here!'* she shouted, he mumbled, *'Why should I?'*
Quo-NP	quote + noun phrase	'*You said you would*,' she reminded him
Vinf	bare infinitive	he dared not *do* it
Vinf_to	'to' infinitive	I love *to visit* them, I tried *to go*
Ving	gerund	she likes *ironing*, I hate *washing* dishes
_0	no complement	*it disappeared, she shouted*
if	whether/if clause	she asked *if I knew,* he wondered *whether she'd found it*
it_constrn	'it' construction	*it seemed there was a mistake, it rests with him to do this*
subj_NP	specifying the noun subject of the headword	(only when needed for sense distinction) *mountain* looms, *disaster* looms
that_0	indicative that-clause	I hear *(that) he's arrived*
that_0_cond	conditional that-clause	he wishes *(that) she would go away*
that_0_subj	subjunctive that-clause	they demanded *(that) he obey them*
wh	wh-clause	I forgot *what I'd said*, she guessed *when you had arrived*, I know *how you feel*
wh-Vinf_to	wh-word + 'to' infinitive	I didn't know *what to say*, watch *how to do it*

Table 10: Syntactic context: verbs

5.3.1. The use of 'X' in syntactic context codes

You will notice that many of the codes include an 'X'. This is used in codes which include a particle or preposition, and the X indicates that a *particular* particle or preposition is required. All codes which include an X are followed by a specific preposition or particle.

For example, the pattern illustrated in the sentence:
It takes a while to ***acclimatise to*** *the humid conditions*
is coded as PP_X (preposition phrase with named preposition) and is followed by '**to**'.

5.3.2. The use of 'Part' in syntactic context codes

The verb codes include four which contain 'Part'. These are:

- Part
- NP Part
- Part_X
- NP Part_X

Part' is used for recording the use of a particle (an adverb or preposition) in a verb which is not a phrasal verb. For example, the verb *amble* can occur in sentences like 'we ambled *across* the wide lawn' or 'they were ambling *along*', and in cases like this, the verb is coded as 'Part'. The addition of 'NP' indicates that a noun phrase comes after the verb and before the particle (as in 'roll it *across* the floor/*up* the hill'). In these codes, various particles are possible, and the examples will show a range of typical instantiations.

The addition of 'X' (cf. 5.3.1. above) indicates that *one specific* particle is required (and the particle is named). For example 'we all filed *in*' (coded 'Part_X in') or 'nail the planks *together*' (coded 'Part_X together'). Where there is no 'X' (in other words, where a range of particles is possible), the code is followed by a 'subcode' which indicates whether the particles show direction (Part_dir) or location (Part_loc). For example, the verb *rush* includes the code 'Part' and subcode 'Part_dir', to cover uses such as 'she rushed *past* me/rushed *up* to him/rushed *into* the room'.

6. Inflections in DANTE

Inflections are not shown explicitly in the Dante database. Lexicographers were briefed to include different forms of nouns or verbs in the example sentences, including:

- singular and plural forms of nouns
- singular and plural uses of nouns with 'zero

plural' forms
- a range of verb forms

Irregular inflected forms (such as *children, mice, addenda, throve*) have their own (short) entry in the database, consisting of a cross-reference to the base form of the lemma.

7. Labels in DANTE

Labels are applied to any item in the database which is not part of the core, 'unmarked' vocabulary of English. Items which – on the evidence of the corpus – are characteristic of a particular text-type or speech community will attract a label. Such items may include entire headwords or lower-level components like grammar codes or examples. but labels are most frequently associated with individual lexical units.

DANTE has six categories of label, allowing us to mark any item according to:

- attitude
- regional variety
- register
- style
- time
- subject field (or domain)

In the entries you see as the result of a search on the website, labels appear in italics within (italicised) square brackets: for example *[offens], [AmE], [journ]*. The exception is domain (or subject-field) labels: see 7.6.

7.1 Attitude

Three labels are available for indicating the attitude of the speaker or writer:

- apprec ('appreciative'), as in: *tireless, slender*
- offens ('offensive'), as in *half-caste, poof*
- pej ('pejorative'), as in *moralize, paltry, bimbo*

7.2 Regional variety

Some of the regional-variety labels in DANTE are idiosyncratic and require explanation. The DANTE database was developed for Foras na Gaeilge as a launchpad for its *New English-Irish Dictionary*, and this has implications for vocabulary coverage and labelling. On the one hand, many major World Englishes (such as South African English or Indian English) are not described systematically (though high-profile usages from any variety will be covered). On the other hand, DANTE provides extensive coverage of Hiberno-English (the variety of English spoken in Ireland), and the corpus resources used by DANTE lexicographers included a specially-created 25-million-word corpus of Hiberno-English.

Consequently, the label 'BrE' (British English) has a different application in DANTE than in standard dictionaries. Conventionally, a 'BrE' label is applied to usages typical of the British Isles as a whole (including Ireland), in contrast to 'AmE' usages. But in DANTE, such usages are labelled 'non_AmE', while 'BrE' is reserved for items typical of usage in mainland Britain but not found in Hiberno-English. Conversely, items specific to Hiberno-English and not familiar to British speakers are labelled 'HibE'.

Examples of items labelled 'BrE' in DANTE include *barnet* (someone's hair or hairstyle), *comp* (a comprehensive school) and *stopping train* (a slow train that stops at every station on the route); none of these is used in Hiberno-English. Examples of items labelled 'HibE' include *brutal* (in the sense of 'awful')*, eejit* (idiot), and *deadly* (very fine or attractive).

The complete set of regional labels in DANTE is:

- AmE: American English
- non_AmE: explained above
- BrE: explained above
- HibE: Hiberno-English
- AusE: Australian English
- Scot: Scottish English

7.3 Register

The four register labels in DANTE are used for indicating the formality level of the labelled item:

- fml ('formal'): words, senses, or expressions that are typical of formal usage: examples include *admonitory, remediable, munificence, ameliorate*
- inf ('informal'): examples include *bolshie, hellish, megabucks, on the razzle*
- vinf ('very informal'): while 'informal' uses are found in a wide range of text types, this label denotes items found only in very informal discourse, typically between people who know each other well or belong to the same social grouping. This equates to what many dictionaries label as 'slang' (a contentious term). Examples include: *charlie* (cocaine), *pants* (as an adjective: very bad), *fanny about* (mess around).
- vulg ('vulgar'): vulgar uses may cause offence and equate to what some dictionaries label as 'taboo' (which we see as an outmoded term and concept). Examples include the familiar four-letter words, and items like *motherfucker, prick, shitless* (as in 'scared/bored shitless').

7.4 Style

7.4.1. Labels for foreign borrowings

The labels in this category are used to mark borrowings that remain noticeably foreign and are often not pronounced in an English manner. Well-integrated borrowings such as *macaroni* are not labelled – though, as is well known, the boundary between the two types is impossible to draw precisely.

The following labels are available:

- Fr (French): *mangetout, ménage à trois, objet d'art*
- Ger (German): *schadenfreude, Zeitgeist*
- Lat (Latin): *modus operandi, non sequitur, obiter dictum*
- Span (Spanish): *mojito, mestizo, paso doble*
- For (any borrowing not covered by the labels above): *feng shui, edamame, perestroika*

7.4.2. Other style labels

This is a somewhat heterogeneous category, and these labels often occur in combination with register, attitude, or domain labels. It is important to stress the difference between a particular *style* of speech or writing and the *domain* which a text belongs to (see 7.6 for domain labels). For example, words like *abatement, predecease* and *heretofore* belong to a legal 'style' of writing ('legalese') and get the style label 'leg'; words like *alibi, foreman* (of a jury), and *bail* are words belonging to the subject field of law, so get the domain label 'Law'.

The following labels are available:

- TM (trademark): *Blu-ray, frisbee, Portakabin, Prozac*
- child (child language): *grown-up, poo*
- drugs (drug-users' slang): *mainline, charlie, re-up*
- euph (euphemism): *nether regions* (genitals), *economical with the truth*, *pass away*
- hum (humorous): *mugshot, nookie, bridezilla*
- iro (ironic): *princely* (sum), *dulcet* (tones) *pearl of wisdom*
- journ (journalese): *beleaguered, probe, wed*
- leg (legalese): *abatement, predecease, heretofore*
- lit (literary): *bounteous, morrow, asunder*
- pc (politically correct language): *person of colour, challenged* (mentally, visually etc)
- prov (proverb): *pride comes before a fall, too many cooks*
- spok (spoken: rarely found in written English): *anyways, and your point is? bro*
- tech (technical usage: often used in combination with a domain label): *macromolecular, meiosis, anisotropic*
- youth (young people's slang): *rad, boyf, respect!*

7.5 Time

DANTE is essentially synchronic, but it includes some items which are in the process of losing their currency and are now rarely heard (labelled 'dat', or dated), and others that are virtually never found in contemporary discourse (labelled 'obs', or obsolete). The latter are included only if users are likely to come across them in classic works of literature.

Examples include:

- dat: *poppycock, betrothed, blotto*
- obs: *apothecary, pox, dropsy*

7.6 Domain

One of DANTE's most valuable features is its extensive use of domain (or subject-field) labels. The editorial team had available to them **156** domain labels, and were encouraged to apply them whenever appropriate. In the entries you see as the result of a search on the website, domain labels appear in capitals within square brackets, in the form of (usually) self-evident abbreviations: thus [BOT] indicates an item from the domain of botany, and [SOCIOL] an item typical of texts about sociology.

IN DANTE's Advanced Search mode, you can search for lexical units with a domain label by clicking the item 'subject field' in the left-hand dropdown list. The right-hand drop-down list includes the labels themselves. Note, however, that this list does not include the full set of 156 domain labels. Instead, it provides a subset of **28** domains, most of which act as superordinates and subsume many related domains. (Obviously, licensed users of the full database can search for any of the 156 domain labels.) Table 11 shows the 28 labels available for searches on the website, and indicates any other labels which these subsume.

Label in drop-down list	Additional labels this covers
Agriculture	botany, horticulture
Art	ceramics, fashion, photography
Artifacts	clothing, cosmetics, furniture, tools.
Calendar	(none: this label covers items like days of the week and names of festivals)
Colours	(none: this label is applied to all colour terms)
Communication	telecommunication
Culinary	(none: this label covers all cooking vocabulary)
Education	(none)
Engineering	aerospace, automotive , chemical engineering, civil engineering, electrical engineering, machinery, mechanical engineering, mining

Finance	accountancy, economics, finance, insurance, tax
Government	(none)
Humanities	architecture, astrology, heraldry, history, sociology, philosophy, mythology
Industry/ Employment	business administration, commerce, construction, fishing, hair-dressing, plumbing, publishing, public relations, surveying, textiles, tourism, transport
IT	(none)
Law	police
Leisure	climbing, collecting, darts, do-it-yourself, table games
Linguistics	(none)
Literature	(none)
Mathematics	measurement units, statistics
Medicine	anatomy, health & fitness, pharmacology, physiology , psychology/psychiatry
Military	air force, army, navy, weaponry
Music	(none)
Nautical	(none)
Performing arts	cinema, theatre, dance
Politics	(none)
Religion	(none)
Science	anthropology, archaeology, astronomy, biochemistry, biology, chemistry, cosmology, dentistry, ecology, electronics, genetics, geography, geology, insects, meteorology, mineralogy, optics/ophthalmology, physics, veterinary science, viticulture, zoology
Sport	American football, archery, athletics, badminton, baseball, basketball, bowls, boxing, cricket, curling, cycling, equitation, fencing, football, Gaelic football, golf, gymnastics, hockey, hurling, horse-racing, hunting, ice hockey, ice-skating, lacrosse, martial arts, motor racing, polo, rowing, rugby, sailing, shooting, softball, surfing, swimming, table tennis, tennis, water sports, windsurfing, winter sports, wrestling

Table 11: Domain (or subject-field) labels

8. Multiwords in DANTE

This section explains how multiword units and expressions are treated in DANTE. Many different types of word+word combination were recorded by the lexicographers using the tags PHRASE, CHUNK, COLLOC, CPD (compound), ITEM (itemiser), and SUPPREP (support preposition). These elements are not searchable using the web interface, but they may appear in the entries returned by a search, so they are explained here.

Phraseology is a ‘scalar’ feature of language. Multiword combinations encompass a huge range of language events, from fixed, opaque idioms (‘for good measure’) to completely open combinations (‘a large house’). DANTE has a number of strategies for recording such items, but the boundaries between each type are not always clear. So the question of where a given combination fits best often comes down to editorial judgment. For example, the recurrent string ‘fit/match/answer a description’ is recorded in DANTE as a ‘phrase’ (a distinct LU with its own definition) but could, arguably, have been treated as a ‘chunk’ (8.2.1). In this case, and many similar cases, there is no ‘right’ answer.

8.1 Phrases, phrasal verbs, and compounds

Idiomatic phrases, phrasal verbs (5.1.3), and compounds are ‘nested’ : that is, they are handled in the entry for the lemma to which they are related: for compounds and phrasal verbs, the first word; for phrases, the first ‘lexical’ word. They appear after the LUs (or senses) of the base form, in a section called (in the database) the ‘Multiword Expression Block’ (or MWEBlk). Thus at the lemma *map*, the eight LUs of *map* itself (four noun senses and four verb senses) are followed (in this order) by:

- two phrases (*off the map, on the map*)
- one phrasal verb (*map out*)

- four compounds (*map maker, map projection, map reading, map reference*)

Each of these is an LU in its own right (so it has its own POS, labels, grammar codes, examples etc). In some cases items of this type consist of more than one lexical unit: the phrase *off the map*, for example, has three separate LUs, each with their own definitions and examples. In the entries shown on the DANTE website, each section (for phrases, phrasal verbs, and compounds) is signalled by a heading in red.

8.2 Multiwords which are not lexical units

DANTE records many other recurrent multiword strings which do not have the status of full lexical units. In addition to 'phases', several other options are available. These are:

- chunks
- collocations
- support verbs
- support prepositions
- itemisers

8.2.1. Phrase or Chunk?

A phrase is a full lexical unit, and (like any LU) requires a definition. Chunks, on the other hand, appear within an LU and do not have their own definitions.

Phrases in DANTE are non-transparent combinations, whose meaning or communicative function cannot be inferred from its components. Phrases span a range of types, from the stereotypical idiom *kick the bucket* (completely opaque), to cases where one or more of the component words has one of its 'usual' meanings, but the meaning of the phrase is still not retrievable: thus *look daggers at* does involve 'looking at' someone, but it is nevertheless not wholly transparent: it therefore needs a definition, and it therefore qualifies as a discrete LU to be treated as a phrase.

The category 'chunk' was introduced for a combination which is non-idiomatic, but frequent enough in the corpus to be worth recording as a significant fact about the lemma it belongs to. In its *form*, a chunk has some of the features of a phrase, in that the selection of words may be idiosyncratic. But its meaning is readily deducible from its component parts: hence it needs no definition, and hence it does not qualify as a separate LU. Whereas phrases appear in the 'Multiword Expression Block' at the end of an entry, chunks are covered in the LU whose meaning they invoke.

Examples of chunks include:

- *go into administration* (at the LU of *administration* that refers to the disposition of insolvent companies etc)
- *by/on one's own admission*
- *out of deference to*
- *on a daily basis*
- *decide for oneself*

8.2.2. Collocation

A collocation is a two-word combination consisting of the lemma and another lexical word (a 'collocate') with which the lemma frequently occurs. For example at the lemma *pool*, the adjective collocates *outdoor, indoor, heated*, and *private* are listed at the noun LU referring to a swimming pool; and the noun collocates *resources, funds*, and *data* are recorded as typical objects of the verb LU meaning 'to put things together for collective use'.

In the entries shown on the website, collocates are listed (following the word COLLOCATES in red), and then provided with examples. Using the 'Word Sketch' function in the Sketch Engine corpus query system, DANTE lexicographers identified and recorded the most frequent collocates for each LU, and the result is a rich and systematic account of collocation in English.

8.2.3. Support verbs

A support verb is a 'light' verb in a verb+noun combination in which the verb makes little semantic contribution. DANTE recognises five support verbs: *do, give, have, make, take*. A support verb+noun combination typically paraphrases the verb cognate of the noun. For example, the combination *take a walk* (where *take* is the support verb) is equivalent to the verb *to walk*. Other examples include:

- *do the packing*
- *give a salute*
- *have a quarrel*
- *make a promise*
- *take a shower*

Support verbs are one of the search parameters available in the Advanced Search mode on the website.

8.2.4. Support prepositions

A support preposition is a preposition which frequently occurs *directly before* a noun. For example, when *peace* refers to 'freedom from disturbance', it is frequently found in sentences like: 'a place where you can do your work *in peace*' or 'spaces where people can walk and cycle *in peace*, away from traffic'. At the relevant LU of *peace*, the support preposition 'in' is shown after the words SUPPORT PREP (in red). Other examples include:

- *at rest*
- *in hysterics*
- *on vacation*
- *by helicopter*

8.2.5. Itemisers

An itemiser is a word that is used in conjunction with a concrete noun to instantiate the idea of 'a piece [of the noun]'. Itemisers are recorded in DANTE when corpus data indicates they are frequent. Examples of itemisers include:

- *a drop of blood*
- *a head of broccoli*
- *a piece/item/article of clothing.*

Different LUs of the same lemma may have different itemisers: thus for the ‘mass’ sense of *chocolate*, a common itemiser is ‘bar’, while for its countable use (‘small sweet or piece of candy’), ‘box’ is often used.

9. Miscellaneous

There are four other fields which are not available as search conditions on the website, but may appear in entries retrieved by a search:

- corpus pattern
- pragmatics
- link
- xref (cross reference)

In all cases, the field is shown in red (in website entries), and the relevant information follows. These are explained here.

9.1 Corpus pattern

Corpus data sometimes reveals recurrent features of a word’s behaviour which are not covered by any of the grammatical or phraseological categories described above. The corpus pattern field is used for recording such information. For example, the verb *gall* often appears in ‘cleft’ sentences such as:
–What galled me even more was her insistence that…
–But what galls many motorists more is the fact that…

Since this is clearly a characteristic feature of the verb, it is recorded in the corpus pattern field as: ‘often in cleft sentences’.

Other examples include:

- *gag*: ‘always in progressive’ (*we were gagging for a drink*)
- *abide*: ‘usually in negative or broad negative environment’ (*I never could abide lobsters*)
- *demote*: ‘often passive’ (*he was demoted in the cabinet reshuffle*)

9.2 Pragmatics

Where appropriate, DANTE records information about a word’s pragmatics: for example, the connotations of a word, or what it tells us about the attitude of the speaker. Sometimes, this can be conveyed through the use of attitude labels (such as ‘apprec’ or ‘pej’: 7.1) or style labels (like ‘euph’ or ‘hum’: 7.4.2). In other cases, none of the available labels is adequate, so the pragmatics field is used. Examples include:

- *charm*: ‘can sometimes have connotations of manipulation’
- *constant* : ‘often expresses annoyance’
- *micromanage*: ‘expresses disapproval’

9.3 Links and XRefs

A ‘link’ is used to cross-refer to another item within the same entry. Most typically, links are used to point to a multiword expression (phrase, compound, or phrasal verb) whose meaning is closely related to the LU where the link appears. The ‘link’ field exists primarily to alert other lexicographers or translators using the database that there is an item relating to that LU further down in the same entry. Examples include:

- *access*: the LU referring to access to computer data or the Internet includes a link to the related compound *access provider*
- *allowance*: the LU meaning ‘the fact of taking something into account’ includes a link to the phrase *make allowances for*
- *auction*: the verb LU includes a link to the phrasal verb *auction off*

A ‘xref’ is used to cross-refer from one headword to another. The most common type of xref is from an ‘empty’ entry to the main entry (e.g. from *center* to *centre*). Otherwise, since most of the information in DANTE is machine-retrievable, explicit cross-references are used only sparingly.

10. References

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

Adapting the dictionary entry structure and DWS configuration for creating a dictionary aimed to be published on paper, online and as electronic dictionary software for PC and mobile

Anna Rylova
ABBYY / Moscow State University
E-mail: Anna_Ry@abbyy.com

Abstract

Publishing one and the same dictionary on paper, as a desktop application, online and mobile application is a common practice in most of publishing houses. But the type of media the dictionary will be published on must be taken into account on the very first stages of developing the dictionary concept, the Style Guide and defining the entry structure. The formats of dictionary publishing give a lot of new challenges to lexicographers who have to keep in mind the dictionary users' needs, apply linguistic theory and always think about how the dictionary is intended to be published. The factor of dictionary publishing media influences not only the editorial decisions, but also a Style Guide and in the same time it affects the Dictionary Writing System design. The DWS must be adopted for the creation of the dictionary entry that will have different structure and visualization in paper and electronic dictionary. In this paper we describe how dictionary data and a dictionary writing system can be adapted for the various formats in which dictionaries can be published today. Our research is based on our own experience of creating dictionaries within our team of lexicographers, developing a professional dictionary writing system ABBYY Lingvo Content and electronic dictionary ABBYY Lingvo (since 1989) for PC, mobile and online.

Keywords: dictionary writing system; dictionary publishing; entry structure

When creating a dictionary a lexicographer has to keep a lot of factors in mind - these are the linguistic theory issues and the dictionary end-user needs. Every time when the entry is being edited, a lexicographer tries to make it relevant and accessible to the dictionary users. Linguistic theory and end-user needs affect every aspect of the lexicographic work and play an important role in a dictionary Style Guide creation process. But in the era of electronic lexicography even more factors come into play. One of the important factors that have a great impact into process of dictionary making is the type of media on which the dictionary is intended to be published.

Publishing one and the same dictionary on paper, as a desktop application, online and mobile application is a common practice in most of publishing houses. But the type of media the dictionary will be published on must be taken into account on the very first stages of developing the dictionary concept, the Style Guide and defining the entry structure.

The common examples of differences between electronic and paper versions of the dictionary are as follows: 1) the electronic dictionary normally include collocations and idioms as separate entries to facilitate the search process while the paper dictionary usually includes the collocations in the entry with a main word; 2) the electronic dictionary can include more examples while the paper dictionary has a limited number of examples because of space limitation; 3) the electronic dictionary includes the full forms of labels while the paper dictionary includes abbreviations; 4) the electronic dictionary usually has the full headword forms in the entry while the in the paper dictionary the headword is normally replaced with the tilde; 5) dictionary entries in electronic and paper dictionaries have different visualization.

Though we are mentioning the electronic dictionaries in general, every type of electronic dictionary has its own specifics. For example electronic desktop applications often include some features for language learners e.g. special functions for words memorizing; online dictionaries have functions of logging the users' queries thus helping publishers to update the dictionary content, and surprisingly the mobile dictionaries have the problem of space (the same as paper dictionaries) because they have limited memory size and mobile gadgets often have a small screen.

So when creating a dictionary entry aimed to be published on paper, online and on handheld devices we have to think the entry structure over very thoroughly. The principle "one database – many dictionaries" give a lot of new challenges to lexicographers who have to keep in mind the dictionary users' needs, apply linguistic theory and always think about how the dictionary is intended to be published.

The factor of dictionary publishing media influences not only the editorial decisions, but also a Style Guide and in the same time it affects the Dictionary Writing System design. The DWS must be adopted for the creation of the dictionary entry that will have different structure and visualization in paper and electronic dictionary. The dictionary makers and DWS developers must work together to adjust the DWS for the needs of publishing the same dictionary on different media. It is preferable when DWS, electronic dictionary software for PC and mobile devices and online dictionary are provided by the same software developer as it reduces the time spent for negotiations and software customization.

In this paper we describe how dictionary data and a dictionary writing system can be adapted for the various formats in which dictionaries can be published today. Our research is based on our own experience of creating dictionaries within our team of lexicographers, developing a professional dictionary writing system ABBYY Lingvo Content and electronic dictionary ABBYY Lingvo (since 1989) for PC, mobile and online.

The most exciting experience of adapting the dictionary data we gained while publishing ABBYY Lingvo Comprehensive English-Russian Dictionary that was published electronically in 1990 and then permanently updated by ABBYY team of lexicographers with a new electronic version (desktop and online) published every one or two years. This dictionary was available on handheld devices (Pocket PC and Palm) since 2003 and on mobile phones since 2006. And it was only in 2007 when the first paper edition has been released. The lexicographers had a lot of new tasks while adapting the electronic dictionary to the paper publishing and the DWS developers helped them to configure the DWS for their needs.

In the presentation we describe the following details:
1) how the entry structure can be adapted for the different kinds of publishing – paper, online and mobile;
2) which DWS functions and which mark up should be used to avoid a lot of manual work and process the dictionary data automatically while preparing it for publishing;
3) what is the best way to make a comfortable visualization of dictionary in the DWS interface so that lexicographer can see how the entry will look like in electronic and paper dictionary.

All the three tasks were solved successfully while preparing ABBYY Lingvo Comprehensive English-Russian Dictionary for paper publication and the other dictionaries (about 40 of them) that were created by ABBYY lexicographers or in collaboration with publishing houses and individual professionals.

The Requirements for the dictionary writing system were worked out in close collaboration with editors, lexicographers, the developers of the electronic versions of the dictionary for different platforms, and the staff of the ABBYY Press publishing house.

First, the key features of each version of the dictionary were identified, which led us the following list of questions:

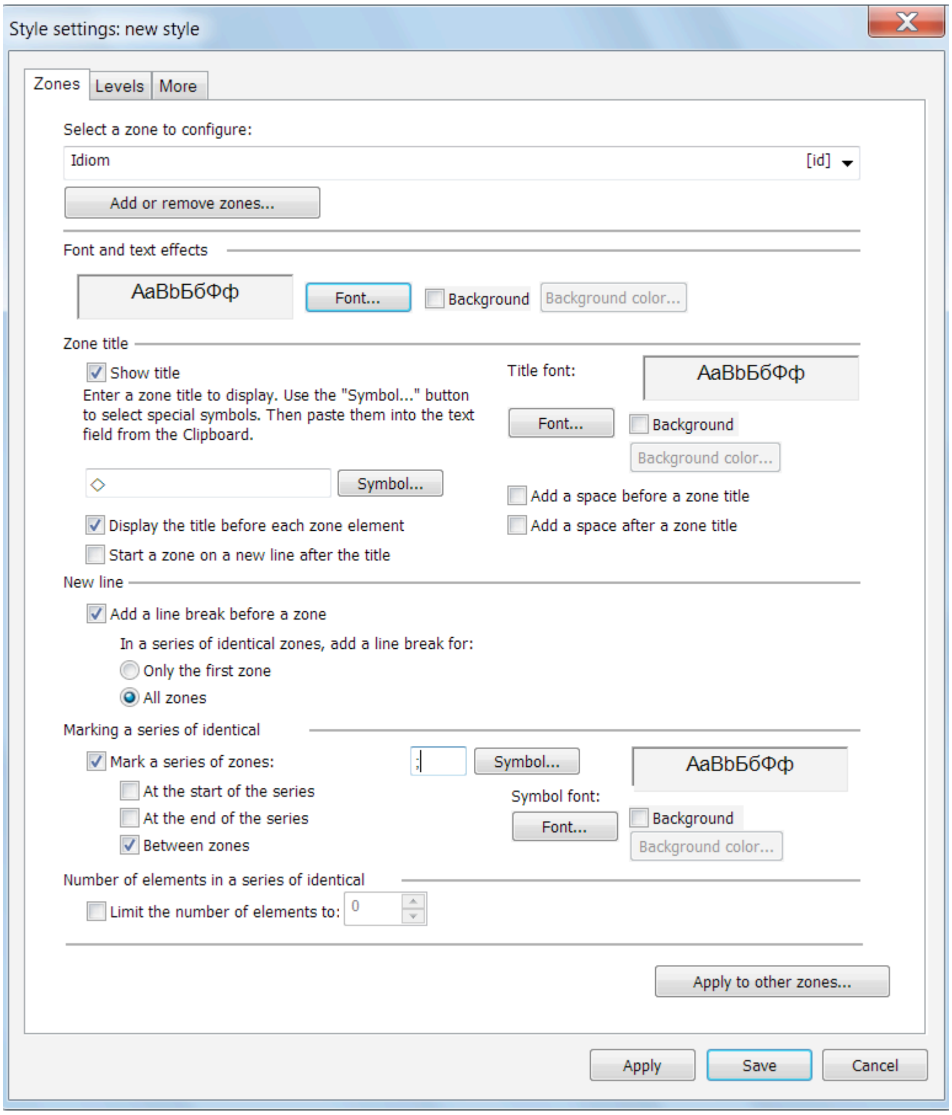
- ✓ What is the optimal size of dictionary entries (space is crucial)?
- ✓ How can we make the layout more user-friendly?
- ✓ How users will interact with dictionaries in different formats?
- ✓ What opportunities each publishing format provides?

  In answering the above questions, we developed different sets of styles and export settings for each entry section.

When creating a new dictionary or preparing an existing dictionary for publishing, the following choices are made for each entry section that is described in the Style Guide. These choices affect how the corresponding section is displayed in the various formats and how entries are exported from the dictionary writing system.

1) How the name of the section will displayed (e.g. “Idiom” for the electronic version and “◇“ for the printed version)
2) How section levels will be numbered (numbers for homonyms, senses etc.)
3) How the sections will appear on-screen and on the printed page (fonts, line breaks, colours)
4) How many same-level sections will be allowed (e.g. no more than 5 examples in the printed version)
5) Which sections will make it into each version of the dictionary (e.g. the etymology section may be excluded from the mobile version)
6) How headwords will be displayed in the body of the entries (they can be displayed in full or replaced with a swung dash)
7) How headword alternatives will be handled (e.g. separate entries in the mobile version and one entry with two headwords in the printed version)
8) How labels will be displayed (e.g. abbreviated in the printed version and spelled out in the electronic version)
9) Which entries will make it into each version (the paper or mobile version may contain fewer entries than the electronic version)

Once the appropriate settings are selected for each entry section, the styles and export settings are selected to suit the output format.



Style settings: new style

Zones | Levels | More

Select a zone to configure:

Idiom [id]

Add or remove zones...

Font and text effects

AaBbБбФф — Font... — □ Background — Background color...

Zone title

☑ Show title

Enter a zone title to display. Use the "Symbol..." button to select special symbols. Then paste them into the text field from the Clipboard.

◇ — Symbol...

☑ Display the title before each zone element

□ Start a zone on a new line after the title

Title font: AaBbБбФф — Font... — □ Background — Background color...

□ Add a space before a zone title

□ Add a space after a zone title

New line

☑ Add a line break before a zone

In a series of identical zones, add a line break for:

○ Only the first zone

● All zones

Marking a series of identical

☑ Mark a series of zones: ; — Symbol...

□ At the start of the series

□ At the end of the series

☑ Between zones

Symbol font: AaBbБбФф — Font... — □ Background — Background color...

Number of elements in a series of identical

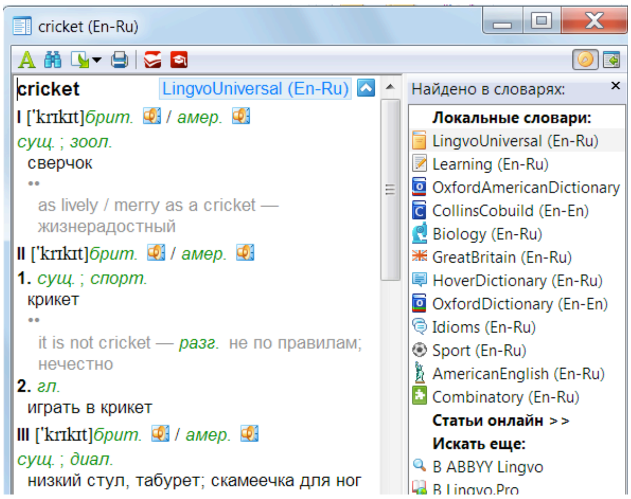
□ Limit the number of elements to: 0

Apply to other zones...

Apply — Save — Cancel

Figure 1. An example of DWS styles settings interface.

Once the lexicographer or editor has selected the settings for each entry section, they can proceed to create dictionary entries in a dedicated window of the dictionary writing system. This window displays the structure of the entry, which the lexicographer needs to fill with content. At any moment, the lexicographer can invoke one of the preview windows and view the entry as it will appear in the printed, mobile, CD, or online dictionary. The lexicographer can easily switch among the preview windows and adjust the display and export settings if required. The lexicographer can also at any moment export an entry into the printed format for proofreading purposes. There is also an option to export the entire dictionary or selected entries into the Lingvo electronic format.



cricket (En-Ru)

cricket — LingvoUniversal (En-Ru)

I ['krɪkɪt]*брит.* / *амер.*
сущ. ; *зоол.*
сверчок
••
as lively / merry as a cricket — жизнерадостный

II ['krɪkɪt]*брит.* / *амер.*
1. *сущ.* ; *спорт.*
крикет
••
it is not cricket — *разг.* не по правилам; нечестно
2. *гл.*
играть в крикет

III ['krɪkɪt]*брит.* / *амер.*
сущ. ; *диал.*
низкий стул, табурет; скамеечка для ног

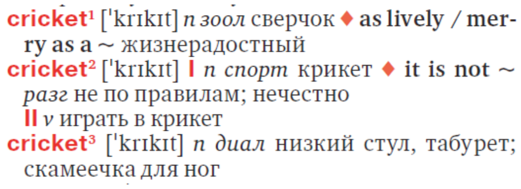
Найдено в словарях:

Локальные словари:
- LingvoUniversal (En-Ru)
- Learning (En-Ru)
- OxfordAmericanDictionary
- CollinsCobuild (En-En)
- Biology (En-Ru)
- GreatBritain (En-Ru)
- HoverDictionary (En-Ru)
- OxfordDictionary (En-En)
- Idioms (En-Ru)
- Sport (En-Ru)
- AmericanEnglish (En-Ru)
- Combinatory (En-Ru)

Статьи онлайн > >

Искать еще:
- В ABBYY Lingvo
- В Lingvo.Pro

Figure 2. ABBYY Lingvo electronic dictionary



cricket^{1} ['krɪkɪt] *n зоол* сверчок ♦ **as lively / merry as a ~** жизнерадостный
cricket^{2} ['krɪkɪt] **I** *n спорт* крикет ♦ **it is not ~** *разг* не по правилам; нечестно
II *v* играть в крикет
cricket^{3} ['krɪkɪt] *n диал* низкий стул, табурет; скамеечка для ног

Figure 3. ABBYY Lingvo paper dictionary

Thus, the lexicographer can easily see how an entry will appear in the various formats while still working on the dictionary.

A new century is posing new tasks for the lexicographers and new software technologies come to help the lexicographers in their everyday work. The collaboration between practicing lexicographers and software developers is very fruitful and helps both parties to facilitate their work and move further to new creative tasks. Thus the collaboration between ABBYY lexicographic team, ABBYY Lingvo Content software developers, ABBYY Lingvo desktop, mobile and online dictionary developers and ABBYY Press publishing house helped to develop new DWS features and a new methodology of dictionary creation.

References

Anokhina, J. (2010). Lingvo Universal English-Russian Dictionary: Making a Printed Dictionary from of an Electronic One. In *Proceedings of 14th EURALEX International Congress*. Leeuwarden, the Netherlands.

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

Dziemianko A. (2010). Paper or electronic? The role of dictionary form in language reception, production and the retention of meaning and collocations. *International Journal of Lexicography,* 23(3), pp. 257-273.

Grefenstette, G. (1998) The future of linguistics and lexicographers: will there be lexicographers in the Year 3000? In T. Fontenelle *et al*. (eds.) *Proceedings of the Eighth EURALEX Congress* Liege: University of Liege, pp. 25-41.

Hockey, S.M. (2000). Dictionaries and lexical databases. In *Electronic Texts in the Humanities: Principles and Practice.* Oxford & New York: Oxford University Press, pp. 146-171.

Heid, U. (2009) Aspects of Lexical Description for Electronic Dictionaries. In S. Granger, M. Paquot (eds*.) eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Kuzmina V., Rylova A. (2009) The ABBYY Lingvo electronic dictionary and the ABBYY Lingvo Content dictionary writing system as lexicographic tools. In S. Granger, M. Paquot (eds*.) eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Moon, R. (2008). Dictionaries and collocation. In S. Granger, F. Meunier. (eds.) *Phraseology. An Interdisciplinary Perspective*. Amsterdam: Benjamins, pp. 313-336.

Rundell, M. (2009) The Road to Automated Lexicography: First Banish the Drudgery… then the Drudges? In S. Granger, M. Paquot (eds*.) eLexicography in the 21st century: New challenges, new applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Rylova, A. (2010) Electronic Dictionary and Dictionary Writing System: how this duo works for dictionary user's needs (ABBYY Lingvo and ABBYY Lingvo Content case). In A. Dykstra, T. Schoonheim, (eds.) *Proceedings of the XIV Euralex International Congress*. Leeuwarden, 6-10 July 2010, Leeuwarden: Fryske Akademy.

Verlinde, S., Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra, T. Schoonheim, (eds.) *Proceedings of the XIV Euralex International Congress*. Leeuwarden, 6-10 July 2010, Leeuwarden: Fryske Akademy.

Pragmatic Components in the Slovene Lexical Database Meaning Descriptions

Mojca Šorli

Trojina, Institute for Applied Slovene Studies
Cesta v Kleče 16, 1000 Ljubljana, Slovenia
E-mail: mojca.sorli@trojina.si

Abstract

The possibility to analyse vast amounts of linguistic data has brought about changes both in methodology as well as in the ways we perceive certain language phenomena. A key insight gained by computational methods in language analysis is undoubtedly the importance of lexical co-occurrence and usage patterns for the description of lexical meaning. Corpus analysis and new methods in the analysis of pragmatic components of meaning have also yielded significant results in areas such as the treatment of semantic prosody. The present paper does not focus on what is traditionally subsumed under connotation or the speaker's attitude (e.g., swear words, pejorative and offensive language, praise, excuses, requests, demands, etc.), but on ways in which the pragmatic (functional) meaning that arises from various contextual features can become an integral part of lexicographic descriptions. This is important for the treatment of all of those lexical items whose meanings reside in their function rather than in their bare lexical-semantic meaning, as this is particularly the case with phraseology and idiomatics. From another perspective, pragmatics turns out to be an effective means of sense discrimination in works of lexical and lexicographic relevance, as will be shown in the continuation.

Keywords: lexicographic description; lexical database; pragmatics

1. Introduction

The Slovene Lexical Database (hereafter SLD)[1] is a monolingual lexical resource that will provide a corpus-driven account of the core vocabulary of the Slovene language, including semantic, syntactic, collocational and phraseological information, supported by illustrative examples. Data from the existing 620-million-word FidaPlus reference corpus of Slovene[2] is recorded and lexicographically treated using the Sketch Engine corpus query tool. The SLD is similar to other modern electronic databases[3] in that it is constructed on the principles of lexicogrammar, but it gives more overall importance to meaning descriptions.[4] Explanations are compiled drawing on those lexicographical practices that appear to be successful in terms of either the rigour of their lexicogrammatical approach or their user-friendliness, or both. Syntagmatics and meaning descriptions have been given more prominence, perhaps intentionally somewhat at the expense of the description of the inherent features and paradigmatic associations of words. [5] In principle, the main goal of a lexical database is not to produce (finalised) meaning descriptions, but to create lexical profiles of words by describing lexicographically relevant information (Atkins; Fillmore; Johnson, 2003); nonetheless, in the SLD, a great deal of thought has been given to the nature and form of lexicographic description as it should be provided in dictionaries, particularly those for upper elementary and intermediate school goers. This is partly due to the specific conditions in the Slovene language community – a small dictionary market and fairly limited human and financial resources in the field of lexicography. With these goals in view, in setting the guidelines for meaning descriptions we have considered corpus-based and pragmatically aware contemporary monolingual dictionaries:[6] MEDAL, LDOCE, and COBUILD are quoted below in order to provide comparison.

[1] Slovene Lexical Database (2008-2012): The database's operation is co-financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. It presently contains 2,500 entries (http://www.slovenscina.eu/Vsebine/En/Aktivnosti/LeksikalnaBaza.aspx).

[2] *www.fidaplus.net*

[3] The SLD is close in scope and methods to the recently compiled DANTE database: http://www.webdante.com/.

[4] *Meaning description* is used with reference mainly to the SLD, while *definitions* are referred to as products of particular (past or future) lexicographic traditions. Throughout the article, *explanation* is used analogously to *meaning description*, in its broadest possible sense.

[5] According to Atkins, Fillmore and Johnson (2003), a complete description of the lexicographically relevant information required for the proper analysis of a keyword would have to include the word's *inherent features* (part of speech class and subclass, semantic type, etc.), its *pragmatic features* (information about users and user communities, contexts of use, emotional affect, etc.), its *paradigmatic associations* (synonymy, antonymy, meronymy, etc.), and its *syntagmatic* or *combinatorial features* (information about the context a word creates or satisfies, expressed in terms of grammatical and semantic phrase types and lexical collocations).

[6] Namely, the Macmillan Dictionary and Thesaurus: Free English Dictionary Online (MEDAL 2010), the Longman Dictionary of Contemporary English (LDOCE 2003) and the Collins COBUILD English Language Dictionary (1995).

1.1 The conceptual framework

Semi-automated electronic databases, such as FrameNet[7] (Fillmore et al., 2003), Corpus Pattern Analysis (hereafter CPA)[8] (Hanks, 2004) and the Cobuild Project[9] (Sinclair, 1987) have all been studied in the construction of the SLD. FrameNet primarily builds ontologies and is concerned with the identification of semantic participants and argument structures by means of predetermined and largely formalised syntactic-semantic categories. It is therefore "more or less limited to recording information about the combinatorial requirements of the words it studies" (Atkins; Fillmore; Johnson, 2003). CPA both records the participant structure of a sentence pattern and provides a schematic explanation of the particular pattern (implicature) that establishes the relationships between the identified participants. The implicature does not pretend to be a dictionary definition, partly because it is ascribed to a particular pattern rather than to a conventional dictionary "sense". Cobuild explanations, on the other hand, are characterised by the clarity and naturalness of the definition language, achieved by describing the meaning of the headword in terms of its typical syntactic patterns and the immediate context surrounding it, and "[u]nlike classical definitions Cobuild definitions make their headwords an integral part of mentioning them, and so deal with the meanings of the words being defined both as entities and activities" (Barnbrook, 2002: 19-20). Metalinguistic information, which traditionally had no place in dictionary descriptions, is now foregrounded (COBUILD: 495): If you **explain** something, you give details about it or describe it so that it can be understood.

1.2 The theoretical background - the Sinclairian lexicographic tradition

The Cobuild definition style is perhaps the most literal transfer into practice of what Halliday (2007: 26) summarises as follows: "In general, it is unwise to assume that meaning is captured in dictionary entries, in the definitions or explanations given against the words. Dictionary definitions can and should be informative and helpful, and, when well written, they provide a paraphrase or explanation of meaning. But the meaning is not necessarily fully contained or exhaustively captured within such a definition. This is not to say that meanings are vague or ethereal. Within the conventions of a particular language, meanings contrast with each other with considerable precision. Words do not mean whatever we want them to mean, but are governed by social convention. Nonetheless, we cannot assume, without qualification, that the wording of a dictionary definition is an ideal representation of what a word means."

1.2.1 Meaning as event vs. meaning as entity

Within a wide range of reflection on the nature of meaning, there have been various attempts to define its complexities, motivated, among other things, by the need to explain language and the ways in which it is used to an average user. The theoretical framework for some of the modern views on pragmatics can be traced back to Piotrowski's (1989: 73-74) formulation: "Thus, on the one hand meaning can be seen as a sort of entity: concept, notion, prototype, stereotype, or fact of culture. On the other hand, meaning can be seen as a sort of activity: skill, knowledge of how to use a word." From this understanding, the so-called "use-mention" dichotomy was derived, built on extensively by Sinclair et al. in the Cobuild project. Hanks adds complexity to the "use-mention" pair by claiming that "[d]ictionaries are much concerned with accounting for what it is that an utterer may expect a hearer to believe" (1987: 20). In his tribute to elegance in lexicography, Rundell (2010: 357) points out how the instability of word senses has long been observed by thinkers about words, summing up Hanks's conclusions on the issue as "it makes more sense to think of meanings as events rather than (as their treatment in dictionaries implies) independently existing entities" (ibid.)

1.2.2 Use and meaning – the metalinguistic approach

The analysis of instances of natural text has long shown that some words are more literally "used" to produce a desired effect, such as to convey the intentions of the speaker, than others; or, as Sinclair (1991: 126) puts it, "[t]he statement may be about what people mean when they use a word or phrase, rather than what the word or phrase means." A well-known example employed by Sinclair to demonstrate how a restatement of meaning becomes a metalinguistic comment on the way the word is used in a context of situation is:
If you call a woman a bitch, you mean that she behaves in a very unpleasant way
vs.
derog. A woman, esp. when unkind or bad-tempered (Barnbrook, 2002: 178).

As Hanks (1987: 203) succinctly put it, "in the most common meaning of this word, what is at stake is the utterer's intention to insult, not the semantic convention associated with the sense. This meaning must be distinguished from the "literal" meaning, which although rather rare is privileged."

2. The Meaning Descriptions in the Slovene Lexical Database

2.1 General principles

The descriptions are formed as full-sentence definitions,[10] with a view to further lexicographical treatment for the purposes of general, and particularly

[7] http://framenet.icsi.berkeley.edu/.
[8] http://nlp.fi.muni.cz/projekty/cpa/.
[9] J. M. Sinclair, 1987 – see References.

[10] Pioneered by the COBUILD 1 project (1987).

student (upper elementary and intermediate school), monolingual audiences. The treatment of lexical data in the SLD sets out to describe individual lexical items, their meanings and usage by means of FrameNet-type "scenarios", which includes defining the range of semantic and syntactic combinatorial possibilities (valencies of each word in each of its senses) (Kocjančič; Zaranšek, 2009). The distinguishing features of the SLD descriptions are: the entry headword is integrated into the definition in its natural context; the syntactic environment of the headword and its semantic participants (semantic roles) are manually annotated: obligatory participants are in block letters for purposes of semi-automated pattern retrieval; to complete the picture, a description of broader circumstances of meaning is provided. Subordinate to the level of argument structure are the levels of grammatical patterns or structures, and collocations; the general rule is not to bend the argument structure too much towards either of these groups of lexical information, but rather to formulate it as a summary of all of them (ibid.).

The SLD meaning descriptions, then, are not yet ready-made dictionary definitions, but rather semantically and pragmatically informed "implicatures" which provide a platform for further work on explanations and definitions tailored to the needs of specific target groups. The policy of identifying the argument structure for each lexical unit[11] and annotating the semantic roles within what, at the same time, have to be adequate and comprehensible (and elegant at that) explanations has resulted in a relatively unique definition language. The aim was to bring together the best of what modern lexical descriptions based on, or influenced by, the contributions of computational lexicography, such as FrameNet, CPA and the Cobuild project, have had to offer. Combining formalisation with efforts to produce intelligible and simple meaning descriptions has resulted in an occasional clash of emphasis, which has had to be resolved independently for each situation and with regard to the reference skills of an average end-user. The fact that the database is intended essentially both for dictionary compilers – who will be compiling a student dictionary – as well as for general users who might be interested in querying somewhat raw linguistic data has added to the challenge.

Based in part on the Cobuild definition taxonomy – which was primarily designed to serve the purposes of the computer processing and formalised accordingly – and on an analysis of the early SLD entries, a new definition taxonomy was built taking into account the specifics of the Slovene language. The main purpose of the taxonomy was as much to provide guidance within the broad spectrum of defining possibilities as to homogenise the choices the compilers were making.

2.1.1 Conciseness and simplicity

The descriptions follow the maxim that the words used in them will be either more precise or easier to understand than the headword is by itself (Barnbrook, 2002: 49). This generally means that rare, polysemous or difficult words, as well as figurative expressions, are avoided in the descriptions, which essentially aim to tell what we already know about the meaning. The concept of "exactness" is secondary in importance to intelligibility, which subsumes brevity, conciseness and simplicity:

a HUMAN **breathes** by taking AIR into his/her lungs and pushing it out again[12]
ČLOVEK ali ŽIVAL **diha** tako, da potegne ZRAK v pljuča in ga nato spet potisne ven

Where there are indications in the corpus data that (pragmatic) circumstances contribute decisively to a particular sense, the headword may require a more extensive description:

an **argument** is a logically derived reason used in a debate to persuade the listeners or the opponents to support you
argument je logična utemeljitev stališča v razpravi, s katero skuša človek pridobiti naklonjenost poslušalcev ali prepričati nasprotnike

2.1.2 The syntactic-semantic description

Meaning descriptions in the SLD are schematically divided into two parts or levels:
a) The participant structure: all of the identified participants and circumstances are assigned semantic types or semantic roles. Syntactic and semantic information is overtly marked in order to enable automatic retrieval of patterns of usage. The assumption is that each meaning is realised within a syntactic pattern consisting of all of the words, expressions and situations in the co-text that contribute decisively to the meaning of a lexical unit.
b) "The scenario" is the level of description that states the general situation of meaning, the relationships between the participants and other sense-discriminating, particularly pragmatic, components of meaning (Gantar et al., 2009: 108).

Obligatory vs. optional elements

Each meaning description includes all of the participants and circumstances that are needed to construct a particular meaning. In the process of identifying obligatory elements, we have also dealt with cases of

[11] *Lexical unit* is used throughout as "a unit of meaning", unless otherwise indicated.

[12] In the paper all of the examples from the SLD are translated into English to aid understanding and listed first. Although sometimes awkward, the translations are intentionally as literal as possible, so that the organisation of the original descriptions remains evident.

null instantiation, but we will not further elaborate on their treatment here. Block letters are used to mark the obligatory participants, while the remaining information appears in lower case (see below) and is considered as part of the "scenario" (in italics). Participants are identified as obligatory if in at least some contexts they are syntactically (or contextually) expressed, i.e., their instantiations are to be found in the corpus data:

if a HUMAN **beseeches** *another* HUMAN, *s/he begs them to help him/her out of a* DIFFICULT SITUATION *or to do something that means a lot to him/her*
če ČLOVEK **roti** *drugega* ČLOVEKA, ga *obupano prosi, da* mu *pomaga iz* STISKE *ali stori nekaj, kar* mu *veliko pomeni*
• *"Please," he* ***besought*** *me, "give me a chance to meet my son ... "*
• *Even as a young woman she was* ***besought*** *by some not to confine herself to the convent, with all that energy, passion and glitter in her eyes.*
• *I burst into tears* ***beseeching*** *and begging her to be more understanding and gentle.*
• *He* ***besought*** *all his friends not to betray his secret to anyone.*

The semantic roles of obligatory participants are annotated in all of the descriptions of verbs as well as of those nouns and adjectives that are "verbal" in nature and therefore construct their meanings with analogous valency patterns:

a **reproach** is a critical expression of dissatisfaction or disappointment, usually in a quarrel, that a HUMAN has endured by another HUMAN
očitek je povzetek nezadovoljstva ali razočaranja, ki ga ČLOVEK izrazi nad ravnanjem drugega Č LOVEKA, navadno med prepiranjem
• *A common* ***reproach*** *to Anna was that she did not show enough interest in the learning skills of her child.*
• *Their children are growing up in a hostile environment, often saturated with mutual* ***reproaches*** *and conflicts.*
• *A severe source of conflict can be mentally or physically handicapped children, especially if the parents are full of* ***reproach*** *for each other.*

This is especially the case with the adjectival and nominal meanings typically activated in the predicative position:

if a HUMAN is **frivolous** s/he does not think enough about the CONSEQUENCES of his/her actions, or does not care about them
ČLOVEK je **lahkomiseln**, če ne razmišlja dovolj O POSLEDICAH svojih dejanj ali mu zanje ni mar

In some situations, however, participants, and especially circumstances, are typically expressed but are not decisive for the realisation of a particular meaning. Such elements are identified as optional and viewed as part of the "scenario". They are, in principle, introduced by the adverb "usually", i.e., a hedge, providing a wider context of situation:

if a HUMAN **flours** FOOD s/he sprinkles it with FLOUR, usually in the process of cooking
če ČLOVEK **pomoka** ŽIVILO, ga potrese z MOKO, navadno v postopku priprave jedi
• *They are then cut to pieces, which we* ***flour*** *with the rest of the flour and place in buttered cookie moulds.*
• *We* ***flour*** *them with buckwheat flour and add stock.*
• *The dough is then* ***floured****, covered and left in a warm place to rise.*

2.2 Metalanguage in full-sentence meaning descriptions

2.2.1 If/when-sentences

The if/when-sentence puts the description into a metalinguistic mode in which the natural usage of the headword is "encoded implicitly within the description text itself rather than explicitly as a separate, densely encoded abbreviation which the user may well ignore" (Barnbrook, 2002: 7-9). In principle, if/when-sentences make the description more explicit, thus facilitating the inclusion of pragmatic components of meaning. The if/when-sentence is also a typical definition type, particularly for verbs, in the SLD:

if a HUMAN or CIRCUMSTANCES **degrade** a HUMAN, his/her EFFORTS, or his/her KNOWLEDGE, they destroy his/her sense of value or importance, or diminish his/her role
če ČLOVEK ali OKOLIŠČINE **degradirajo** ČLOVEKA, njegove NAPORE ali ZNANJE, ga razvrednotijo, mu vzamejo veljavo ali zmanjšajo njegovo vlogo
COBUILD: Something that **degrades** someone causes people to have less respect for them. (*...the notion that pornography degrades women... //When I asked him if he had ever been to a prostitute he said he wouldn't degrade himself like that).*

There is enough flexibility in the guidelines to prevent a forced and inappropriate use of this type – if/when-sentences can be cumbersome and therefore inappropriate for some meanings – however, we abide by the rule on full-sentence definition. While acknowledging the advantages of the Cobuild strategy in describing words or phrases that typically occur in quite limited contexts, Rundell (2010: 361) is critical of its application where contextual features are not especially salient, as it "can sometimes lead to definitions which mislead the reader by overspecifying typical contexts of use (Rundell, 2006: 330-331)." On providing arguments in favour of full-sentence definition, Barnbrook (2002: 55) also acknowledges the fact that "[t]he adequacy of the contents of any individual dictionary is a separate consideration." These objections are indeed more relevant for dictionaries. If/when-sentences often create the need to use more pronouns and anaphoric expressions, which can be at the expense of elegance. Each solution in the SLD is subject to consideration from the perspective of the potential proliferation of anaphoric and deictic elements.

2.3 Pragmatics and the definition strategies

Various aspects of lexicographic description have been studied in the SLD, including the potential for rendering pragmatic components an integral part of meaning description, for which so-called "projection" proves to be extremely useful.

2.3.1 The "projection" principle

This description is characterised by reported speech or by the so-called "report" element of the co-text in the left side of the definition (Sinclair, 1991: 126-127), which re-establishes the traditional lexicographic equation as a comment on usage rather than as a description of the intrinsic meaning of the headword. The label "projection" was taken from Halliday (Barnbrook, 2002: 151-152). Hanks (1987: 204) links the projection principle – and the use of a strategy such as "If you say that ..." or "If you call someone a ... " – directly to the description of figurative senses and phraseology or idiomatic expressions (see section 2.4):

if we say that TREES and BUILDINGS **soar** we mean that they rise up very high
če rečemo, da DREVESA ali ZGRADBE **silijo** navzgor, menimo, da segajo zelo visoko

Some headwords and their meanings thus need to be treated with special attention to what, in actual fact, "people mean" when they use them in writing or speech. The SLD meaning descriptions attempt to make this sometimes very subtle layer of meaning as evident as possible within the explanation itself, rather than using labels, usage notes and so on to convey comments on usage. It is understood that this information is an inseparable part of meaning. An alternative strategy is to use a "something is an expression for something" formula which generally has been avoided:

a **consumer** is an expression used for someone who regularly buys and uses goods or services, especially with regard to his/her rights
potrošnik je izraz za posameznika, ki redno kupuje in uporablja trgovske izdelke in storitve, zlasti kadar so v ospredju njegove pravice
COBUILD: A **consumer** is a person who buys things or uses services. (*...claims that tobacco companies failed to warn consumers about the dangers of smoking.// ...improving public services and consumer rights.)*

To express specific circumstances of usage in this case, MEDAL, for example, uses the label Economics, the second part of the description specifying usage in relation to the grammatical feature "singular with plural meaning", which indicates that "the consumer" is viewed as a category of people (economics. "someone who buys and uses goods and services. The expression the consumer is often used for referring to consumers as a group"). To the definition "someone who buys and uses products and services", LDOCE adds a usage note: "A consumer is anyone who pays for goods and services. This word is used especially when you are talking about people's rights (*Consumers have a right to know what they are buying*)".

2.3.2 Circumstances of meaning – hedging

Pragmatic information is often located in the circumstances of meaning. In view of the fact that a high degree of granularity is presupposed in the SLD meaning descriptions, pragmatic components, along with the semantic-syntactic behaviour of words, become prominent indicators of meaning (nuances) and often play an important role in the process of sense discrimination. The above examples show that – seemingly for reasons of the structure of natural discourse – the pragmatic background will often fit naturally into the end part of the description. Early attempts to annotate pragmatic elements in the descriptions for purposes of automatic retrieval were abandoned due to seeming inconsistencies in the lexicographic treatment of pragmatic information. It turned out that this type of information is quite naturally and consistently located in the semi-formalised parts of descriptions beginning in "usually" (also "especially"), which provide typical situations of meaning. Where relevant, the definitions are given below from MEDAL and LDOCE in order to provide comparison of the ways in which they describe (pragmatic) circumstances:

if a HUMAN **distorts** INFORMATION, FACTS or someone's STATEMENT, s/he intentionally presents them in a way that is no longer accurate or true, usually because s/he wants to hide something or to harm someone[13]
če ČLOVEK **izkrivlja** PODATKE, DEJSTVA ali IZJAVE, jih namenoma prireja ali navaja neresnično stanje, navadno zato, ker hoče kaj prikriti ali komu škodovati
• *Such comments **distort** the truth in an insidious way and are covertly destructive.*
• *During the interrogation he presumably **distorted** the truth, gave false information, but above all he openly lied about how he'd dealt with the problem himself.*
• *A while ago a group of renowned Danish scientists accused him of **distorting** facts to make them fit with his own theories thus violating the research community ethic.*
MEDAL: to change something such as information so that it is no longer true or accurate (*The paper was accused of distorting the truth// The programme presented a distorted picture of her life.*)

[13] Examples of usage would be needed to adequately support the existing meaning descriptions throughout the text, but they could not, unfortunately, always be listed for reasons of limited space.

LDOCE: to report something in a way that is not completely true or correct (*His account was badly distorted by the press*)

The section introduced by "usually" typically contains information on cause, reason, intention, manner or other circumstances of meaning – this strategy is adopted analogously in many monolingual dictionaries, including COBUILD, MEDAL and LDOCE. In the SLD, the question often arises whether a circumstance is to be interpreted as typical or obligatory given that participants and circumstances must be explicitly identified as obligatory (set in block letters) or optional (in lower case). In order to be either one or the other, the element must recur in the corpus data; however, the evidence is not always unequivocal, especially as we move away from evident syntactic patterning to more elusive contextual clues. Sometimes components of meaning have to be inferred from the co-text, sometimes even based on our general knowledge of the world. Below are some examples of unclear situations, which are illustrated by the descriptions in the EFL dictionaries. Here we touch upon the issue of semantic prosody (more on this in section 2.5).

The example below shows that the "for pleasure" part is interpreted in the SLD as a non-obligatory participant, while MEDAL presents it as a "core" element of the definition, adding instead as an extension "especially to a series of places", which provides more detailed information on the type of experience in store:

a HUMAN **cruises** around the SEA or a RIVER by BOAT, usually for pleasure or as part of a holiday
ČLOVEK s PLOVILOM **križari** po MORJU ali REKI, navadno za zabavo ali preživljanje prostega časa
MEDAL: to sail in a ship for pleasure, especially to a series of places (*The first week was spent cruising around the Baltic// They'll be going cruising the Greek islands next week.*)
LDOCE: to sail along slowly, especially for pleasure (*We were cruising in the Caribbean all winter.// an evening spent cruising the River Seine*)
COBUILD: If you **cruise** a sea, river, or canal, you travel around it or along it on a cruise. (*She wants to cruise the canals of France in a barge.// During their summer holidays they cruised further afield to Normandy and Brittany. (*A **cruise** is a holiday during which you travel on a ship or boat and visit a number of places).

Another example of discrepancy follows where the "by force" element is recognised in the SLD and in LDOCE as an added circumstance, while in MEDAL and COBUILD this is a central component of the "scenario":

if an INDIVIDUAL or a group of PEOPLE **overthrows** a HUMAN or a REGIME they take away his power, usually by using force
če POSAMEZNIK ali SKUPINA LJUDI **strmoglavi** DI KTATORJA, PREDSEDNIKA ali VLADO, mu navadno z uporabo sil e odvzame moč
MEDAL: to force a leader or government out of their position of power (*Her father was overthrown in a military coup in the seventies.*)
LDOCE: to remove a leader or government from power, especially by force
COBUILD: When a government or leader **is overthrown**, they are removed from power by force. (*That government was overthrown in a military coup three years ago.// ...an attempt to overthrow the president).*

2.4 Phraseology – figurative meanings and idiomatic expressions

Many, but not all, figurative meanings in the SLD fall under the phraseology section. Figurativeness is neither necessary nor a sufficient, but merely a typical condition for listing a lexical unit under phraseology. Often, in order to describe (figurative) meaning and idiomatic expressions, there is a need to adopt what Hanks (1987: 203) has called a "displacement strategy" while referring to the Cobuild tradition. This is particularly important where a phraseological unit allows for a literal interpretation, which "must be guarded against" (ibid.). Hanks quotes the example: If you twist someone round your little finger, they will do anything that you want them to do. For Hanks, "that is open to the objection that it is either a false statement about English or a false statement about the world, or both." Certain notions go with a conventional interpretation of figurative meaning, and this must be "indicated on both the left-hand and the right-hand side of the explanatory equation (ibid.)": If you say that you can twist someone round your little finger, you mean that they will do anything that you want them to do.

Attempts have been made in the SLD to identify those figurative meanings that require displacement by using the strategy "If you say that ..." as opposed to those which presumably do not. Hard and fast rules in this matter cannot be applied, given that the interpretation of metaphors and figurative language is largely subjective. This is potentially a point where the intuitions of a lexicographer must be exploited, which, of course, does not guarantee the best of outcomes:

if we say that a MUSICAL INSTRUMENT **squeaks**, we find the sound coming from it unpleasant, such as if somebody cannot play properly
če rečemo, da GLASBILO **cvili**, se nam zdi zvok, ki prihaja iz njega, neprijeten, na primer, ker nekdo ne zna igrati

if someone says the **time flies** s/he has a feeling that time passes quickly[14]

[14] In the descriptions of phraseology, we have not explicated obligatory participants (not in block letters). Also, the description techniques have not been prescribed to the same

če kdo izreče, da **čas beži**, ima občutek, da čas hitro mineva
MEDAL: used for saying that time seems to be passing very quickly
LDOCE: used to say that time seems to pass very quickly (*Time flies when you're having fun*)
COBUILD: If you say that **time flies**, you mean that it seems to pass very quickly. (*Time flies when you're having fun.)*

The distinction is easier to grasp if displacement is paralleled with the actual literal usage (the second example is treated in the phraseology section (no block letters)):

if a HUMAN is **following** another HUMAN s/he is walking or driving closely behind to see where they are going
če ČLOVEK **sledi** KOMU, hodi ali se vozi za njim, da bi ugotovil, kam gre
vs.
if somebody says that s/he is not **following** somebody s/he wants to convey that s/he cannot understand what they are saying
če kdo reče komu, da mu ne **sledi**, sporoča, da ne razume, kaj želi povedati

if NATURE or CIRCUMSTANCES **conspire** against a HUMAN s/he has a feeling that they are operating in disaccord with his or her interests or wishes, usually in a critical moment
če se NARAVA ali OKOLIŠČINE **zarotijo** proti ČLOV EKU, ima ta občutek, da
delujejo v nasprotju z njegovimi interesi ali hotenji, nava dno v odločilnem trenutku

The following expression is highly idiomatic, but does not seem to require displacement:

if somebody is always **a step ahead of their time**, s/he has new ideas or does things long before other people do making them show disapproval, distrust or, rarely, enthusiasm
če je kdo vedno **korak pred časom**, pomeni, da je z dejanji ali v mislih pred svojimi sodobniki, kar izzove sodbo okolja, bodisi tako da zbuja nezaupanje, nejevoljo ali redkeje navdušenje
COBUILD: If someone is **ahead of their time** or **before their time**, they have new ideas a long time before other people start to think in the same way. (*He was indeed ahead of his time in employing women, ex-convicts, and the handicapped.//His only fundamental mistake, he insists, is that he was 20 years before his time*).
The criteria of (non-)compositionality considerably affect the way in which pragmatics comes into play in phraseology. It has been recognised that "units of meaning associated with metaphors – *metaphoremes* – must obligatorily have a pragmatic function" (Cameron and Deignan, 2006, in Philip, 2009). In corpus linguistics terminology, this pragmatic element is the semantic prosody (ibid.).

2.5 Semantic prosody

The analysis of concordance lines enables the retrieval of typical patterns of meaning not only in the immediate co-text but also in the wider co-text of the headword, where subtle, on the surface less obviously recurring, elements of meaning may be identified. According to Sinclair (1996a: 34),[15] a semantic prosody expresses attitudinal meaning and is on the pragmatic side of the semantics/pragmatics continuum: "it shows how the rest of the item is to be interpreted functionally." An ethereal, but perhaps even more frequently cited, definition of semantic prosody comes from Louw (1993: 157): "a consistent aura of meaning with which a form is imbued by its collocates." Semantic prosodies add meaning that goes beyond the meaning already expressed by word-semantics, requiring a close examination of contexts of use and components of meaning that are not always detectable in the immediate surroundings of the headword, or, as Philip (2009) puts it, "[c]orpus texts facilitate the retrieval of recurrent patterns, but they do so at the expense of the context of situation in which the language under study was originally uttered. Semantic prosodies, therefore, have to be inferred by extracting information from the cotext which allows a picture of the context of situation to be built up." The fact that semantic prosody is somewhat elusive and not always present has given some linguists reason to discard it as "a figment of corpus linguists' imaginations" (ibid.), similar to the scepticism that permeates pragmatic meaning in general. Nonetheless, corpus evidence shows that semantic prosody, like meaning on the whole, cannot be identified purely with introspection. As Louw explicitly states: "semantic prosodies are a collocational phenomenon and one which is preferably to be regarded as recoverable computationally from large language corpora rather than intuitively" (2000: 48).

From the above, and particularly from the practical analysis of corpus data, it emerges that semantic prosodies are often difficult to describe "clearly and succinctly, and this may well explain the widespread tendency to speak loosely of positive/negative prosodies rather than attempt to articulate the semantic prosody more precisely" (Philip, 2009). In view of this fact, the question arises as to how, and to what extent, to include semantic prosodies in the descriptions of meaning, as such inclusion will inevitably increase their length and complexity:

extent, thus allowing the compilers more individual freedom in deciding which strategy best suits a lexical unit.

[15] Semantic prosody was first used and presented to the research community by Sinclair (1996b) and Louw (1993).

if a HUMAN **equates** SOMETHING with a PHENOMENON, CONCEPT, or CHARACTERISTIC s/he thinks that they are the same things, usually failing to see the difference either as a result of ignorance or intentionally, due to prejudice
če ČLOVEK **enači** KAJ s POJAVOM, POJMOM ali LASTNOSTJO, meni, da gre za enake stvari, pri tem pa navadno spregleda bistvene razlike, bodisi zaradi nevednosti ali namerno, zaradi predsodkov
- *In the process of compiling a draft of the final document the Arabic countries renounced the demand to **equate** Zionism with racism.*
- *In Western countries Muslim faith is **equated** with terrorism and all the Arabian nations are treated as potential suspects who have to prove their innocence.*
- *Because she **equates** good sex with love she persists with her partner even if nothing but sex is good in the relationship.*

Pre-corpus lexicographic descriptions generally included little or no pragmatic information. Electronic text corpora have made a considerable difference in the selection of illustrative material which tends to show typical usage. In their definitions, most contemporary dictionaries as yet fail to convey the complexities of semantic prosodies (or avoid them), but typically imply them in the examples of (typical) usage. This strategy works well on the presumption that implicit information is lexicographically sufficient:

MEDAL: to consider something to be the same as something else (*These people seem to equate honesty with weakness.//Don't make the mistake of equating high test scores and intelligence*.)
LDOCE: to consider that two things are similar or connected (*Most people equate wealth with success*.)
COBUILD: If you **equate** one thing with another, or if you say that one thing **equates** with another, you believe that they are strongly connected

Each of the examples listed contains either colligational or collocational information about the semantic prosody," i.e., "seem to equate", "don't make the mistake", and "most people". The semantic prosody could be summarised as: "(people) give equal importance to things that are not the same because they cannot, or will not, see the difference." Collocationally, juxtapositions of "honesty" and "weakness", and "wealth" and "success" also contribute to the construction of meaning, based on the conventional associations of these word patterns. These, of course, are subtle indicators that can only be identified as such in the context of the whole situation, and, particularly, when analysed against a vast collection of data.

While semantic prosodies are often equated with the so-called "semantic preference", some studies (Philip, 2009) have shown that the term semantic prosody can be used loosely incorporating what, in effect, are two different levels of meaning analysis, of which the first is word-centric and the second delexical, functional, phraseological or contextual. By bringing the associations back into a real context of situation, the latter facet of semantic prosody is inextricably pragmatic in nature: *where*, *when*, *why* and *to whom* something means what it does (ibid.).

3. Conclusions

Semantic prosody builds along the semantics/pragmatics continuum, and, unlike "communication background" (Verschueren, 1999: 47), is not a pragmatic backdrop on which we could look for infinite implicit meanings; on the contrary, it is a result of empirically identifiable elements of the meaning structure, albeit on the furthest boundaries of a lexical unit of meaning. This has serious implications for the analysis of corpus data and for the selection of the default span of concordance lines, as well as, and not least, for the way lexicographic descriptions of meaning are constructed. Pragmatic information is an integral part of an (extended) unit of meaning, identifiable only by examining its repeated occurrences in corpus data. Although in the present paper the focus is on pragmatic components that can be abstracted from contextual features, in the SLD we have also addressed the questions of word connotation and emotive and attitudinal meaning that can be associated with words per se. Regardless of the theoretical stance, what is obvious in the process of lexical analysis is the difference in the difficulty of describing semantic prosody as opposed to allocating collocations or attributing semantic roles. In our view, in a lexical database it is vital to provide all of the information on the headword that is retrievable from the corpus data – semantic prosody may be difficult to describe lexicographically, but when present it is an integral part of the wording that cannot be severed from the co-text or context. The question remains how to integrate it into dictionary meaning descriptions for the benefit of their users.

The Slovene Lexical Database is still in the process of compilation. Apart from providing information on the inherent features, grammatical patterns and syntagmatics of words it has also explored the possibilities for constructing meaningful lexicographic descriptions that can serve as a basis for dictionaries targeted at younger audiences. Special consideration has been given to ways in which pragmatics, as a component of meaning, can become an integral part of the definition.

4. References

4.1 Dictionaries and Databases

Collins COBUILD English Language Dictionary. (Second Edition). Sinclair, J. (ed.) (1995) London: HarperCollins. (COBUILD)

The DANTE Database. Accessed at: http://www.webdante.com/

Longman Dictionary of Contemporary English, Longman: Accessed at: http://www.ldoceonline.com/. (LDOCE)

Macmillan Dictionary and Thesaurus: Free English Dictionary Online. Oxford: Macmillan Education. Accessed at: http://www.macmillandictionary.com/. (MEDAL)

4.2 Other Literature

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

Atkins, 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 Festscrift for Patrick Hanks*. Kampala: Menha Publishers, pp. 267-297.

Atkins, S., Fillmore, C. & Johnson, C.R. (2003). Lexicographic Relevance. Selecting information from corpus evidence. *International Journal of Lexicography*, 16(3), pp. 251-280.

Barnbrook, G. (2002). *Defining language: A local grammar of definition sentences. Studies in Corpus Linguistics*. Amsterdam: John Benjamins Publishing Company.

Cameron, L., Deignan, A. (2006). The emergence of metaphor in discourse. *Applied Linguistics* 27, pp. 671-690.

Fillmore, C., Johnson, C.R. & Petruck, R.L. (2003). Background to Framenet. *International Journal of Lexicography,* 16(3), pp. 235-250.

Gantar, P., Krek, S. (2009). Dictionary Definitions - A Different View: to Describe, to Explain or to Paraphrase? (Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti?). In M. Stabej (ed.), *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Znanstvena založba FF UL in Center za slovenščino kot drugi/tuji jezik, pp. 151-159.

Gantar, P., Grabnar, K., Kocjančič, P., Krek, S., Pobirk, O., Rejc, R., Šorli, M., Šuster, S. & Zaranšek, P. (2009). *Standard for the compilation of a lexical unit (Indicator 6): Communication in Slovene Project: Style Guide: the Slovene Lexical Database*. Kamnik. Slovenia. Available at: http://www.slovenscina.eu/Media/Kazalniki/Kazalnik6/SSJ_Kazalnik_6_Specifikacije-leksikalna-baza_v1.pdf.

Halliday, M.A.K., Yallop, C. (2007). Lexicology. A Short Introduction. London & New York: Continuum.

Hanks, P. (1987). Definitions and explanations. In J.M. Sinclair (ed.) *Looking up. An Account of the COBUILD Project*. London and Glasgow: Collins ELT.

Hanks, P. (1994). Linguistics Norms and Pragmatic Exploitations, or Why Lexicographers Need Prototype Theory, and Vice Versa. In: F. Kiefer, G. Kiss & J. Pajsz (eds.) *Papers in Computational Lexicography: Complex '94*. Budapest: Hungarian Academy of Sciences, pp. 89-113.

Hanks, P. (2004). Corpus Pattern Analysis. In G. Williams, S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress*, *Volume I*. Lorient, Université de Bretagne Sud, pp. 87-97.

Kocjančič, P., Zaranšek P. (2009). Slovene Lexical Database: the organizing principles of the argument structure. In P. Cantos Gómez, A. Sánchez (eds.) *A Survey on Corpus-based Research*. Murcia: AELINCO, Asociación Española de Lingüística de Corpus, pp. 293-306.

Louw, B. (2000). Contextual Prosodic Theory: bringing semantic prosodies to life. In C. Heffer, H. Sauntson (eds.) *Words in Context: A tribute to John Sinclair on his retirement*. Birmingham: University of Birmingham, pp. 48-94.

Louw, W. E. (1993). Irony in the Text or Insincerity in the Writer?: The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis & E. Tognini Bonelli (eds.) *Text and technology: in honour of John Sinclair*. Amsterdam: Benjamins, pp. 157-176.

Philip, G. (2009). Why prosodies aren't always: Insights into the idiom principle. *Corpus Linguistics Conference. Liverpool. Great Britain.* http://ucrel.lancs.ac.uk/publications/cl2009/ (Accessed 17 October 2011).

Piotrowski, T. (1989). Monolingual and Bilingual Dictionaries, Fundamental Differences, In M.L. Tickoo (ed.) *Learners' Dictionaries: State of the Art*. Singapore: SEAMEO Regional Language Centre, pp. 72-83.

Rundell, M. (2006). More than One Way to Skin a Cat: Why Full-Sentence Definitions have not been Universally Adopted. In E. Corino, E, C. Marello & C. Onesti (eds.) *Proceedings XII Euralex International Congress. Torino, Italy, September 6-9 2006.* Alessandria: Edizoni dell'Orso, pp. 323-338.

Rundell, M. (2010). Defining Elegance. In G.-M. de Schryver (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festscrift for Patrick Hanks*. Kampala: Menha Publishers, pp. 349-375.

Sinclair, J.M. (1987). *Looking Up: Account of the Cobuild Project in Lexical Computing (Collins Cobuild dictionaries)*. London and Glasgow: Collins ELT.

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

Sinclair, J.M. (1996a). The search for a unit of meaning. TEXTUS IX.

Sinclair, J.M. (1996b). The Empty Lexicon. *International Journal of Corpus Linguistics 1*, pp. 99-119.

Tickoo, M.L. (ed.) (1989). *Learners' Dictonaries: State of the Art*. Singapore: SEAMEO Regional Language Centre.

Verschueren, J. (2000). *Razumeti pragmatiko*. (The original: *Understanding Pragmatics*, 1999). Ljubljana: Založba/*cf.

Evaluating e-resources for Japanese language learning

Irena Srdanović
University of Ljubljana
Aškerčeva 2, Ljubljana 1000, Slovenia
E-mail: irena.srdanovic@ff.uni-lj.si

Abstract

The aim of this paper is to evaluate and compare electronic resources for the Japanese language in terms of their usability for Japanese language learners. The paper focuses on a computer assisted language learning (CALL) system developed to support learners in reading Japanese language texts, the Japanese Language Reading Tutorial System “Reading Tutor”, and the e-dictionary server “WWWDJDIC” with its extended functions. Each system was introduced and used in the classroom for various teacher-planned reading and comprehension tasks. The paper outlines the advantages and disadvantages of the systems that were identified during classroom usage by university students enrolled on a Japanese language course, and discusses the results of a survey conducted in order to explore student opinion, preferences and usage habits regarding the resources. In addition, the paper also notes and briefly addresses some other e-dictionaries, e-resources and tools covered in the survey, such as the pop-up support dictionary tools “Rikai-chan”, “Perapera-kun”, and “Popjisyo”, the Japanese Reading System for Multi-Lingual Environments “Asunaro”, and the Writing Support System “Natsume”, as well as the corpus query system “Sketch Engine” with “Word Sketches”, which automatically extracts collocational information for the Japanese language.

Keywords: e-resources; CALL systems; Japanese language learning; evaluation; reading and comprehension tasks; student survey

1. Introduction

The Japanese language is peculiar for its writing system as it uses three sets of characters: two syllabic, *hiragana* and *katakana*, which are relatively easy to master, and a few thousand Chinese characters, *kanji*, which require a long learning process. The Japanese Ministry of Education prescribes 2,136 characters as being the most essential for common use and everyday communication.[1]

Chinese characters in Japanese writing are used either independently, or in combination with *kana* letters, or in various combinations with other Chinese characters, and typically they have two or more readings. It is a frequent situation for learners to not know how to read an unknown Japanese word and, thus, must first consult a Chinese character dictionary, where characters are arranged according to form, radicals, and the number of strokes, prior to looking-up the meaning in a dictionary arranged according to either an alphabetical or *kana* order. This is a time-consuming process and an obstacle to Japanese language students when trying to read Japanese texts, which is especially the case with using traditional paper dictionaries.

Along with the advances in computer and internet technologies, various computer assisted language learning (CALL) systems, online tools, and dictionary servers have been created to support the reading of electronic Japanese texts. They provide significant support to Japanese language learners through various functions: ranging from simple or complex search methods based on Chinese radicals, characters and their combinations, to quick information about the readings and meanings of unknown words in the form of pop-ups, lists of entries, sentence analysis, text analysis. Naturally, the quality of such systems is dependent on the electronic dictionaries and the tools of morphological analysis incorporated within the systems. Moreover, the overall usability of such systems greatly depends on the user interface, customization options, and the possibilities of interacting with other language resources, such as corpora and lexical databases, including various learning aids and materials, as well as other incorporated functions.

This paper focuses on an evaluation of e-resources that support reading skills; the Japanese Language Reading Tutorial System “Reading Tutor” and an e-dictionary server with extended functions “WWWDJDIC”. The paper also notes some other tools relevant for Japanese language learners.

2. Evaluation targets and methods

The evaluation of CALL systems is a challenging task and requires the involvement of both teachers and students.

Chapelle (2001) introduces a number of principles for the evaluation of CALL systems. One of the principles is the need to provide judgmental and empirical analysis within an evaluation. According to Chapelle, there are three levels of analysis and three objects of evaluation: CALL software, teacher-planned CALL activities and learner performance during the CALL activities.

Hubbard (2006) describes the evaluation of CALL software as a three-stage process consisting of (a)

[1] Details on the current *jōyō kanji* list (常用漢字, Chinese characters in regular use) are provided at the web site of the Japanese Ministry of Education: http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2010/10/08/1298254_02.pdf

selection: investigating a piece of CALL software to judge its appropriateness for a given language learning setting, (b) implementation: identifying ways it may be effectively implemented in that setting, and (c) assessment: assessing its degree of success and determining whether to continue using or to make adjustments to the implementation for future use.

This paper only briefly covers the selection and the implementation of the resources in order to concentrate on the third process of assessment. The targets of the evaluation are e-resources for support in reading Japanese texts. These e-resources were demonstrated during a Japanese text processing course organized for fourth grade students of Japanese studies at the University of Ljubljana. The students' knowledge of Japanese is at the intermediate to advanced levels.

During the class, a number of teacher-planned e-resource activities were organized. Students used the systems to undertake various reading and comprehension tasks and they were asked to provide some feedback on the difficulties that they had encountered or to describe advantages of the used tools. Thus, the systems were evaluated based on the student feedback and issues identified as being either inappropriate or absent from the content, as well as the coverage of the systems, in adopting a judgmental approach to the evaluation.

In addition, quantitative comparisons were carried out on the results for content coverage and for the accuracy of the lexical information within the systems, to also provide an empirical basis to the evaluation. Although an overall evaluation of the e-resources is not conducted, the paper focuses on aspects deemed to be important for performing the teacher-planned reading and comprehension tasks.

Finally, a survey was conducted to gather information concerning student usage preferences, usage habits and their overall opinions and judgment of the target e-resources and other e-resources for Japanese language learners.

3. Japanese CALL systems and other e-resources supporting reading of Japanese

CALL systems are typically created with the objective of assisting one or more of the four foreign language learning skills: reading, writing, listening or speaking. The targets of this paper are CALL systems and other e-resources that have been created to support the reading of online texts in Japanese.

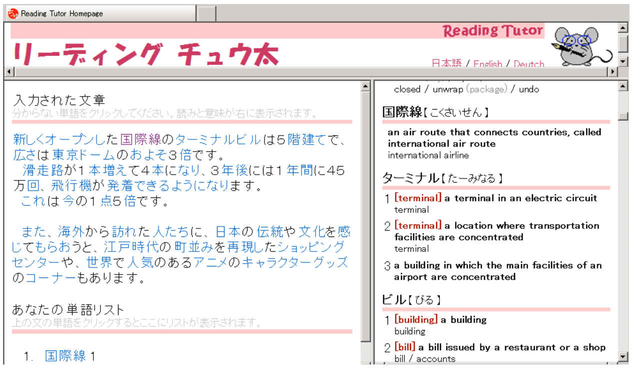
3.1 Reading Tutor

The Japanese Language Reading Tutorial System "Reading Tutor"[2] is a CALL system created to help Japanese language learners to start with reading in Japanese and to improve their reading skills (Kawamura, 2001). It was created at Tokyo International University and is freely available on the internet. The tool consists of four main functions:

- Japanese-English Dictionary Tool, where all words have an explanation in English, their English equivalents, and furigana (indicating the reading of the Japanese words).
- Japanese-Japanese Dictionary Tool, where all words have an explanation in Japanese, their English equivalents, and furigana.
- Vocabulary Level Checker, which analyzes all the words in a text according to their level of difficulty. These levels correspond to the four levels of the Japanese Language Proficiency Test.
- Kanji Level Checker, which analyzes the kanji (Chinese characters) that appear in the text according to their level on the Japanese Language Proficiency Test.

The Japanese-English and Japanese-Japanese Dictionary Tools operate so that users copy and paste any Japanese text into the toolbox, and then a user-friendly interface is provided with the original text on the left side and the lexical items with their descriptions, readings, translations on the right side. When clicking a word in the text, the word appears on the right side with all its lexical information (Figure 1).

In the background, the tools use the Japanese-Japanese and Japanese-English EDR dictionary and morphologically analyse texts into segments using the ChaSen morphological analyzer. In addition to the language pairing of Japanese-English, there are also pairings of Japanese-German, Japanese-Dutch, and Japanese-Slovene.[3] Each pairing employs the appropriate dictionary for the respective languages. In addition, the site provides a collection of reading materials and quizzes.



入力された文章
分からない単語をクリックしてください。読みと意味が右に表示されます。

新しくオープンした国際線のターミナルビルは５階建てで、広さは東京ドームのおよそ３倍です。
　滑走路が１本増えて４本になり、３年後には１年間に45万回、飛行機が発着できるようになります。
　これは今の１点５倍です。

　また、海外から訪れた人たちに、日本の伝統や文化を感じてもらおうと、江戸時代の町並みを再現したショッピングセンターや、世界で人気のあるアニメのキャラクターグッズのコーナーもあります。

あなたの単語リスト
上の文の単語をクリックするとここにリストが表示されます。

1. 国際線 1

closed / unwrap (package) / undo

国際線【こくさいせん】
an air route that connects countries, called international air route
international airline

ターミナル【たーみなる】
1 [terminal] a terminal in an electric circuit
terminal
2 [terminal] a location where transportation facilities are concentrated
terminal
3 a building in which the main facilities of an airport are concentrated

ビル【びる】
1 [building] a building
building
2 [bill] a bill issued by a restaurant or a shop
bill / accounts

Figure 1: The Reading Tutor user interface

[2] http://language.tiu.ac.jp/

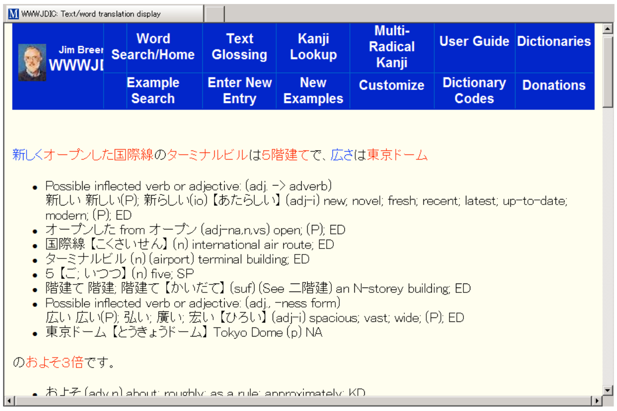
[3] The Japanese-Slovene Dictionary Tool is also available from http://nl.ijs.si/jaslo/chuta/ (Hmeljak-Sangawa & Erjavec, 2010).

3.2 WWWJDIC

The online Japanese dictionary service WWWJDIC has been developed by Jim Breen and others at the Electronic Dictionary Research and Development Group at Monash University. The dictionary server is associated with the JMdict/EDICT and KANJIDIC projects.[4] It offers various functions to Japanese language learners, such as:

- Word Search with lexical entries including furigana, translation, examples, part of speech annotation and pronunciation. It also offers various links to other e-resources: example sentences from the Tanaka corpus, verb conjugations, Japanese WordNet, Japanese Wikipedia, Google search and Google images, goo and ALC dictionary servers, and lessons from JapanesePod101.com
- Text Glossing, which is function most similar to the Reading Tutor, and supports learners in reading Japanese language texts. The main difference is in the display, as WWWJDIC divides a text into sentences and then provides the lexical entries for each word in a sentence including translations and other information (**Figure 2**).
- Kanji Lookup with animation of kanji stroke order and links to various Chinese character dictionaries.
- Multi-Radical Kanji to search for a kanji based on its components which are classified into a set of 250 basic shapes.

In addition, the server provides example search, word lookup in various bilingual dictionary combinations with Japanese language specialized dictionaries, customization options for its user interface, the ability to add new entries or examples, and to search using romanized Japanese. The server has six mirror sites and provides online help and up-to-date information about changes on the server. It can be also accessed on Japanese mobile phones. The main dictionary within the server is EDICT, which has been constantly improved over the last twenty years, and which has been integrated into a number of other e-resources that are being used recently, as Rikai-chan and Perapera-kun (see 0).



新しくオープンした国際線のターミナルビルは5階建てで、広さは東京ドーム

- Possible inflected verb or adjective: (adj. -> adverb)
  新しい 新しい(P); 新らしい(io) 【あたらしい】 (adj-i) new; novel; fresh; recent; latest; up-to-date; modern; (P); ED
- オープンした from オープン (adj-na,n,vs) open; (P); ED
- 国際線 【こくさいせん】 (n) international air route; ED
- ターミナルビル (n) (airport) terminal building; ED
- 5 【ご; いつつ】 (n) five; SP
- 階建て 階建; 階建て 【かいだて】 (suf) (See 二階建) an N-storey building; ED
- Possible inflected verb or adjective: (adj., -ness form)
  広い 広い(P); 弘い; 廣い; 宏い 【ひろい】 (adj-i) spacious; vast; wide; (P); ED
- 東京ドーム 【とうきょうドーム】 Tokyo Dome (p) NA

のおよそ3倍です。

- およそ (adv,n) about; roughly; as a rule; approximately; KD

Figure 2: The WWWJDIC user interface

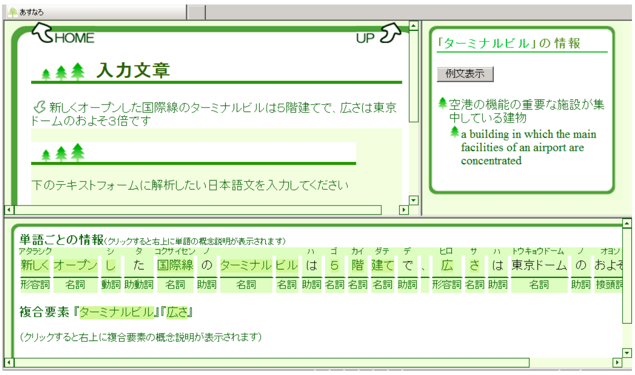
3.3 Asunaro

The Japanese Reading System for Multi-Lingual Environments "Asunaro"[5] is a reading-support tool for a range of languages: English, Chinese, Thai, Malaysian and Indonesian. The system has been created at the Tokyo Institute of Technology and is also freely available on the internet (Nishina et al., 2004). It has the following components:

- Morphological analysis of sentences with information about part of speech categories and meanings.
- Syntactic analyses of sentences using various methods (i.e., tree diagrams, box structures, and dependency structures).
- Example-sentence display. Sentences can also be morphologically and syntactically analyzed and the meaning of each lexical item is provided.
- Compound word elements with their idiomatic meanings.

The tool operates so that a user type in or copy in a sentence and the system provides the requested analyzes (Figure 3). In the background, this system also uses the Japanese-English EDR dictionary and the morphological analyzer ChaSen. Furthermore, it uses XPath to extract compounds, idiomatic phrases, and proverbs. The tool CaboCha is used for the dependency-structure analysis.

Although Asunaro is a reading support system, it was not evaluated within the present study. Because of its sentence-oriented user interface, it seemed less appropriate for undertaking the reading tasks planned during the course.



新しくオープンした国際線のターミナルビルは5階建てで、広さは東京ドームのおよそ3倍です

下のテキストフォームに解析したい日本語文を入力してください

空港の機能の重要な施設が集中している建物
a building in which the main facilities of an airport are concentrated

単語ごとの情報(クリックすると右上に単語の概念説明が表示されます)

アタラシク		シ	タ	コクサイセン	ノ			ハ	ゴ	カイ	ダテ	デ		ヒロ	サ	ハ	トウキョウドーム	ノ	オヨソ
新しく	オープン	し	た	国際線	の	ターミナル	ビル	は	5	階	建て	で	、	広	さ	は	東京ドーム	の	およ
形容詞	名詞	動詞	助動詞	名詞	助詞	名詞	名詞	助詞	名詞	名詞	名詞	助詞		形容詞	名詞	助詞	名詞	助詞	接頭詞

複合要素『ターミナルビル』『広さ』

(クリックすると右上に複合要素の概念説明が表示されます)

Figure 3: The Asunaro user interface

3.4 Other e-resources

Although a number of other CALL systems exist, such as e-dictionaries, corpus-query systems, and e-learning tools, which were noted during the course and within the survey, they are not evaluated in the present paper. This is because either they are not created specifically for reading-skill support or they were not planned for within

[4] http://wwwjdic.com/

[5] http://hinoki.ryu.titech.ac.jp/asunaro/index-j.php

the present course curriculum, and, thus, were not thoroughly presented and used within the classes. Such e-resources include:

- Natsume Writing Support System, a CALL system that supports learners' writing skills in Japanese. (Abekawa et al., 2011).[6] The tool provides frequent and statistically relevant collocational relations across various types of corpora.
- Sketch Engine, a corpus-query system where the main function is the presentation of word sketches; a summary of the collocational and grammatical relations of a particular word (Kilgarriff et al., 2004; Srdanović et al., 2008; Kilgarriff et al., 2010).
- Tools that provide online dictionary lookup in the form of pop-ups on a particular web-page, such as PopJisyo,[7] Rikai-chan,[8] and Perapera-kun.[9]
- Sanseido dictionaries, Yahoo dictionaries, Wikipedia, Google translator.

4. Teacher-planned task and tools evaluation

The course curriculum included demonstrations of and the utilization of the Reading Tutor, Asunaro, WWWJDIC and Natsume e-resources during the classes. The students had to perform a number of tasks to become familiar with the tools and their support function for reading and writing in Japanese. The teachers selected these tools to be part of the curriculum because they have been used for a number of years and have proven to be useful and supportive to Japanese language learners.

This section of the paper presents the reading and comprehension tasks that were mainly carried out using the two reading support resources: Reading Tutor and WWWJDIC. Specifically, it presents the issues highlighted by the students, and compares how successful the resources are in terms of content coverage to support the students in undertaking the tasks.

4.1 Reading and understanding news articles online

One of the tasks was to choose online newspaper article(s) in Japanese (between 3,500-4,500 characters in length), read them using one or more e-resources and write a resume of the content in Slovene. The resume part of the task was used to verify if students adequately comprehend the respective texts.

The students reported that they were satisfied with the support provided by the e-resources. However, the students also expressed various preferences about the use of these e-resources, from various perspectives including their content and coverage, the user interfaces, the ease-of-use, and whether or not internet connection was necessary. The students reported on a number of instances where the tools were unable to assist: including compound nouns, compound particles, names of books, people, places, colloquial terms, specialized terms, and words written in katakana. In such cases, the students frequently has to resort to searching for help in other resources; such as with dictionaries or e-dictionaries with more detailed information on the lexical entries, their usages and meanings, as well as Google search and Wikipedia for explanations of some terms, toponyms and proper names.

Overall, the students performed the tasks successfully with only minor reading and comprehension errors for the read materials.

4.2 Quantifying the content coverage of the resources

To quantify and compare the content coverage of the resources, the output results of the target resources for one newspaper article of about 1,400 words,[10] were examined in detail, by counting incomplete, misinterpreted and missing analyses.[11] Case particles, such as *ga, wo, kara*, and the topic particle *wa*, were excluded from the list, because the target resources do not actually provide descriptions for them. The auxiliary verb *da* and its forms are also excluded because it is not listed and covered in the dictionaries.

A summary of the results is presented in Table 1, which indicates some of the weak points of both resources in terms of their content coverage and accuracy. In total, 25 issues were found for WWWJDIC and 56 issues for Reading Tutor.

The weakest areas for WWWJDIC are various types of words (such as verbs, particles, conjunctions, adverbs, and nouns) written in hiragana, such as かかる *kakaru* 'to take', まわりの *mawari no* 'around'. Such words are not found in the dictionary. Also, the resource usually only partially covers compound particles, such as なければならなくなりました *nakereba naranaku narimashita* 'it became so that we should (do)' , and their meanings are not clear from the results.

The results for Reading Tutor clearly reflect ChaSen's fine-grained approach to morphological analysis, as the tool is incorporated within the CALL system and is used to analyze texts. Accordingly, there are numerous cases when compounds were divided into segments and the

[6] http://wombat.ryu.titech.ac.jp
[7] http://www.popjisyo.com/
[8] https://addons.mozilla.org/ja/firefox/addon/rikaichan/
[9] http://perapera.wordpress.com/perapera-kun/

[10] http://cgi2.nhk.or.jp/kdns/mwakari/mwakari_detail.cgi?id=244&y=2010&s=asc&w=%89H%93c (visited 27.01.2011.)
[11] Refer to Srdanović (2011) for information on coverage of collocational relations in these and some other computer-assisted language learning systems and resources for the Japanese language.

Reading Tutor provided translations for each segment separately, such as 成田 *Haneda* ‘Haneda [toponym]’ and 空港 *kuukou* ‘airport’, for 成田空港 ‘Haneda airport’, or 観光 *kankou* ‘tourism’ and 客 *kyaku* ‘guest’ for 観光客 *kankoukyaku* ‘tourists’. In most cases, it is not difficult to understand the whole compound from its segments, but, naturally, there were also exceptions. This issue becomes even less user-friendly when one part of the compound had already appeared within the text and is therefore separated by a number of entries from the other part of the compound in the entry definitions view of Reading Tutor.

	Type of issue	Example	No of issues
WWWJDIC (25)	partial match (compounds, phrases)	ことになる，とする，なければならなくなりました, 人気のある	6
	wrong match	チャンギ, としている	2
	no match (words written in hiragana)	ほしい，いったい，できる，１つ, ことし, もらう, かかる,まわりの	17
Reading Tutor (56)	partial match (compounds, phrases)	江戸時代，玄関口，～階建て, 本格的な	22
	wrong match and translation	ハブ, ローコスト	2
	no match	ではないか,だけでなく,万,羽田,チャンギ,ある	17
	matched but without translation	およそ,直接,お隣,チェックインカウンター,1 つ, 東京ドーム, ～倍	15

Table 1: Classification of the issues identified for WWWJDIC and Reading Tutor

In some cases, it seems likely that this kind of incomplete information elicits student mistakes when reading. For example, Reading Tutor presents the compound 玄関口 as 玄関 *genkan* ‘entrance’ and 口 *guchi / kou / kuchi* ‘gate/mouth etc.’ Based on this information, students can incorrectly read the compound as *genkankou* or *genkankuchi* instead of the correct *genkanguchi*.

There were also words (such as particles, compound particles, conjunction, verbs written in hiragana, and toponyms) that were not found in Reading Tutor, such as だけでなく *dakedenaku* ‘not only that’, 万 *man* ‘ten thousand’. Examples of words that are linked to Reading Tutor’s dictionary but did not have definitions are 東京ドーム *Toukyou Doumu* ‘Tokyo Dome’, チェックインカウンター *chekkuinkauntaa* ‘check-in counter’, and およそ *oyoso* ‘about’.

In addition, it should be noted that neither of the resources covers the frequent auxiliary verb いる *iru* ‘to be’ and its conjugations well. In the case of WWWJDIC, the form is not linked and is described as a part of a main verb; there is no definition or description provided for the verb. In the case of Reading Tutor, the auxiliary verb is not marked as such but it always provides the following misleading definition/translation: “to do something with deep commitment; heartily / [respect] intensely / profoundly / deeply”. The other frequent auxiliary verb of ある *aru* ‘to be’ is well covered in WWWJDIC but not in Reading Tutor.

Given that there were a little more than twice the number of issues for Reading Tutor than for WWWJDIC, we may conclude that WWWJDIC performs better in terms of the accuracy and completeness of its lexicon entries.

5. Students survey

At the end of the two-semester course, a survey was conducted that asked the students about their preferences and habits concerning e-resource usage, and about their opinions of the two target e-resources, as well as other resources that they are fond of using. Nine respondents completed the six-page long survey.

The students all responded positively to the general question of whether they like to use e-resources as a support for learning Japanese. Each student listed the one or more e-resources that he/she most typically and most frequently use. The lists differed from student to student. Table 2 provides a summary of student responses, which indicates that the Reading Tutor followed by WWWJDIC were listed as most frequently used by the students. This might also be due to the fact that these were given the most attention during the class, compared to Natsume (and Rikai-chan, which was suggested as an add-on for Natsume).

The next group of resources, which were listed by three students, are Rikaichan, Natsume and Perapera-kun. It is rather surprising that Perapera-kun is so high on the list as it was not demonstrated or used in the class. It seems that it is preferred by some students and therefore it will be considered for inclusion within future evaluations. Clearly, the students also typically use online services, such as Jisho.org, Yahoo dictionaries and alc.co.jp with multiple monolingual and bilingual Japanese dictionaries.[12]

Table 3 shows the e-resources that were chosen by students as the most preferred support for the reading task in Japanese. Reading Tutor is the most preferred, with students stating that it is very systematic, clear, and easy-to-use. Also, it is regarded as being useful because

[12] The fact that the Asunaro reading support system was not listed probably reflects its limitation in only allowing searches for one sentence at a time, which indicates that it is not sufficiently effective for doing the planned reading tasks.

it provides various senses for a word, although it does not provide appropriate entries for all words (such as proper names). The next choices are Rikai-chan and Perapera-kun, both providing pop-up dictionary support, which they are praised for: [13]

- (for Rikai-chan) “Its advantage is that you can read and understand any word while reading a text, all you have to do is to place your mouse on the unknown word.”, “It supports more languages.”
- (for Perapera-kun) “It is useful since it immediately translates words while reading text.”, “The entries are detailed, with good definitions and the grammatical properties of words.”

Most often used e-resources	No of students
Reading Tutor	6
WWWJDIC	5
Rikai-chan	3
Natsume	3
Perapera-kun	3
Jisho.org dictionaries	2
Yahoo dictionaries	2
Popjisho	1
Google translator	1
alc.co.jp dictionaries	1

Table 2: E-resources listed by students as being the most frequently used; some students listed more than one resource

Students also stated that one advantage of the tools is that they are very quick. Also, some minor disadvantages were noted: it is not possible to use it without an internet connection, one cannot use it for reading pdf files (for Rikai-chan), it doesn’t always properly cover some compound words, the definitions exist only in English (for Perapera-kun).

One of the students chose Popjisho, also a pop-up dictionary support, as most preferred.

Most preferred e-resources for reading	No of students
Reading Tutor	4
Rikai-chan	3
Perapera-kun	3
Popjisho	1

Table 3: E-resources listed by students as being the most preferred resource for reading in Japanese

Table 4 presents the e-resources that were chosen by students as being their most preferred support tool for writing in Japanese. Almost half of the students stated that they do not use e-resources for writing support. Three of the students indicated Natsume as being their most preferred, although some of them also wrote that they use additional resources (such as Perapera-kun, Rikai-chan and Google) in order to read and understand word combinations provided by Natsume. Natsume was also praised for its exhaustive word-usage information, but it was also criticized for its lack of clarity. Interestingly, WWWJDIC was not listed as a most preferred tool for reading but was for writing. It was described as being useful for writing because of its rich dictionary content, with examples and inflectional forms.

Most preferred e-resources for writing	No of students
(not using e-resources for writing)	4
Natsume	3
WWWJDIC	2
Perapera-kun (along with Natsume to check translations)	1
Google translator	1
Rikai-chan (along with Natsume)	1
Google, Wikipedia	1

Table 4: E-resources listed by students as being their most preferred resource for writing in Japanese

Furthermore, the students were asked to rate some of the e-resources on a 5-point scale in terms of the following four categories; a) usable for learning Japanese, b) easy and practical user interface, c) coverage and accuracy, and d) additional functions/data.

Figure 4 presents the results of the survey for the Reading Tutor and WWWJDIC resources.

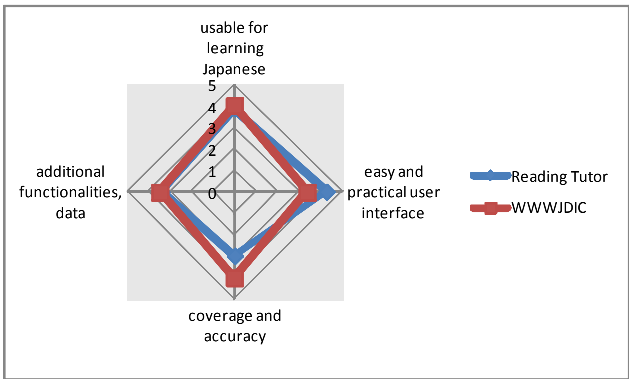


Figure 4: Rating survey results for Reading Tutor and WWWJDIC

The results are high for all four categories indicating the students’ level of satisfaction with the resources. While the students rated both resources similarly in terms of usability for learning Japanese and additional functions, there are also some noticeable differences in the rating for the easy-to-use/practical user interface and the coverage and accuracy of the resources. Reading Tutor

[13] The author of the paper translated students’ survey responses from Slovene to English.

was evaluated as being better in terms of its user interface - easy-to-use and practical. On the other hand, WWWJDIC was rated higher for its dictionary coverage and accuracy.

The pattern of ratings is in line with student comments about the advantages and disadvantages of these resources, as summarized in Table 5 and Table 6.

As these tables indicate, the majority of positive comments (C1– C7) for Reading Tutor refer to its user interface and its easy-of-use. Only C8 refers to the content of the dictionary. On the other side, only two negative comments (C1, C2) refer to the interface, technical aspects of the tools, and the rest (C3 – C7) refer to the lack of data in the resource's dictionary.

Reading Tutor	
Positive comments	**Negative comments**
(C1) Easy to use, systematic	(C1) Difficult to read if text is too long
(C2) Clear interface, easy to search for words	(C2) Need to click on the words
(C3) While reading, it is easy to search for unknown words and their readings.	(C3) Some words not found
(C4) We can copy text into it (also pdf data that is not readable by Rikai-chan, for example).	(C4) Some translations are not correct, especially for proper names
(C5) Easy to use	(C5) Rather poor dictionary, a lot of words (especially colloquial) not found
(C6) Easy to use, no need to move from one window to another	(C6) Some words not covered, lots of words divided into morphemes
(C7) Very easy to use	(C7) Poor vocabulary
(C8) Able to provide definitions of words in more languages	

Table 5: Advantages and disadvantages of Reading Tutor

WWWJDIC	
Positive comments	**Negative comments**
(C1) Good coverage of inflections, compounds, sayings	(C1) Also lists words that appear in example sentences
(C2) Good dictionary, specialized dictionaries, links to other applications	(C2) Not easy to use, web page hard to follow
(C3) Good coverage of inflections, good vocabulary (phrases, sayings)	(C3) Some words / meanings not covered
(C4) Lots of words	(C4) Cannot find all the words I search for
(C5) Kanji search through radicals	
(C6) Multiple ways to search for kanji, their readings etc.	
(C7) Lots of sentence examples	

Table 6: Advantages and disadvantages of WWWJDIC

In contrast, the positive comments for WWWJDIC mostly refer to the coverage of the dictionary content (C1-C4), and some to the tool's additional functions (C5-C7). Half of the negative comments address the not-easy-to-use interface (C1, C2), while the other half note missing data with the dictionary.

Finally, the survey also confirms that students frequently prefer to use a given e-resource in combination with other resources in order to recheck some meanings, readings, translation, as well as to search for entries not covered by other e-resources.

6. Conclusion

The evaluation presented in this paper confirms that the targeted e-resources for Japanese are supportive of learners in carrying out reading tasks in Japanese and that these Japanese language learners are positive towards using the resources. The research also sheds some light on aspects of the targeted resources and their tools that should be improved. Most frequently, misinterpretations occur because of overly-fine segmentation of Japanese words or combination of words, leading students to wrong readings of kanji characters and wrong translations. This is especially the case for proper names, toponyms, compounds and idiomatic expressions, and for foreign words written in *katakana*. The integration and use of multiple resources, with incorporated multiple dictionaries, proves to be the most efficient in achieving complete and accurate results.

It is interesting to notice that the students rather prefer Reading Tutor as a tool for reading support, even though WWWJDIC was shown to have better performance in terms of content coverage. This indicates that the user interface plays a very important role in selecting support resources for language learning. Japanese language learners also showed a preference for using pop-up dictionary support. Considering that some of the resources, for example Perapera-kun, use the data available at the WWWJDIC server, one may expect this to be a promising combination for language learners.

There is a gap, which we should be aware of, between how supportive an e-resource can be to language learners, on one hand, and how the resource actually enhances the language skills of a learner, on the other hand. Being supportive is helpful and time-sparing in most cases, and, in the long run, that is expected to help in enhancing learner skills. However, support can also lead to over-dependency, which is not inevitably connected to skill enhancement. This is an area where e-resources could be further evaluated and improved in order to create environments that enable learners to directly progress in acquiring their language skills.

7. References

Abekawa, K., Hodoscek, B., Nishina, K. (2011). BCCWJ wo riyou shita nihongo sakubun shien shisutemu Natsume no hyouka. *Tokutei ryōiki kenkyū "Nihongo kōpasu" Heisei 22 nendo kōkai wākushoppu yokōshū*. Tokyo: Monbukagakusyoo kagakukenkyuuhi tokuteiryooiki kenkyuu 'Nihongo koopasu' sookatu ban, pp. 507-512.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge: Cambridge University Press.

Hmeljak-Sangawa, K., Erjavec, T. (2010). The Japanese-Slovene dictionary JaSlo: Its development, enhancement and use. *Cognitive Studies / Etudes Cognitives,* vol 10, pp. 203-216.

Hubbard, P. (2006). Evaluating CALL software. In L. Ducate, N. Arnold (eds.) *Calling on CALL: From Theory and Research to New Directions in Foreign Language Teaching*. San Marcos: CALICO.

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I. & Tiberius, C. (2010). A Quantitative Evaluation of Word Sketches. In *Proceedings of the XIV Euralex International Congress*. Leeuwarden: Fryske Academy. 7pp. (Available at: http://nlp.fi.muni.cz/publications/kilgarriff_xkovar3_etal/kilgarriff_xkovar3_etal.pdf)

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the Xlth EURALEX International Congress*, Lorient, France, pp. 105-116.

Nishina, K., Okumura, M., Abekawa, T., Yagi, Y., Bilac, S. & Fu, L. (2004). Asunaro CALL System: Combining Multilingual with Multimedia. In *International Symposium on Large-scale Knowledge Resources LKR 2004*, pp. 69-72.

Srdanović, I., Erjavec T. & Kilgarriff, A. (2008). A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing*) 15/2, pp. 137-159. (Also available at: http://www.jstage.jst.go.jp/article/imt/3/3/3_529/_article).

Srdanović, I. (2011). Collocational Relations in Japanese Language Textbooks and Computer- Assisted Language Learning Resources. *Acta Linguistica Asiatica*, 1(1), pp. 85-97. (available at http://revije.ff.uni-lj.si/ala/index).

The Reversed Dutch-Slovene database: Shortcomings and Some Contrastive Linguistic Issues

Anita Srebnik
University of Ljubljana
Aškerčeva 2, 1000 Ljubljana, Slovenia
E-mail: anita.srebnik@guest.arnes.si

Abstract

In 2007 a Dutch-Slovenian dictionary was compiled for the purpose of decoding. The paper describes the process of reversing it and its result; the converted Slovenian-Dutch database and the subdivision of entries into distinct groups. A general review of the converted database addresses some contrastive relations between Dutch and Slovene; valuable encoding information for the new dictionary, two-way translation dynamics, and the discovery of some inconsistencies of the source Dutch-Slovenian dictionary. The shortcomings of the reversing process, like the absence of multi-word headwords and the parts of headwords inside brackets will have to be resolved in the next conversion round. The main problem to be solved is insufficient lemmatizing and POS-tagging of Dutch separable verbs used separately in a context − which proved to be a universal issue of tagging Dutch texts. The analysis of the converted database is only one step towards a reversed dictionary. Prior to the inclusion of the converted and post-edited database into a reversed dictionary, a monolingual analysis of Slovenian has to be carried out by various monolingual sources and tools.

Keywords: bilingual dictionaries; reversed database; contrastive analysis; two-way translation dynamics

1. Introduction

There are a very few lexicographic resources for the language pair Dutch/Slovene. Next to two pocket-size dictionaries and a Slovene-Dutch dictionary on the internet, the only more comprehensive dictionary is a Dutch-Slovene dictionary (Srebnik, 2007) in paper form. Accordingly, a new Slovene-Dutch dictionary is still waiting to be compiled.

The main purpose of reversing a dictionary is to “maximize the abundance of information” (Krek *et al.*, 2008) in a source dictionary, and to reuse the already established cross-linguistic equivalents. It therefore seemed reasonable to reverse the existing Dutch-Slovenian dictionary before starting to compile a companion Slovenian-Dutch dictionary to gain maximum benefit from the information already contained in a monodirectional dictionary. The reversal was done automatically by Amebis, a software company for language technologies involved in NLP.

The purpose of reversing is rather twofold; not only can a reversed database serve as a database for a new Slovenian-Dutch dictionary, but at this stage of our research it is a valuable source for the analysis of Slovenian as seen from the Dutch perspective, taking into account the lexicographic context. The reversed database represents only “the mirror image” (ibid.) of the Dutch-Slovenian contrastive relation. The analysis of contrastively relevant aspects can contribute to well-founded decisions when compiling a new dictionary.

2. The source

The source of the dictionary data that has been reversed is an XML database of the printed Dutch-Slovene Dictionary (Srebnik 2007). It is monodirectional and primarily intended for Slovenian users, learners of Dutch, to help them with understanding and translating from Dutch into Slovenian. The main function is thus passive or receptive, the secondary function, however is rather active or productive; by means of grammatical information and illustrative examples a more skilled user could produce written or spoken texts in Dutch. The dictionary is not a corpus-based work, but has been compiled by means of numerous available paper and electronic resources involving Dutch, mainly bilingual and monolingual dictionaries and Dutch learning methods. The macrostructure comprises 11,200 of the most frequent and common lexemes, which belong to the basic vocabulary of the Dutch. Therefore, no specialized, archaic, or dialectal variants of language are included. In the microstructure stress is laid on numerous examples of usage, and on significant context in which a headword appears. As the author of the dictionary, I based the decision about the inclusion of a fair amount of pregnant context on the experience that for a learner it is not enough to know only the meaning of the word, it is equally important to be able to see it in a broader context − and in an ideal case to use it actively.

The mono-directional concept dictated the selection of Dutch headwords and illustrative examples that were often based upon expected difficulties in the Slovene translations. It is more interesting to observe what kind of “mirror image” of Slovenian would show up in the reversed database.

The dictionary was compiled with the help of an editor called ADICTED (Anita's DICTionary EDitor), which was developed for the Dutch-Slovenian dictionary at the Institute for Dutch Lexicology (INL) in Leiden with an underlying XML format.

3. The reversal process

3.1 Other languages

Given a vast amount of dictionary compilation software available nowadays, the reversal of a bilingual dictionary is technically much easier than ever before. Several reversal projects have been completed so far, with various software and diverse outcome (see e.g. Geisler, 2002; Honselaar and Elstrodt, 1992; Tamm, 2002; Krek *et al*., 2008; Maks, 2007; Prinsloo and de Schryver, 2002; Newmark, 1999; Veisbergs, 2004; Martin and Tamm, 1996). Seen from the results that the reversion of the monodirectional Oxford DZS English-Slovenian Dictionary has yielded, the reversal can be rewarding and provide the lexicographer with useful information for the new dictionary and relevant contrastive data for further analysis.

In some cases, especially in the Dutch experience of reversing dictionaries with the OMBI, a tool for creating and editing bilingual dictionaries, the reversal is planned from the outset and should be anticipated in the design of a bilingual dictionary, e.g. Dutch-Finnish v.v., Dutch-Estonian v.v., Dutch-Turkish v.v., Dutch-Arabic v.v., Dutch-Hungarian v.v., Dutch-Polish v.v., Dutch-Italian v.v., and Dutch-Swedish v.v. (Martin and Tamm, 1996; Laureys, 2007).

The reversibility doesn't need to be integrated into the compilation process from the outset, and can nonetheless yield successful results, as has been proved by Krek's team who reversed the English-Slovene dictionary.

3.2 Dutch-Slovenian database

Also reflections upon the converted Dutch-Slovenian database reveal that the underlying XML structure forms a flexible base for a conversion process even if the reversing has not been planned from the beginning when compiling the L1 (source language) → L2 (target language) dictionary. The XML structure allows arbitrary sequencing of different information categories. Before reversing the lexicographer is supposed to define the sequence of XML elements with information categories which are coded according to a specific system, and consequently the previous elements are labelled anew.

The most obvious reversal is between a Dutch headword in a source dictionary which now becomes a translation equivalent of the previous Slovenian translation equivalent and vice versa and between illustrative example in Dutch which now becomes a translation of the Slovenian example.

However, the conversion process enables the tracing back of every new element to its position in a source dictionary, especially previous Slovenian translation equivalents of new Dutch translation equivalents.

Here is an example of an illustration of a reversed dictionary entry for *alarm* (alarm) in an XML format:

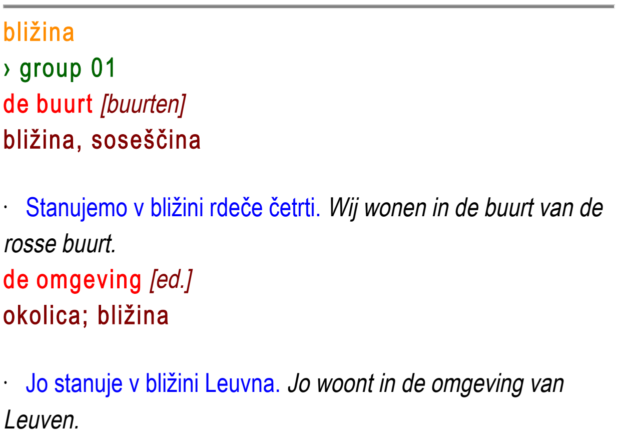
```
<article>
<hw>alarm</hw><hwx>alarm</hwx>
<pron><IPA/></pron>
<prio>01</prio>
<gen>het</gen>
<tText>alarm</tText>
<tText2>al<L>a</L>rm</tText2>
<pl>ed.</pl>
<Ann>
<trx>alarm, preplah</trx>
<part>1</part>
</Ann>
<e>
<eMarker/>
<eFromL>lažni alarm</eFromL>
<eToL>loos alarm</eToL>
<Ann>
<hwe>loos</hwe>
<parte></parte>
</Ann>
</e>
</article>
```

Figure 1: Slovenian-Dutch database, XML-format

In the next phase both the Slovenian and Dutch texts had to be POS-tagged and lemmatized. For the Slovenian, this has been done fully automatically by the proprietary tagger owned by the Amebis company. It was also necessary to POS-tag and lemmatize the Dutch part of dictionary examples in order to be able to detect the existence of translation equivalents. This has been done at the INL in Leiden by means of FROG (Dutch morpho-syntactic analyzer and dependency parser).

After that the dictionary was ready to be converted following the routine of XML elements defined by the lexicographer. This operation was done automatically by the Amebis company.

Additionally, an XSL file was created to enable the lexicographer to visualize the data chosen to be seen in a user-friendly form. The final result is the following:



bližina

› group 01

de buurt *[buurten]*

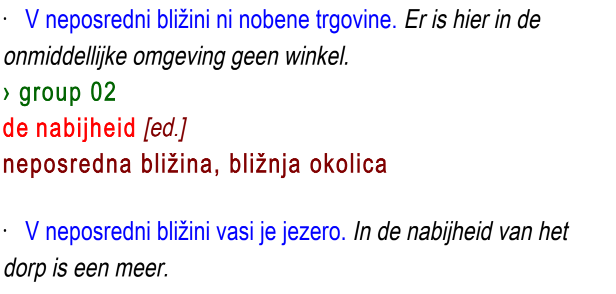
bližina, soseščina

· Stanujemo v bližini rdeče četrti. *Wij wonen in de buurt van de rosse buurt.*

de omgeving *[ed.]*

okolica; bližina

· Jo stanuje v bližini Leuvna. *Jo woont in de omgeving van Leuven.*



· V neposredni bližini ni nobene trgovine. *Er is hier in de onmiddellijke omgeving geen winkel.*

› group 02

de nabijheid *[ed.]*

neposredna bližina, bližnja okolica

· V neposredni bližini vasi je jezero. *In de nabijheid van het dorp is een meer.*

Figure 2: Slovenian-Dutch database, XSL-file

Taking into account that the recent reversion has only been the first attempt towards a new Slovene-Dutch reversed database and that in the short time available for evaluation a few draw-backs showed up, we will briefly compare the source dictionary and its reversed database in terms of numbers:

Dutch-Slovene dictionary
headwords: 11.117
examples of usage: 9.153
translation equivalents: 13.117

Slovene-Dutch reversed database
headwords: 11.465
examples of usage: 39.716
translation equivalents: 26.192

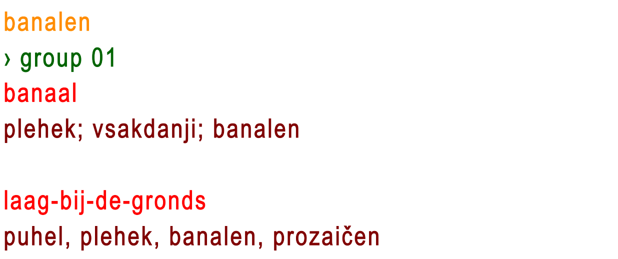
The high number of headwords is expected to be higher after the next reversion round because of the absence of multi-word headwords and those headword parts, which are in between brackets. The latter got partly lost in the process, so that for example the Slovenian translation of the Dutch noun *berouw* (remorse) *kes(anje)* which is a compressed form of two variants of the same lemma, *kes* and *kesanje*, was only registered as one word (*kes*).

The high number of examples of usage in the new database stands out most since the Slovene example is automatically listed under a new Slovene headword if it contains that headword. Still, this was done consistently during the automatic process, which will be illustrated later on.

3.3 Organizing principles of the converted database

Krek *et al*., (2008) invented an innovative approach to the organizing principles of the reversed dictionary database which resulted in an enhanced reversing process with built-in categorization of the material. Their Slovenian-English database was a much more extensive database with a more complex and detailed structure. The research resulted in the article 'The Funny Mirror of Language,' (ibid.) where they introduced four distinct groups into which the new entries are subdivided. The same model for reversion has been applied for the Slovene-Dutch converted database. The groups are clear-cut and most helpful as a starting point for any kind of contrastive analysis. Otherwise the reversed material would be too unorganized and less accessible, and valuable information would be more difficult to find. The categories are the following drawing on the above mentioned article by Krek *et al.*:

1. group “one to one”: the new Slovenian headword appears as a one-to-one translation of the new Dutch candidate for the translation equivalent. The corresponding Slovenian examples from the entire database where a one-to-one translation appears in the Dutch part of the example are grouped under each equivalent. E.g.:



banalen

› group 01

banaal

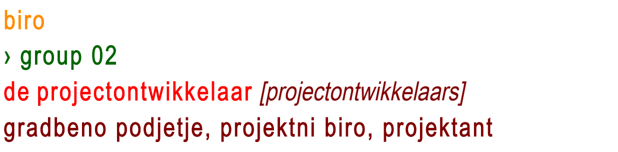
plehek; vsakdanji; banalen

laag-bij-de-gronds

puhel, plehek, banalen, prozaičen

Figure 3: Translation group 1

2. group “one to multi-word + base form”: in this case the new Slovenian headword appears as a part of the multi-word Slovenian translation equivalent in the Dutch-Slovenian dictionary and is used in its base form. E.g.:



biro

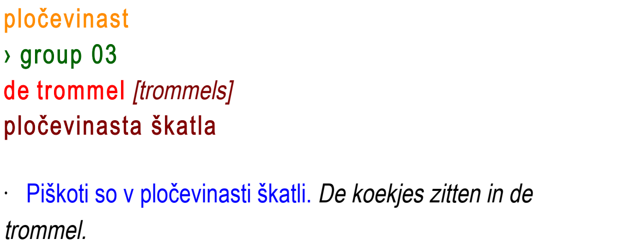
› group 02

de projectontwikkelaar *[projectontwikkelaars]*

gradbeno podjetje, projektni biro, projektant

Figure 4: Translation group 2

3. group “one to multi-word + inflected form”: in this case the new Slovenian headword appears as a part of the multi-word Slovenian translation equivalent in the Dutch-Slovenian dictionary and is used in one of its inflected forms. E.g.:



pločevinast

› group 03

de trommel *[trommels]*

pločevinasta škatla

· Piškoti so v pločevinasti škatli. *De koekjes zitten in de trommel.*

Figure 5: Translation group 3

4. group “no translation”: the new Slovenian headword, e.g. adjective *dobro* (good), is used in the Slovenian part of the example: *To pravim v tvoje dobro.* (I am saying

this for your own good.), but none of the Dutch translation equivalents from previous groups 1 to 3 (goed, lekker, knap, wel, zoet) are used in the Dutch part of the example: *Dat zeg ik om je eigen bestwil*.

According to Krek *et al*. (2008), "the last group is seen as particularly useful since it exposes contrastively interesting cases where in the English-Slovenian dictionary (in our case the Dutch-Slovenian), lexicographers had to find a solution that did not include the most common translation equivalents for a particular headword."

4. General assessment of the first reversion round

4.1 Contrastive relations between Dutch and Slovene

As the reversion will have to be activated once again with more narrowly defined input because it did not yield expected results for the analysis, the results of the first general review which follow here can not be elaborated and complete. Below are some first reflections upon the contrastive relations in the reversed database between the Dutch and Slovenian:

4.1.1 Valuable encoding information for the new dictionary

A new dictionary is supposed to be compiled in the first place for Slovenians to actively use Dutch, so extra information is needed about grammatical and collocational behavior of a Dutch lemma. An instance of good usage examples, found under the headword *akademija* (academy) would be: *študirati na akademiji za likovno umetnost* (to study at the art academy) → *studeren aan de kunstacademie, aan/op de kunstacademie zitten.* A Slovenian speaker learns how to properly use the lemma, and that next to a formal translation (*studeren aan de kunstacademie*) he can also use a frequent non-formal combination with two different prepositions (*aan* or *op*) and the verb *zitten*.

A different, but again a very useful type of usage examples which can be directly applied to a new dictionary was found under the headword *bencin* (petrol). The Slovenian sentence *Avto porabi liter bencina na osem kilometrov. / Avto porabi 12,5 litrov bencina na sto kilometrov* (The car does 8 km to the litre.), the first sentence being a literal translation of Dutch and the second a commoner version in Slovene (still the use of a decimal number in this context is unusual, because it is an exact translation), is rendered into Dutch with: *De auto loopt één op acht.* A sentence which would never occur to a Slovene speaker if he translates more or less literally from his own language. The Dutch sentence could be a pragmatic formula, because of its frozen structure and transparent meaning, which makes it a good candidate for a reversed dictionary.

In the database there is an abundance of examples with an encoding function. It is interesting and quite confronting to see, as the only author of the source dictionary, why so many examples of usage are so readily insertable in a new, reversed dictionary while we have converted a primarily passive dictionary and are going to work on an active one. The reason lies in the already mentioned secondary, active or encoding function, and, admittedly, in a personal urge of the lexicographer not merely to explain but also to teach how to use Dutch.

4.1.2 A two-way translation dynamics

As Šorli (2009) and Krek *et al*., (2008) emphasize in their findings about the reversed database, there is the "simple truth about the dynamics of the translation process: *L1 (Source Language) → L2 (Target Language) does not equal L2 (Source Language) → L1 (Target Language)*." The lexicographers should take into account some specific issues "if they are to avoid falling into traps set continually by the reversed perspective. Typically, the L1 content will be a self-contained semantic unit and, ideally, rendered into L2 with an equally natural and/or lexically frozen semantic unit. However, in many cases the levels of this naturalness and fixedness differ, sometimes considerably. The key problem is not so much that of semantic equivalence, but rather that of equivalence in terms of typicality/ frequency." Below are some examples:

- multi-word lexical units as entry headwords:

umazana posoda (a washing up) = *de vaat, imeti slavnostno večerjo* (to have a dinner, a formal meal or to celebrate something) = *dineren* in one of its meanings*, ostati doma* (to stay at home) = *thuisblijven.* The Slovenian multi-word lexical units are quite acceptable as a translation equivalent in a source dictionary, but the lexical ties between their individual items are not strong enough to be used as headwords.

- insufficiently contextualised lexical strings as illustrative examples:

jata škurhov (a flight of whimbrels)= *een vlucht regenwulpen*

The context in the Dutch-Slovenian dictionary is determined by the Dutch perspective. In the reversed dictionary this example is not suited to the needs of a Slovene speaker. Firstly, this sort of migratory birds is extremely seldom in Slovenia, secondly, the corpus analysis reveals that it is mainly used with adjectives *veliki* (big) or *mali* (small), and never with *jata* (a flight), and thirdly, the example of usage is quite acceptable in the explanatory source dictionary since it does not only refer to a common bird in the Netherlands but also to a title of a very famous Dutch novel:

- idiomatic lexical strings:

An idiom can often be translated by an idiom in another language: *het topje van de ijsberg* (the tip of the iceberg) *= vrh ledene gore, To nima ne repa ne glave.* (There's neither rhyme nor reason to it.) = *Daar is geen touw aan vast te knopen., Malo je čez les.* (He's got a screw loose.) *= Hij ziet ze vliegen.* But quite often, the translations in a source dictionary can be too generic or explanatory; *prenehati s čim* (make an end of something) = *een punt achter iets zetten*, *zmesti se = de kluts kwijt zijn, obvladati kaj* (to master something) *= iets onder de knie krijgen* to be considered as the source idioms for the new dictionary.

4.1.3 Shortcomings of the source dictionary

"The mirror image" of Slovenian can confront the lexicographer with inconsistencies in the source dictionary or simple mistakes which he sometimes could not have been able to notice if the base had not been reversed. In this way he can improve the source dictionary. A quick look at the headword *antologija* (anthology), which is rendered into Dutch by the near synonyms *anthologie* and *bloemlezing*, reveals that in the source dictionary *anthologie* had been translated by *antologija, izbor*, and *bloemlezing* by *zbirka, antologija* which means that both strings of translation equivalents lack one more, *zbirka* must be added to the equivalents of *anthologie* and *izbor* to those of *bloemlezing*.

5. Shortcomings of the reversion process

As already mentioned, there have been some shortcomings discovered so far, which need to be improved during the next reversion round. Unfortunately, not all of them are due to the reversion process.

One of the most conspicuous drawbacks is the complete absence of Dutch illustrative examples containing separable verbs, when the prefix is used separately from the verb. Those sentences together with their Slovene translations do not appear under the new Slovenian headword, which is a translation of that compound verb. Sometimes they appear under another lemma in the sentence, but never under a lemma of that verb. For example, the Dutch sentence with the verb *nadoen* (imitate, copy) *Marja svojo starejšo sestro v vsem posnema.* (Marja copies her older sister in everything.) = *Marja doet haar oudere zusje in alles na.* does not, but should appear under the headword *posnemati* (imitate). The reason for that is, that lemmatizing of the Dutch text was carried out word for word. "The parts of a separable compound verb /.../ are therefore each allocated their own lemma." (Van Eynde, 2004). The prefix in a sentence like above is placed at the end of the sentence and tagged as a preposition in a final position. In this way the verb is not recognized anymore as *nadoen*, but only as *doen*. In Van Enyde's *Protocol for POS tagging and lemmatizing* we can read the following: "'Separability' is not included because allocating the values needs a full syntactic analysis." This is one of the major reversion drawbacks and at the moment it is not clear yet, how to fix it so as to avoid time-consuming manual work.

It also has to be found out why the sentences with separable verbs are arbitrarily distributed under some lemma's and why the others are simply left out. For instance, the usage example in the previous paragraph appears as an illustrative sentence under the headword *star* = *oud* (old), *vse* = *alles* (everything), and *v* = *in* (in), but not under the headword *sestra* = zus (sister) and, as already mentioned, under the separable verb *posnemati* (imitate).

Another issue to be resolved during the next conversion round is the use of brackets containing affixes especially in Slovene verbs, but also other word classes. During the conversion only parts of the words outside brackets were considered new headword candidates, consequently a lot of headwords were left out. The use of brackets is an economical way of spelling Slovene imperfective/perfective verb pairs. Translation equivalents of the Dutch *knetteren* are listed as: *(za)prasketati*, *(za)pok(lj)ati*, *and (za)hreščati*. Written out fully, they would yield eight translation equivalents contributing to eight new headwords in a reversed database: *zaprasketati*, *prasketati*, *zapokljati*, *pokljati*, *zapokati*, *pokati*, *zahreščati*, and *hreščati*. Verb prefixes together with the verb root convert the imperfective verb into a perfective one.

Because the parts (usually verb prefixes) in brackets were not merged with the rest of the words, illustrative examples got lost as well. The sentence *V kuhinji sta vse prepleskala na rumeno.* (They painted the whole place yellow in their kithcen) = *Ze hebben in hun keuken de hele boel geel geschilderd.* appears four times: under the headwords *kuhinja* (kitchen)*, v* (in)*, na* (on) and *ves* (all), but not under the headword *prepleskati* (to paint). The reason lies in the spelling of the Slovenian translation equivalent *(pre)pleskati,* where *(pre)* has been ignored by the conversion programme. This issue is going to be resolved by means of regular expressions which will enable merging parts of the word into one.

Another issue touches upon entry division into the above four categories, but not accurately implemented in the conversion process. Only a few examples of usage that fell into the fourth category do actually belong there. What exactly has triggered the incorrect categorization remains to be discovered so that the next conversion round will yield better results.

A more detailed and comprehensive analysis of contrastive relations can only be carried out after the next reversal, when the data will be more complete – with more headwords due to the inclusion of multi-word translation equivalents and due to merging parts of words inside brackets into new words. The issue of the fourth category "no translation" must be resolved, and the Dutch lemmatizing and POS tagging of separable verbs which are used separately in usage examples, must be corrected.

6. Conclusion

A general review of the converted Slovene-Dutch database addresses some contrastive relations between the languages and the shortcomings of the reversing which will be resolved in the next conversion round. An analysis of a reversed database represents only one step towards a new reversed dictionary, and can only be done thoroughly after the next conversion round. It remains to be seen to what extent are the results of an automatic reversion applicable for the production of the L2-L1 dictionary. All this taken into account, the sources and tools for the analysis of Slovenian, like the Gigafida Reference Corpus of Slovenian (http://demo.gigafida.net/), the Slovenian Lexical Database (LBS), the Word Sketch Engine, a corpus tool that analyses a word's grammatical and collocational behaviour, may adjust the "distorted image of Slovenian" (Krek *et al*., 2008) determined by the Dutch. The monolingual analysis must be done prior to the use of the converted Slovenian-Dutch database in a post-editing phase.

7. Acknowledgements

The reversion would never have been done without the encouragement and continued support of Simon Krek adnd his professional expertise. I am also grateful to the Amebis company and its director Miro Romih for reversing my dictionary and Iztok Kosem from the Trojina Institute for all his support and advice.

8. References

Geisler, C. (2002). Reversing a Swedish-English dictionary for the Internet. In L. Borin (ed.) *Language and Computers, Parallel Corpora, Parallel Worlds.* Amsterdam: Rodopi, pp. 123-133.

Honselaar, W., Elstrodt, M. (1992). The electronic conversion of a dictionary: from Dutch-Russian to Russian-Dutch. In H. Tommola, K. Varantola, T. Salmi Tolonen & J. Schopp (eds.) *EURALEX '92 Proceedings I-II; Papers submitted to the V EURALEX International Congress on Lexicography in Tampere, Finland. Part I.* University of Tampere: Department of Translation Studies. Studia Translatologica, pp. 229-237.

Krek, S., Šorli, M. & Kocjančič, P. (2008). The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XII EURALEX International Congress.* Barcelona: Universitat Pompeu Fabra, pp. 535-542.

Laureys, G. (2007). Optimizing Procedures for the Making of Bilingual Dictionaries and the Concept of Linking Contrastive Lexical Databases. *International Journal of Lexicography,* 20(3), pp. 295-311.

Maks, I. (2007). OMBI: The Practice of Reversing Dictionaries. *International Journal of Lexicography,* 20(3), pp. 259-274.

Martin, W. (2007). Government Policy and the Planning and Production of Bilingual Dictionaries: the 'Dutch' Approach as a Case in Point. *International Journal of Lexicography,* 20(3), pp. 221-237.

Martin, W., Tamm, A. (1996). OMBI: An editor for constructing reversible lexical databases. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, & C. Röjder Papmehl (eds.) *EURALEX '96 proceedings I-II. Papers submitted to the VII EURALEX international congress on lexicography in Göteborg, Sweden.* Göteborg: Göteborg University, pp. 675-687.

Newmark, L. (1999). Reversing a One-Way Bilingual Dictionary. *Dictionaries,* 20, pp. 37-48.

Prinsloo, D.J., de Schryver, G.-M. (2002). Reversing an African-Language Lexicon: *the Northern Sotho Terminology and Orthography No. 4* as a case in point. *South African Journal of African Languages,* 2., pp. 161-185.

Srebnik, A. (2007). *Nizozemsko-slovenski slovar / Nederlands-Sloveens woordenboek.* Ljubljana: DZS.

Šorli, M. (2009). Pridobivanje podatkov o slovenščini za izdelavo slovensko-tujejezičnih slovajev. In M. Stabej (ed.) *Infrastruktura slovenščine in slovenistike. Simpozij Obdobja 28.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 359-369.

Tamm, A. (2002). Reversing the Dutch-Estonian Dictionary to Estonian-Dutch. In A. Braasch, C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress. Vol I.* Copenhagen: CST, pp. 389-399.

Van Eynde, F. (2004). Protocol for POS tagging and lemmatizing. Accessed at:
http://www.tst.inl.nl/cgndocs/doc_English/topics/annot/pos_tagging/tg_prot_en.pdf (accessed: 29 September 2011)

Veisbergs, A. (2004). Reversal as Means of Building a New Dictionary. In G. Williams, S. Vessier, (eds.) *Proceedings of the XI EURALEX International Congress.* Lorient: UBS, pp. 327-332.

Online Dictionaries for immigrants in Greece: Overcoming the Communication barriers

Anna Vacalopoulou, Voula Giouli, Maria Giagkou and Eleni Efthimiou
Institute for Language and Speech Processing, R.C. "Athena"
Artemidos 6 & Epidavrou, Maroussi, Greece
Email: {avacalop; voula; mgiagkou; eleni_e}@ilsp-athena-innovation.gr

Abstract

In this paper we describe on-going work aimed at the creation of a suite of specialized Language Resources (LRs) intended for users not previously targeted at, namely, adult immigrants in Greece. The ultimate goal being to help them integrate in the Greek society, we aim to provide support touching at basic linguistic, social and everyday issues. The suite comprises: (a) bilingual dictionaries integrating a grammar component; (b) sample typical dialogues, relevant to communicative situations that the target group is most likely to cope with; and (c) a multilingual parallel text corpus that adheres to domains that are of interest to the target group. These LRs will be integrated into a web interface coupled with advanced search mechanisms that will provide innovative accessibility options for visually impaired users. The paper describes the intended LRs suite elaborating on the corpus compilation and processing, as well as on the dictionaries macro- and micro-structure the focus being on the methodological principles underlying selection and organization of the dictionary entries.

Keywords: bilingual dictionaries; immigrants in Greece; user needs and requirements; corpora

1. Introduction

The unprecedented growth of immigrant population in Greece over the last decade has led to the adoption of policies aimed at their smooth integration and social inclusion. In this context, language education plays a central role within the action plans and measures taken. The document reports on work still in progress within the framework of *eMiLang* project. It elaborates on the bilingual dictionaries that are being developed in order to support the communicative needs of immigrants in Greece. Section 2 outlines the project scope and aims, which ultimately guided *dictionary design* and *platform functionalities*, whereas Section 3 briefly describes the target group with respect to main characteristics, and their needs and requirements. Section 4 elaborates on the dictionary specifications (macrostructure and microstructure), the emphasis being on the provisions taken towards addressing the specificities of the target group. The lexicographic considerations taken into account in lemma selection are presented in section 5, whereas the primary data (corpora) that comprise the sets of bilingual textual collection and the methodology adopted for collecting them are presented in section 6. In the last section, we present conclusions and future work.

2. The framework: eMiLang project

The *eMiLang* project aims to develop a *digital infrastructure* tailored to support adult immigrants in Greece to overcome the communication barriers in their everyday interactions, and in administrative, social and educational settings. The ultimate goal is to assist both immigrants and policy makers in their joint efforts for smooth integration of the target groups to the Greek society. The intended infrastructure encompasses two inter-related pillars: (a) the development of LRs, namely **specialized multilingual parallel corpora** in the form of informative material and **bilingual dictionaries** (*partly* extracted from these corpora), and (b) the implementation of a multilingual, multimedia web **interface** designed so as to integrate the digital content (dictionaries and informative material). This interface will also offer advanced search mechanisms and information retrieval capabilities. Finally, a news aggregator will be integrated into the system, offering digital information services to the users.

3. The target group: needs and requirements

As it is evident, dictionary design in terms of language coverage, entry selection and presentation mode is *user-oriented*. The user perspective in dictionary making is considered along the following axes: (a) users' *reference needs*; (b) their *proficiency level* and *background knowledge*; (c) their reference *skills* and *strategies*; and (d) effectiveness of dictionary use training (Varantola, 2002).

To infer the needs and requirements of the target group, an investigation had to be conducted in order to primarily identify their profile(s) and respective needs. One major difficulty in this task was the inability to perform proper analyses employing appropriately designed questionnaires and tests as proposed by mainstream lexicographic research (Atkins, 1998). This was mainly due to the fact that locating the intended users at such an early stage of the dictionary-making process and persuading them to participate in any type of survey was extremely difficult, since these people whose upmost concern is to struggle for a living in a new and unknown environment. Instead, we opted for postponing any immediate contact with the target group until a first version of the platform was available for on-line user feedback elicitation during the pilot use phase. In fact, this approach is consistent with what has been called "simultaneous feedback" from the target users to the compilers (De Schryver et al., 2000). Based on the assumption that in most cases, user feedback usually comes too late because it can at best be considered for implementation in the revised edition of the dictionary, this approach caters for the identification of prospect user needs and preferences by launching/testing pre-final dictionaries coupled with questionnaires. In this way, hypotheses may be tested, and refinements and modifications can be implemented where needed and in the light of feedback obtained by the users during dictionary development.

Thus, to initialize dictionary compilation, consultation of official, general-purpose statistical data took place. As a matter of fact, there are relatively few data available detailing immigrants in Greece and their characteristics. And apart from the sparse quantitative and qualitative surveys on immigration (Baldwin-Edwards, 2004; 2008), the only sources available were the *2001 Census* survey, along with figures obtained from *Eurostat (*http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat). Information thus obtained reveals the principal immigrant nationalities in Greece as being *Albanian*, *Bulgarian*, *Georgian*, *Romanian*, *Russian*, *Ukrainian*, *Polish*, *Pakistani*, and *Egyptian*. Moreover, regarding age, the vast majority of immigrants fall within the range 15-64. Dependent employment has also been recorded as the principal reason for award of residence permits (68% of the total). Following this, roughly equal at 12% each, are family reunification and self-employment, whereas very few immigrants enter Greece for study purposes. Moreover, as far as *immigrants' presence in the Greek labour market* is concerned, Census data regarding male immigrants' main occupations, the principal employment has been in building construction, followed by agriculture, industry and tourism. Female employment is dominated by occupations such as housekeeping and cleaning, and also employment in sectors such as agriculture, and tourism.

The characteristics of the immigrant population with respect to educational level and language literacy were also obtained from the aforementioned sources. According to *Census 2001* survey and *Eurostat* data, the vast majority of immigrants in Greece are of educational levels ranging from medium to law. More specifically, statistical data show that immigrants in Greece mainly fall in one of the following three groups with this respect. The first group comprises immigrants who have completed secondary education before entering the country; the second class consists of those who failed to progress beyond primary school. Both classes are populated with people originating from European countries (Albania, Bulgaria, Poland, and Serbia). The last one, which comprises immigrants from countries in Africa and Asia, includes those who are classed as illiterate. Additionally, most of the immigrants were reported as having little or no knowledge of the Greek language prior to entering the country. Moreover, one can safely infer that the target group has almost no prior experience in the use of dictionaries or other linguistic resources, and that they have low to medium level of computer literacy.

Another interesting fact was the highly visible increase in the number of immigrant children recorded in state schools, especially since the mid-1990s. And although this seems to be a difficult problem to tackle per se, the difficulties immigrant parents face when communicating with their children's tutors have been reported as also being a problematic issue (Baldwin-Edwards, 2008).

From all the above, we conclude that the intended target group is diverse in terms of nationality, level of literacy and language proficiency in Greek, yet the tendency is for lower a level. On the basis of users' profile, their needs and requirements were identified or inferred. The ultimate goal being to help immigrants integrate in the Greek society, we aim to provide support touching at basic linguistic, social and everyday issues along the following axes:

(a) communicative needs in official settings (as for example, in dealing with the Greek authorities, applying for a green card, etc.);
(b) communicative needs in social settings;
(c) communicative needs in order to cope with every-day issues (as for example travelling and transportation, etc.);
(d) language learning in formal or informal settings;
(e) familiarization with the general cultural and social context.

4. Dictionary Specifications

The nationalities identified determined the languages to be covered by the bilingual dictionaries, namely: Greek–Albanian (EL-AL), Greek–Arabic (EL-AR), Greek–Bulgarian (EL-BG), Greek–Chinese (EL-CH), Greek–English (EL-EN), Greek–Polish (EL-PL), Greek–Romanian (EL-RO), Greek–Russian (EL-RU), and Greek–Serbian (EL-SR).

Furthermore, the specifications recommended that the whole process of dictionary compilation be corpus-based; this refers to headword selection (in order to identify the appropriate vocabulary), sense selection and distinction, and collocations and usage examples extraction. Finally, the specifications stressed the importance of user-friendliness of the dictionary and ease of access as a basic feature of the underlying platform, since it addresses the needs of people as outlined above and, to this end, meta-language should be kept to a minimum.

In the following sections the implementation of these basic guidelines will be presented in more detail.

4.1 Macrostructure

4.1.1 Types of entries

Each bilingual dictionary will comprise around 15,000 entries which cover mainly the basic vocabulary of Greek. And although a formal complete list of basic vocabulary is still missing for the Greek language, in the current implementation, the basic vocabulary is conceived as one which comprises not only the most frequent items but also less frequent words and phrases that are relative to everyday life.

Another substantial category of lemmas is the one often occurring in official, administrative or other documents that the target group is likely to come up with during their stay in Greece, as for example when applying for a residence permit, etc. To this end, selected technical vocabulary, that is, terms pertaining to domains/subject fields that are of utmost interest to the target group have been included as well.

Because of the fact that the target group is generally expected to lack basic encyclopaedic information about Greece, this dictionary also contains proper nouns. These include the names of: (a) geographical entities (i.e., cities, islands, regions etc.), (b) official bodies (i.e., ministries and other official organisations), and (c) geopolitical entities (*Ηνωμένα Έθνη = United Nations*). Both official bodies and geopolitical entities are quite often expressed by acronyms which are also retained in the lemma list.

In terms of form, this dictionary contains two main categories of entries: single-word and multi-word lemmas. Multi-word entries may include expressions (*καλό ταξίδι = have a nice trip*), collocations (*χαρτί υγείας = toilet paper*), etc.

Different alternatives[1] of the same word or phrase are separate entries which are interlinked with each other. For instance, *Ολυμπιακοί Αγώνες* (Olympic Games) and *Ολυμπιακοί* (Olympics) are two separate dictionary entries linking to each other. Similarly, *κινητό τηλέφωνο* (mobile phone) and *κινητό* (mobile) are also listed as stated above. The most 'complete' form of such entries is considered as the main entry and contains the rest of the information in this dictionary. The secondary entry/entries serve as cross-references to the main entry. When two entries linked by cross-reference belong to different registers, the main entry is the most formal type, as more likely to occur in official and/or state documents. In the case of acronyms, the main entry is the full name of the entity (*Ευρωπαϊκή Ένωση = European Union*), the acronym (*ΕΕ = EU*) being a cross-reference. Acronyms are normalised for easy reference and are thus all written without dots among letters.

Although none of the cross-references are fully developed entries, certain types of information are included, so that users can access them immediately, without having to follow the cross-reference link. These are: hyphenation, pronunciation, link between masculine and feminine nouns, the three forms of adjectives (i.e. masculine, feminine and neutral) and domain (see section 4.2).

4.1.2 Lemma distinction

The main criterion for lemma distinction is morphology. Thus, the following are listed as different entries: *Οκτώβριος* and *Οκτώβρης (=October), μέρα* and *ημέρα (=day), εβδομάδα* and *βδομάδα (=week)*. The next criterion for lemma distinction is part of speech. Thus, homographs belonging to different parts of speech form separate entries (*άρρωστος, άρρωστη, άρρωστο = ill*, *άρρωστος = patient*). Because of the difficulties arising from the fact that Greek is a highly inflectional language, the past participle of verbs is treated lexicographically as an adjective, thus forming a separate entry (*αγαπημένος*, *αγαπημένη*, *αγαπημένο = beloved or favourite, p.p. of the verb αγαπώ = love*; *χαμένος*, *χαμένη*, *χαμένο = lost, p.p. of the verb χάνω = lose*).

Along similar lines, all types of word derivatives are separate entries. Thus, adverbs (*αργά = slowly, διαφορετικά = differently*) are different entries from the respective adjectives (*αργός, αργή, αργό = slow, διαφορετικός, διαφορετική, διαφορετικό = different*).

All single-word entries appear in the 'base' form, in the way that would be expected to appear in regular monolingual dictionaries: for verbs, this is the first person singular present in the active voice; for nouns, the singular nominative; for adjectives and past participles, the nominative positive (in this case, in the masculine, feminine and neutral); for adverbs, the positive. Exceptions to the above rules occur when what is usually considered as the 'base' form is either ungrammatical or particularly infrequent in Greek (*χιονίζει = it snows*, the third instead of the first person, *λεφτά = money*, the plural instead of the singular, *συναχώνομαι = catch a cold* instead of *συναχώνω = cause somebody to catch a cold*).

Nouns referring to professions or other people's activities form two different entries (masculine and feminine) as, in most cases, their morphology differs (*δάσκαλος* and *δασκάλα = teacher, ιδιοκτήτης* and *ιδιοκτήτρια = owner, ταξιτζής* and *ταξιτζού = taxi driver*). Exceptions to the above rule would be nouns with identical masculine and feminine forms (*ηθοποιός = actor* and *actress*; *ταμίας = male* or *female cashier*; *υπουργός= male* or *female minister*).

The comparative and superlative of certain highly frequent adjectives and adverbs also form separate entries. Thus, *λιγότερος*, *λιγότερη*, *λιγότερο = less* as well as *περισσότερος*, *περισσότερη*, *περισσότερο = more* appear separately from *λίγος = little* and *πολύς = much*, respectively.

[1] For alternative spellings of words or phrases, the main entry follows the official school grammar spelling whereas other spellings are cross-references.

4.2 Microstructure

4.2.1 Meanings and examples of usage

As this dictionary is mainly targeted toward starter learners of Greek who are in need of speedy learning, it has been decided that only basic meanings would be included in it. Meanings however are neither defined nor directly translated; they are implicitly presented through one or more examples of usage, which bear the informative load. Examples of usage are thus a core element of the dictionary.

Furthermore, examples in this dictionary are carefully selected so as to reflect not only the different meanings but also the most basic forms of usage, grammar and/or collocation. Thus, for instance, the active and passive of verbs are presented in separates when voice differentiates meaning as well; the same stands for verbs used with different prepositions etc.

As the emphasis of this dictionary has been to include as much information as possible but in the most user-friendly way possible, examples have been selected so as to be as interesting as possible to the target group. To this end, a combination of different corpora has been used. A large part of the examples for the basic vocabulary was extracted from the Hellenic National Corpus, although usually shortened and/or simplified to suit the target group level.

In terms of length, examples are short and contain no excess information. They usually consist of one simple sentence, although some dialogue is included to exemplify everyday phrases, such as greetings or asking for information. Apart from accelerating the learning process, the brevity criterion also simplifies the ambitious work of translating everything into 9 languages.

As it is customary in most multilingual dictionaries, examples also play the role of describing each meaning, due to lack of definition. This has placed additional difficulty in selecting the right example for each meaning. For instance, an example of the verb *αγωνίζομαι = struggle* would be *Αγωνίστηκε πολύ, για να καταφέρει αυτό που ήθελε = She struggled a lot to get what she wanted.*

Last but not least, taking into account the great variety of backgrounds from which the target group of this dictionary comes, extra care has been taken toward political correctness. All examples are free of any social, political, racial, national, and religious or gender bias.

4.2.2 Communicative/subject domains

Each meaning/example of the entry words is categorized in broad domains that reflect certain communicative contexts an immigrant in Greece may be involved in. As noted above immigrants are a special case of language learners, i.e. their needs are those of a summer week tourist and of an active citizen at the same time. An immigrant has, for example, to go shopping or book an apartment, to register a child in a public school and object to the employer when labour legislation is violated. In this view, the domains have to be detailed enough to cover as many possible different communicative needs and comprehensive enough to facilitate usability. An additional factor that has led to the categorisation of entries into domains is that, according to studies, users rarely go through the list of senses for each dictionary entry, usually selecting the first meaning (Lew, 2004). It is, therefore, suggested that users will be more likely to identify the appropriate meaning of multi-sense entries when these are clearly categorised into domains. These domains are:

- **Education**, e.g. *πανεπιστήμιο* (university), *ΑΕΙ* (acronym for Higher Education Institution)
- **Labour – insurance**, e.g. *επίδομα ανεργίας* (unemployment allowance), *ημερομίσθιο* (wage)
- **Law, justice and public safety**, e.g. *ποινικός κώδικας* (penal code), *δικηγόρος* (lawyer)
- **Finance**, i.e. anything related to money and the economy, including taxation, bank transactions etc.
- **Public administration – politics**, i.e. vocabulary that does not fall into any of the above categories and concerns administration, bureaucracy, the government, the political framework etc., for example, *βουλή* (parliament), *πιστοποιητικό οικογενειακής κατάστασης* (civil status certificate)
- **Transportation and travel**, i.e. vocabulary related to urban transport and travelling in general
- **Geography**, which will include an extensive list of countries, nationalities and languages, as well as all the major Greek cities and areas
- **Physical condition and health**, i.e. parts of the body, diseases, doctors, etc.
- **Science and technology**, i.e. computers and technological gadgets, some widely used scientific fields and terms
- **Environment**, i.e. flora and fauna, geomorphology, weather, ecology, etc.
- **Culture, recreation and the media**, i.e. vocabulary from the arts, hobbies and spare time, television and the media in general
- **Relations – family**, i.e. words for family and social relations
- **House and accommodation**, i.e. parts of a house, furniture and appliance, as well as vocabulary relevant to accommodation in general, e.g. hotels and rooms to let.
- **Public holidays and Greek traditions**, comprising the most common Greek holidays and celebrations, as well as culture specific traditions that an immigrant is unfamiliar with.

Finally, the most populated domain is, as expected, **general vocabulary**. For educational reasons mainly, part of the general vocabulary will be further subcategorized into distinctive vocabulary groups such as:

- numbers
- clothing and accessories
- food and cooking
- time
- space
- colours
- measurement units
- everyday interaction (informal words and expressions).

4.2.3 Additional entry information

Information for each dictionary entry includes phonetic transcription, pronunciation (audio file), hyphenation, alternative entry types (cross-references), elementary grammatical information (i.e. the masculine, feminine and neutral type for all adjectives and past participles) and examples of usage. Each example is translated into 9 languages, with the entry word/phrase highlighted in the example.

Apart from entries themselves, examples are also pronounced in Greek and Bulgarian using a synthetic voice. This is meant to help people with vision or literacy problems on the one hand and the vast majority of people who are not familiar with the Greek alphabet on the other.

As far as hyphenation is concerned, it is included for all single-word entries, in an attempt to help users compose hand-written or electronic texts. What is more, hyphenation in Greek is not arbitrary, so this feature is expected to help more advanced users familiarise themselves with the basic rules of hyphenation in Greek.
All multi-word entries are interlinked with each of their components (excluding functional words such as prepositions, conjunctions and articles). Not only does this feature help in easy reference, but it also has a pedagogical added value, as most of the words contained in phrases are inflected types of lemmas. Thus, users are guided to link each individual type to the base form of the lemma.

5. Lemma Selection Methodology

Dictionary entries were semi-automatically selected from a variety of sources, including (a) a large (POS-tagged and lemmatized) reference corpus of the Greek language, namely the Hellenic National Corpus (http://hnc.ilsp.gr/), (b) the Greek counterpart of the specialized multilingual parallel corpus, and (c) from already existing dictionaries and glossaries, customized to better suit the user needs (communicative situations and relevant vocabulary, etc.). As it has been noted above, a proportion of the entries is part of what can be conceived as the *basic vocabulary* of Greek. This does not only mean the most frequent items attested in the HNC, but also less frequent words and phrases that are relative to everyday life, and which are used to populate the domains described above (such as *μαξιλαροθήκη = pillowcase* or *πάνα = nappy*).

Similarly, a corpus-based methodology has been employed for the semi-automatic selection of entries which belong to a more technical vocabulary with the use of NLP lingware (see section 7 below), coupled with manual correction and selection of the most frequent/appropriate terms.

Finally, this dictionary follows the closed vocabulary concept, thus including every word in the examples as an entry itself for easy reference. This has led to adding a considerable amount of entries ad hoc and keeping a better balance, in terms of content, between everyday vocabulary and the administrative jargon of the public service.

6. Corpora and Linguistic Processing

The role of corpora in the project is two-fold: (a) to provide linguistic evidence and aid linguistic introspection and (b) to form the multi-lingual informative textual material. Two types of corpora were consulted in this respect: the Hellenic National Corpus (HNC), a large reference corpus of the Greek language, and the *eMiLang* specialized corpus. As it has already been mentioned, the former was used to extract the source language (EL) material (headword selection, sense discrimination, usage examples), whereas the latter has already been used for the extraction of terms adhering to the domains catered for in the project. Additionally, it will form the data pool for the development of the informative multilingual material.

More precisely, the *eMiLang corpus* comprises texts that adhere to domains that are of interest to the target group, namely: *administrative/legal*, *health, education, transport* and *civilization*. The texts have been selected from various sources over the Internet: official websites of public bodies, organizations, the EU portal, etc. This corpus currently amounts to 172K words. As far as balance is concerned, this was achieved for the domains *health, education* and *transport* (c. 30K words each). The domain *administrative/legal* outperformed the other three (c. 95K words) because of the availability of data and data sources. Data pertaining to the *civilization* domain were the most difficult to collect (only 17K words) due to the strict Intellectual Property Restrictions, and they were only kept for the off-line part of the corpus.

One peculiarity of the textual collection at hand is that it will also form the data pool for the creation of the informative material. To this end, documents containing information that is dated or obsolete, were only retained to form the off-line corpus from which linguistic evidence was extracted, and they were appropriately marked so as to be used with consciousness thereof.
A metadata scheme for the efficient representation of the corpus data along with the encoding of the linguistic annotations has been implemented for the efficient

management and retrieval of the textual data. This scheme is compliant to widely accepted standards so as to ensure reusability of the resource at hand, namely the specifications of the Text Encoding Initiative (TEI). Metadata elements have been deployed which encode information necessary for text indexing with respect to text title, source, author, publication date, etc. (bibliographical information) and for the classification of each text according to text type/genre and topic. Metadata elements for catalogue descriptions compatible with the specifications proposed by the ISLE[2] Meta Data Initiative (IMDI) were also added manually to the whole corpus in view of rendering the corpus searchable by prospect users.

After text selection and documentation, extended manual validation (where appropriate) was performed. Normalization of the primary data was kept to a minimum so as to cater, for example, for the anonymisation of the official documents (that is the deletion of person names) and for the conversion of collected files to a format appropriate for further processing.

Text processing was then applied via an existing pipeline of shallow processing tools for the Greek language (Papageorgiou et al., 2002). These processing steps include: (a) part-of-speech (POS) tagging and lemmatization; (b) Named Entity (NE) Recognition; and (c) Term extraction. The Greek POS-tagger has been developed in-house and is based on Transformation Based Learning architecture. Trained on Greek textual data from various sources (newspapers, internet, etc.) it assigns Part-of-speech labels to words in a sentence. Following POS tagging, lemmas retrieved from a Greek morphological lexicon were assigned to every word form. At the next stage, Named Entity Recognition was performed on a subpart of the corpus using a Maximum Entropy Named Entity Recognizer (MENER), a system compatible with the Automatic Content Extraction (ACE) scheme (http://www.itl.nist.gov/iad/mig/tests/ace/), catering for the recognition and classification of the following types of NEs: person (PER), organization (ORG), location (LOC) and geopolitical entity (GPE). For the purposes of the current project only NEs of the types (LOC) and (ORG) were retained. A Greek Term Extractor (TE) was finally used for spotting terms and idiomatic words. TE proceeds in three pipelined stages: (a) morphosyntactic annotation of the domain/specialised corpus, (b) corpus parsing, i.e., identification of syntactic constituents using a pattern grammar endowed with regular expressions and feature-structure unification, and (c) lemmatization. The tool employs a hybrid methodology, in that statistical evaluation of candidate terms skims valid domain terms, lessening, thus, the over-generation effect caused by pattern grammars.

7. Conclusions and Future Work

We have hereby presented work still in progress targeted at the development of on-line dictionaries for immigrants in Greece. Initial considerations involve entry selection and entry organisation the ultimate goal being to better suit the intended users' needs.

Future work involves the implementation of a platform that will be user-friendly, featuring search functionalities for easy access to the entry via the lemma or the word form. To this end, a tool will be integrated, which links each inflected form to a very large morphological lexicon of Greek. This is expected to be of enormous help to the lookup process. Moreover, fuzzy-matching techniques will also be employed, and users who misspell words will be presented with a list of correct spelling alternatives from which they can choose. This is one of the features adding to the pedagogical nature of this dictionary.

8. Acknowledgements

The research within the project "*eMiLang: Improved Informative Digital Services Targeted on Immigrants*" leading to these results has received funding from the Greek state, under the framework of *O. P. Digital Convergence (2007-2013).*

9. References

Atkins, B.T.S. (1998). *Using Dictionaries: Studies of Dictionary use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.

Baldwin-Edwards, M. (2008). *Immigrants in Greece: Characteristics and Issues of regional distribution*. MMO Working Paper No. 10, Jan. 2008.

Baldwin-Edwards, M. (2004). *Statistical Data on Immigrants in Greece*. Athens: Mediterranean Migration Observatory and IMEPO.

De Schryver, G.M., Prinsloo, D.J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. *Lexikos 10*, pp. 1–31.

IMDI, Metadata Elements for Catalogue Descriptions, Version 2.1, June 2001.

Lew, R. (2004). *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English*. Poznań: Motivex.

Papageorgiou, H., Prokopidis, P., Giouli, V., Demiros, I., Konstantinidis, A. & Piperidis, S. (2002). Multi-level, XML-based Corpus Annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pp. 1723-1728.

TEI Guidelines for Electronic Text Encoding and Interchange (http://www.tei-c.org).

Varantola, K. (2002). Use and Usability of Dictionaries: Common Sense and Context Sensibility?. In M.-H. Correard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Grenoble, France: EURALEX, 2002.

[2] International Standards for Language Engineering

Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations

Orsolya Vincze, Margarita Alonso Ramos, Estela Mosqueira Suárez, Sabela Prieto González
Universidade da Coruña
Campus da Zapateira, s/n, 15071 A Coruña
E-mail: {ovincze|lxalonso|estela.mosqueira}@udc.es

Abstract

This paper provides an insight into ongoing research focusing on the exploitation of data from learner corpus in order to enhance the performance of an automatic tool aimed at the correction of collocation errors of L2 Spanish speakers. The procedure adopted for collocation annotation is described together with the main difficulties involved in the annotation task, such as the problem of distinguishing collocations from other kinds of idiomatic expressions and from free combinations, the problem of correction judgment, and the problem of assigning concrete error types. It is shown that the fine-grained typology used in the course of error annotation sheds lights on certain collocation error types that are generally not taken into account by automatic error correction tools, such as errors concerning the base of the collocation, target language non-words, and grammatical collocation errors.

Keywords: collocations; learner corpus; error typology; Spanish as second language; Computer assisted language learning (CALL)

1. Introduction

The present paper forms part of a research project that aims at the development of a CALL environment for learning Spanish collocations. The intended CALL environment is conceived of as a flexible and dynamic tool, which will provide an integrated interface combining several resources such as a collocation dictionary (see Alonso Ramos et al., 2010; Vincze et al., 2011[1]), corpora, and an automatic correction tool.

Following Hausmann (1989) and Mel'čuk (1998), we hold that collocations are restricted binary combinations of two lexical units, where one of the two elements, the base, conditions the choice of the other, the collocate. These idiomatic combinations are considered a major challenge for L2 acquisition. In fact, the difference in collocational knowledge has been found to constitute an important factor that contributes to the difference between native and non-native language use (e.g. Howarth, 1998; Granger 1998; Higueras García, 2006).

Previous work suggests that a CALL environment focusing on collocations can profit from data on learners' actual language behavior obtained from corpus research (Shei and Pain, 2000; Chang et al., 2008). In order to gain information on the collocation knowledge and typical errors of Spanish L2 learners, we annotated correct and erroneous collocations in a portion of the CEDEL2 corpus (Lozano, 2009), a corpus containing essays written by English mother tongue Spanish L2 learners.

This paper is structured in the following way. We start, in Section 2, with a brief review of previous work in relation to collocations in the two main research fields concerning our study: learner corpus and correction tools inside a CALL environment. Section 3 provides a description of the collocation annotation procedure we adopted, and looks into the three main difficulties involved in the task: 1. recognition of collocations, 2. correction judgment, and 3. error type annotation. Following this, in Section 4, we highlight some characteristics of collocation errors observed in the corpus that generally have not been attended to by automatic correction tools; these are: 1. the location of the error, 2. L1 interference and 3. grammatical collocation errors. Finally, in Section 5, we draw some conclusions on the presented work and give future lines of research.

2. Previous work

As we have mentioned, the object of the present study lies at the crossroads of two fields of research. Within the first of these, learner corpus research, collocations have been the subject of a considerable number of studies since Granger's (1998) seminal work. These constitute quantitative and qualitative studies comparing learners' and native speakers' collocation production. However, none of them goes into such detail in terms of error analysis as our own research. Some of the most recent studies in this field are Nesselhauf (2005), Martelli (2007) and Thewissen (2008).

Similarly, in the field of CALL, we find several proposals aiming at creating an automatic tool for the correction of collocation errors. Notably, some of these, such as Shei and Pain (2000), Liu (2002) and Chang et al. (2008), make use of error analysis data coming from learner corpora. The first proposal uses corpus data to build an error library to enhance the performance of the system, while Liu (2002), in addition to this, also exploits her observation that most erroneous collocates can be related to their correct counterparts through

[1] Diccionario de Colocaciones del Español, available at: www.dicesp.com

semantic relations established in WordNet (Fellbaum, 1998). Finally, re-examining Liu's (2002) data, Chang et al. (2008) emphasize that the great majority of learner collocation errors can be accounted for by L1 interference. Therefore, their system uses bilingual dictionaries to check synonymous translations of erroneous collocates in order to suggest likely corrections.

3. Corpus annotation

With the aim of studying the collocation production of L2 Spanish learners, we manually annotated 100 essays from the CEDEL2 corpus, amounting to 46420 words. The corpus annotation task was carried out by native speakers of Spanish, following a well-defined procedure, and making use of an elaborate typology devised by Alonso Ramos et al. (2010, see below) for labeling collocation errors. However, as the low inter-annotator agreement shows, the annotation task posed a significant challenge, mainly due to the notoriously fuzzy interpretation of the notion of collocation.

3.1 The annotation process

The corpus annotation was carried out by two main annotators, whose annotations were merged and revised by a third, consensus annotator. Annotations lacking consensus were resolved in the following way: Cases of collocations judged to be correct by one and incorrect by the other main annotator were checked against corpus data by the consensus annotator. When at least five cases of the given collocation were found in the *Corpus de Referencia del Español Actual* (CREA), it was considered a correct combination. Cases that could not be resolved using this method were sent to three independent annotators, and subsequently treated according to the majority vote. Finally, dubious annotations and conclusions on merged annotations were discussed in weekly annotation sessions supervised by an expert annotator.

As for the success of the annotation process, we should note that despite the well-defined annotation procedure and the weekly sessions to comment on criteria concerning the new annotations, we were able to achieve only a slight increase in inter-annotator agreement, which however remained considerably low throughout the whole of the annotation process: an average of about 30% during the first weeks and average of about 50% over the last weeks.

This issue is mainly related to the problem of recognizing collocations in the learner texts. In what follows we discuss this and other difficulties affecting the annotation process, such as the problems of correction judgment, and the problems of interpreting errors.

3.2 Problems of recognizing collocations

The problem of recognizing collocations in the learner texts can be ascribed to the difficulty of establishing clear and, most importantly, operational criteria for delimiting the notion of collocation. In practice, this results in the annotators having difficulty in telling collocations apart from free combinations, on the one hand, and from idioms, on the other hand.

For instance, it is quite straightforward to agree on that *buena nota* 'good grade' is a collocation, given that the semantic characteristics of a noun like *nota* 'grade' call for a qualification adjective. This is not so in the case of the combination *buena comida* 'good food', where the meaning of the noun *comida* 'food' does not necessarily require qualification. Consider, however, the combination *comida rica* 'delicious food', where the adjective, *rico* 'delicious', has a rather restricted use; it is the adjective prototypically chosen to speak about good food. From our point of view, combinations such as *comida rica* should be considered collocations, and, consequently, other less idiomatic combinations, containing less restricted adjectives appearing with the same noun, such as *buena comida* 'good food' or even *comida fantástica* 'fantastic food' will be considered collocations as well. An example for the difficulty of distinguishing collocations from idioms is the case of *darse cuenta* 'realize', which should be treated as a non-compositional expression, given its frozen syntactic structure. It was mistaken for a collocation by the annotators due to the fact that the verb *dar* 'give' is often used in light verb constructions, as in *dar un paseo* 'take a walk', *dar consejos* 'give advice', etc.

We also noticed that correct collocations often passed unnoticed by annotators until an incorrect counterpart of the same combination was found. An example for this is the case of *país de origen* 'country of origin', which was not annotated as a collocation until the erroneous combination *países maternos* lit. 'mother(ly) countries' was found in the corpus. At the same time, any error was bound to be perceived as a collocation error by the annotators. For instance, the free combinations *lleno *con historia* lit. 'full with history' and *recorrimos *por la isla* 'we travelled all over the island' were both annotated in the first stage of the annotation process, probably because the preposition errors made them more salient[2].

3.3 Problems of correction judgment

The main issues here are the individual permissiveness

[2] Note that we do not treat as collocations word combinations consisting of a lexical element and its governed preposition (e.g. *depende de* 'depend on'), often referred to in the literature as *grammatical collocations* (Cf. Benson et al., 1986). However, prepositions governed by a member of a collocation (e.g. *tener miedo de* lit. 'have fear of') are considered to form part of the expression as a whole, therefore, when erroneous, they are annotated as grammatical collocation errors (see below, in Section 4.2).

of annotators, on the one hand, and the challenge posed by language variation, on the other hand. Differences in the individual permissiveness of each annotator towards unusual language use led to lack of consensus in judging a lexical combination correct or incorrect. It also appears that annotators tend to be less permissive with non-native speakers in terms of creative or unusual language use than they would be with native peers.

The problem of language variation was noticed especially in the case of collocations typically used in Latin American Spanish. They were judged, at first sight, as incorrect by the annotators, however corpus data showed that these combinations are actually in use in other Spanish-speaking countries. Consequently, these expressions were annotated in the corpus as correct collocations, specifying the language variant they belong to. For example, the combination *hice las reservaciones* 'I made the reservations' was perceived as incorrect, given that European Spanish uses the form *reserva* 'reservation' and not *reservaciones.* We also find differences, for instance, in the use of collocate verbs such as in *tomar clases* lit. 'take classes' used in America (see in (1)) and the combination expected by European speakers of Spanish in the same context: *ir a clase* lit. 'go to class'.

(1) Empecé *tomando clases* de una española
lit. I started by *taking classes* from a Spanish women

Finally, the limitations inherent to written text, such as missing intonation pattern, sometimes also caused difficulties in the interpretation of the text itself.

3.4 Problems of interpreting errors

Three kinds of problems constituted a challenge when labeling errors with the specific error categories. Firstly, given that the error type labels, to some extent, reflect how the erroneous expression relates to its correction, cases when more than one correction was possible resulted problematic. For instance, in the sentence in (2), the combination *hizo gorditas* can be corrected either for a collocation *ponerse gordas* lit. 'put one selves fat' or a single verb *engordar* 'gain weight'. In the first case the error should be described as the use of an incorrect collocate (*hacer* instead of *ponerse*), while in the second case, it should be described as the use of an erroneous analytical form (*hacer gorditas*) instead of a single lexical item (*engordar*).

(2) el viaje no **nos hizo gorditas*
lit. the trip didn't *make us fatty*
we didn't *gain weight* during the trip

Secondly, some incorrect collocation-like combinations produced by the learners turned out to be literal translations of combinations in the native language that have no collocation equivalent in Spanish. For instance, the erroneous form **humo de segunda mano* corresponds to the English collocation *secondhand smoke*, which can only be translated to Spanish by a complex phrase expressing the same meaning without constituting a phraseological expression: *humo del tabaco de otras personas* 'smoke from other people's cigarette'. On the contrary, some expressions used by the learners do not constitute collocations themselves; however the correct form to be used should be a collocation in Spanish. An example for this case can be seen in (3) where the expression using the copulative verb and the adjective *curioso* 'curious' should be corrected as a collocation: *tengo curiosidad* lit. 'I have curiosity'.

(3) **estoy curiosa* conocerlo
lit. *I'm curious* to get to know it

Thirdly, two coexisting category labels had to be allowed in the cases where the source of the error could not be determined unambiguously. For instance, in the case of the incorrect collocation **hice citas* lit. 'I made appointments', the annotators found it feasible to treat the error both as a direct translation from English and as a generalization error, whereby the generic verb *hacer* 'make/do' is used instead of the correct and more restricted *concertar* 'arrange'.

4. Exploiting corpus data for a learning tool

We have already mentioned that the error typology (Alonso et al. 2010) we used in the annotation task allows for a more detailed error annotation than the coarse-grained typologies used in other learning tools focusing on collocations (Chang et al. 2008; Shei and Pain 2000). In these, only lexical errors affecting the collocate are taken into account, and the main type of error foreseen is that resulting from L1 lexical transfer. With these limitations, a learning tool aimed at the automatic recognition and correction of collocation errors would have difficulties in identifying some of the error types inherent in our typology. In what follows, we will show some particular features revealed by our detailed error analysis.

4.1 The collocation error typology

Our error typology distinguishes three parallel dimensions. The first, “location” dimension captures whether the error concerns one of the elements of the collocation (the *base* or the *collocate*, following Hausmann's (1989) terminology) or the collocation as a whole. The second dimension models *descriptive* error analysis and distinguishes between three main types of error: lexical, grammatical and register error, the first two of which are further detailed in several subtypes. Finally, the third dimension represents *explanatory* error analysis: it concerns the source of the error, described by the main categories of transfer errors, that is, errors reflecting L1 interference and interlanguage errors, resulting from the incomplete knowledge of the L2 without L1 interference.

4.2 The "location" of errors

In contrast with the general approach in automatic collocation error correction, we have taken into account any error affecting either member of the collocation or the expression as a whole, as captured by the “location” dimension of our typology. As a result, we have annotated not only erroneous collocates (4), but also erroneous bases (5). The latter case would pose a difficulty for systems that correct collocation errors merely verifying the correctness of the collocate. For example, in the case of the collocation in (4), where the collocate is incorrect, a search for collocate verbs of the base *regla* 'rule', similarly to Liu's (2002) or Chang et al.'s (2008) proposal, restricted to those synonymous or sharing translation synonyms with *interrumpir* 'to interrupt', would likely return the correct combination. However, when it is the base that is erroneous, such as in (5), the same strategy, a search for co-occurring verbs with the base *gol* 'goal (in sport)' will not be effective. Note that, we have also found cases where both the collocate and the base are incorrect, as in (6).

(4) **interrumpir una regla* 'interrupt a rule' instead of *romper una regla* 'break a rule'

(5) **lograr un gol* 'achieve a goal (in sport)' instead of *lograr un objetivo* 'achieve an aim'

(6) *pasar un testemuño 'pass a testimony (from Portuguese)' instead of *dar testimonio* 'give testimony'

There is a total number of 445 erroneous collocations among the 1401 collocations annotated in the corpus. For now, we will limit our analysis to those affected by lexical errors, a total number of 266 collocations. As for the “localization” dimension, we find lexical errors of the collocate in the highest number, affecting a total of 174 collocations (61%), however a still large proportion, 61 collocations (21%) have erroneous bases, while 50 expressions (18%) contain a lexical error that is considered to affect the collocation as a whole. These numbers suggest that a CALL system aimed at correcting collocation errors efficiently, shouldn’t be limited to collocate errors, but should also foresee lexical errors concerning the base of the collocation.

Lexical errors affecting the collocation as a whole are of various kinds: an otherwise correct combination can be used in an incorrect sense, as in (7) where, in order to express the correct meaning, the combination *aliviar el estrés* ‘ease the stress’ should be substituted for *aumentar el estrés* ‘increase the stress’.

(7) al oirlo hablar, tengo que apagar al aparato para no **aliviar el estrés*
when I hear him speak, I have to turn off the television in order to not *to ease the stress*

Furthermore, we also considered here incorrect collocation-like expressions that should be correctly expressed by a single word (8), or, as we have seen above, by a non-idiomatic expression (9) and cases of incorrect single-word forms standing instead of a collocation (10).

(8) **poner apasionado* 'make passionate' instead of *apasionar* 'to fascinate'

(9) **humo de segunda mano* ‘secondhand smoke’ instead of *humo del tabaco de otras personas* ‘smoke from other people’s cigarette’

(10) **misenterpretación* 'misinterpretation' instead of *mala interpretación*

The correction of these kinds of expressions may pose further difficulties for an automatic tool.

4.3 L1 influence

Out of the 284 lexical collocation errors found in the corpus (note that a collocation can contain more than one error), 67% were found to be transfer errors, while 33% were annotated as interlanguage errors. This is in line with the findings of other authors such as Liu (2002), Nesselhauf (2005) etc. Our corpus data also corroborates the hypothesis that automatic tools such as Liu (2002), Chang et al. (2008) and Futagi (2010) make use of, that is, in most lexical collocation errors, the erroneous element can be conceived of as a synonym or a translation synonym of its correct counterpart for correction purposes. Remarkably, we find this is true both in the case of L1 transfer and interlanguage errors. Nevertheless, we would like to highlight a few error types that do not fit into this picture.

In the case of L1 transfer errors, we found the example shown in (11). We assume that the word *colegio* ‘primary school’ is used instead of *universidad* ‘university’ due to its formal resemblance to the English word *college*. This case shows that errors resulting from the phenomenon commonly known by language learners and teachers as 'false friends', that is the confusion of formally similar but semantically not necessarily related word forms, might be taken into account in language tools.

(11) Hemos **licenciado en el colegio* en la vecina ciudad
Lit. We *earned a degree in the primary school* in the neighbor town

Other phenomena concern the use of lexical elements that constitute non-words in the target language. Firstly, a small group of transfer errors (amounting to less than 6%) involve the use of a L1 lexical item, as in (12), or a lexical item from a L2 different from the target language (TL), see example (6) above. These forms are sometimes adapted to TL orthography and morphology as in (13), where the erroneous form *misenterpretaciones* stands instead of the Spanish collocation *malas interpretaciones* lit. 'wrong interpretations'. Secondly, among interlanguage errors we find cases of Spanish non-words, we assume to be the result of an erroneous derivation process. For instance, in (14) a non existent wordform **frescar* is derived instead of *refrescarse* 'cool down'.

(12) En Oaxaca se puede **ir de hiking*
Lit. In Oaxaca one can *go hiking*
(13) el trama del libro es una sarta de **misenterpretaciones*
Lit. the plot of the book is string of *misinterpretations*
(14) Las *temperaturas* cambian y **frescan* un poco
Lit. The *temperatures* change and *cool down* a bit

4.4 Grammatical errors

Learner tools aimed at the correction of collocations in general do not take grammatical errors into account at all. An exception to this is Futagi (2010) where article and inflection errors are considered, although merely with the aim of enhancing the performance of collocation extraction from learner texts. Our approach is clearly different from this, given that, from our point of view, certain grammatical errors should be considered as proper collocation errors, due to the fact that they affect the correct formulation of the lexical combination.

Grammatical collocation errors are rather frequent in the corpus, they concern 212 (44%) of the 478 erroneous collocations annotated. In what follows we show examples for each class of grammatical collocation error:

-determination error: **tomar sol* instead of *tomar* el *sol* 'to sunbathe'
-incorrect government: **montar a bicicleta* instead of *montar en bicicleta* 'to ride a bike'
-incorrect gender: **mente abierto* instead of *mente abierta* 'open mind'
-incorrect number: **estamos en vacación* instead of *estamos de vacaciones* 'we are on holiday'
-incorrect external government: **en buen humor* instead of *de buen humor* 'of good humor'
-pronominal verb error: **muero de ganas* instead of *me muero de ganas* lit. 'I am dying from desire [to do something]'
-word order error: **reputacion mala* instead of *mala reputación* 'bad reputation'

5. Conclusions and future work

The present paper has provided an insight into ongoing research focusing on the exploitation of data from learner corpus in order to enhance the performance of an automatic tool aimed at the correction of collocation errors of L2 Spanish speakers.

As we have shown, collocation annotation in corpus is not a straightforward process; the difficulties discussed in more detail are the problem of telling collocations apart from other kinds of idiomatic expressions or from free combinations, the problem of correction judgment, and the problem of assigning concrete error types. We have also demonstrated that the fine-grained typology we used for error annotations sheds lights on certain error types that are generally not taken into account by automatic error correction tools, such as errors concerning the base of the collocation, target language non-words, and grammatical collocation errors.

As for future investigation, our goal is to annotate a comparable corpus of native speakers of Spanish in order to compare the collocation knowledge and use of the native and non-native groups. We also plan to exploit our data on typical collocation errors for automatically generating activities for practicing collocations.

6. Acknowledgements

The work described in this paper has been carried out in the framework of the project COLOCATE, partially funded under the contract number FFI2008-06479-C02-01 by the Spanish Ministry of Science and Innovation (MICINN) and the FEDER Funds of the European Commission.

7. References

Alonso Ramos, M., Nishikawa, A. & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In S. Granger, M. Paquot (eds.) *Elexicography in the 21st Century: New challenges, new applications. Proceedings of eLex 2009.* Cahiers du cental 7. Louvain-la Neuve: Presses Universitaires de Louvain, pp. 369-374.

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor, G., Vázquez, N., Mosqueira, E. & Prieto, S. (2010). Towards a Motivated Annotation Schema of Collocation Erros in Learner Corpora. In N. Calzolari et al. (eds.) *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10). Paris: ELRA, pp. 3209-3214.

Benson, M., Benson, E.E. & Ilson R. (1986). *The BBI combinatory dictionary of English*. Amsterdam/Philadelphia: John Benjamins.

Chang, Y., Chang, J., Chen, H. & Liou, H. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3), pp. 283–299.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Futagi, Y. (2010). The effects of learner errors on the development of a collocation detection tool. In *AND'10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data.* New York: ACM, pp. 27-34.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A.P. Cowie (ed.) *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, pp. 145-160.

Hausmann, F.J. (1989). Le dictionnaire de collocations. In F.J. Hausmann *et al.* (eds.) *Wörterbücher – Dictionaries – Dictionnaires*, vol. 1. Berlin: de

Gruyter, pp. 1010-1019.

Higueras García, M. (2006). *Las colocaciones y su enseñanza en la clase de ELE*, Madrid, Arco Libros.

Howarth, P. (1998). The phraseology of learners' academic writing. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 161-186.

Liu, L.E. (2002). *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English*. Master's thesis, Tamkang University, Taipei.

Lozano, C**.** (2009). CEDEL2: Corpus Escrito del Español L2. In C.M. Bretones Callejas et al. (eds.) *Applied Linguistics Now: Understanding Language and Mind / La Lingüistica Aplicada Hoy: Comprendiendo el Lenguaje y la Mente.* Almería: Universidad de Almería, pp. 80-93.

Martelli, A. (2007). *Lexical Collocations in Learner English: A Corpus-Based Approach.* Alessandria: Edizioni dell' Orso.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 23-53.

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam/Philadelphia: John Benjamins.

Real Academia Española. *Corpus de referencia del español actual.* Accessed at: http://www.rae.es.

Shei, C.C., Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2), pp. 167–182.

Thewissen, J. (2008). The phraseological errors of French-, German- and Spanish-speaking EFL learners: evidence from an error-tagged learner corpus. In *Proceedings from the 8th Teaching and Language Corpora Conference (TaLC8)*. Lisbon: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, pp. 300-306.

Vincze, O., Mosqueira E. & Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In I. Boguslavsky, L. Wanner, (eds.) *Proceedings of the 5th International Conference on Meaning-Text Theory*, pp. 275-286. Available at: http://meaningtext.net/mtt2011.

DutchSemCor: Building a semantically annotated corpus for Dutch

Piek Vossen[1], Attila Görög[1], Fons Laan[2], Maarten van Gompel[4], Rubén Izquierdo[3], Antal van den Bosch[4]

[1]VU University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
[2]ISLA, University of Amsterdam, Science Park 904. 1098 XH Amsterdam, The Netherlands
[3]Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
[4]Radboud University Nijmegen, P.O. Box 9013, 6500 HD Nijmegen, The Netherlands
E-mail: p.vossen@let.vu.nl, a.gorog@let.vu.nl, fons.laan@gmail.com, proycon@anaproy.nl, r.izquierdo@uvt.nl, a.vandenbosch@let.ru.nl

Abstract

State of the art Word Sense Disambiguation (WSD) systems require large sense-tagged corpora along with lexical databases to reach satisfactory results. The number of English language resources for developed WSD increased in the past years, while most other languages are still under-resourced. The situation is no different for Dutch. In order to overcome this data bottleneck, the DutchSemCor project will deliver a Dutch corpus that is sense-tagged with senses from the Cornetto lexical database. Part of this corpus (circa 300K examples) is manually tagged. The remainder is automatically tagged using different WSD systems and validated by human annotators. The project uses existing corpora compiled in other projects; these are extended with Internet examples for word senses that are less frequent and do not (sufficiently) appear in the corpora. We report on the status of the project and the evaluations of the WSD systems with the current training data.

Keywords: Semantic annotation; Word Sense Disambiguation; Machine Learning

1. Introduction

State of the art Word Sense Disambiguation (WSD) systems require large sense-tagged corpora along with lexical databases to reach satisfactory results. While the number of English language resources annotated at the level of lexical semantics has increased in the last decade, data is still scarce for most other languages, Dutch included. In order to overcome the data bottleneck, DutchSemCor[1] is aiming to deliver a one-million word Dutch corpus that is sense-tagged with senses and domain tags from the Cornetto lexical database (Vossen 2006 and Vossen et al. 2007, 2008). The Cornetto database has over 92K lemmas and almost 120K word-senses. It includes both a wordnet and a database with lexical units which provide rich morphosyntactic, semantic and combinatoric information. Synsets in the wordnet part consist of sets of lexical units. The Dutch wordnet is linked to the Princeton WordNet (Fellbaum 1998), SUMO (Niles and Pease 2002) and Wordnet Domains (Magnini and Cavaglià 2000).

In DutchSemCor about 300K examples have so far been manually tagged by two annotators, resulting in 25 examples on average per sense. The examples mainly come from existing corpora collected in the projects CGN (Eerten, 2007), D-Coi, and SoNaR (Oostdijk et al., 2008). These corpora have already been annotated morpho-syntactically in previous projects. In some cases the annotators of DutchSemCor could not find sufficient examples in these corpora for certain word senses. A web search tool was therefore developed to find additional examples on the Dutch Internet and add these to the data. At the moment of writing, our project is entering the final phase in which the remainder of the corpus will be automatically tagged using different WSD systems. The output of the systems will be validated by human annotators through co-training. When sufficient precision is reached by the WSD systems we automatically annotate the complete corpus not yet manually annotated. In this paper, we describe the project and our approach, and we report on the results so far: both in terms of the manual annotation and the performance of the WSD systems. In Section 2, we describe the work on preparing the corpora. Section 3 describes the manual annotation protocol. In Section 4, we describe the annotation tool that was developed. Finally, in Section 5, we present two WSD systems and their estimated performance.

2. Corpus Selection and preparation

The most comprehensive corpus currently available for the Dutch language is the SoNaR corpus. SoNaR is the successor of the D-Coi Project (funded by STEVIN) and aims to contain at least 500 million words of written Dutch. This corpus was selected as the logical primary basis for DutchSemCor. The corpus is fully tokenised, part-of-speech tagged, and lemmatised. Another corpus is CGN which contains about nine million words of transcribed spontaneous Dutch adult speech.

Though SoNaR is large, it still does not contain sufficient examples for certain senses, even though the lexicographers agree it is a valid sense. For this reason,

[1] http://www2.let.vu.nl/oz/cltl/dutchsemcor/

the DutchSemCor corpus is augmented with manually selected web snippets. A special web-based tool was developed to allow for the searching of such fragments. Human annotators enter a search query and the system passes the request to a search engine (either mediated through WebCorp.co.uk[2], or directly). The results are presented on a screen and human annotators select the samples they want to annotate. After selection, snippets are automatically tokenised, part-of-speech tagged and lemmatised using Frog[3] and made available in the corpus annotation tool for assigning the sense.

The final DutchSemCor corpus will thus be a superset of SoNaR, CGN, and the manually-selected Web snippets. We integrated into the corpus representation format FoLiA (Format for Linguistic Annotation[4]) the ability to annotate lexical semantic senses, along with their annotators and confidence.

3. Manual annotation

The DutchSemCor corpus is split into two parts that are handled in different ways. The first part of about 300,000 word tokens is annotated manually in a traditional way (compare OntoNotes & SemCor): a group of 8 human annotators analyzed and tagged an average of 25 examples per sense of the 3,000 most frequent and most polysemous words of the Dutch language (65% nouns, 23% verbs and 12% adjectives). The procedure was supported by a knowledge-rich tagging system (SAT, see next section).

During manual annotation, two annotators consider the same lemmas and KWIC index examples of the reference corpus to annotate. Each tagged sentence and every annotator action is recorded in a separate database. From the database we regularly derive the annotation statistics and status (see Figure 1). The table shown in Figure 1 contains information and scores for each annotated word, such as number of annotators, number of senses, number of annotations, overlap, agreement, and proportion of annotation per sense. The total agreement/disagreement proportion per word results in the overall Inter-annotator Agreement (IA) which is our quality measure. If the IA is less than 80%, annotators examine the disagreements and improve the annotations until an IA of 80% or higher is reached.

In previous projects such as OntoNotes (Sameer and Nianwen, 2009) similar cycles have been used to reach high IA scores. To our knowledge no further criteria have been applied in these projects. Our aim is to not only obtain an IA score of 80% or higher, but also to deliver a large corpus that is sufficiently diverse in terms of syntactic and semantic patterns. We are trying to reach high diversity by implementing different filters which make use of constituency patterns, semantic roles, collocational information, and domain labels. This way we not only guarantee rich and interesting data for purposes of linguistic research but also a semantic corpus with optimal variation for machine learning. Text fragments with a large syntactic and semantic diversity can better serve WSD techniques and yield better results when used for bootstrapping.

In order to ensure an optimal coherence in the annotation we have frequent meetings with the annotation team. In these meetings we reflect on problems of different origins (possible mistakes in the lexical database, difficult sense distinctions, senses not represented in the corpus). We also discuss co-occurrence strategies to find word meanings directly in the corpus or on the Internet as well as to group examples and to discover figurative and idiomatic uses. Another purpose of the discussions is to gain insight into the peculiarities of the Dutch language and to teach annotators to validate their language instincts using different word meaning tests (e.g. zeugma, cross readings). In the initial phase, these meetings were held bi-weekly for reasons of training and tool-testing. At present, they take place once a month.

Current results of manual annotation:

- PoS: nouns, verbs and adjectives
- number of annotated lemmas: 2,589
- number of word senses: 10,172
- number of overlapping annotations[5]: 255,625
- IA[6]: 93%
- Coverage 1[7]: 77%
- Coverage 2[8]: 86%.

[2] http://www.webcorp.org.uk/

[3] http://ilk.uvt.nl/frog

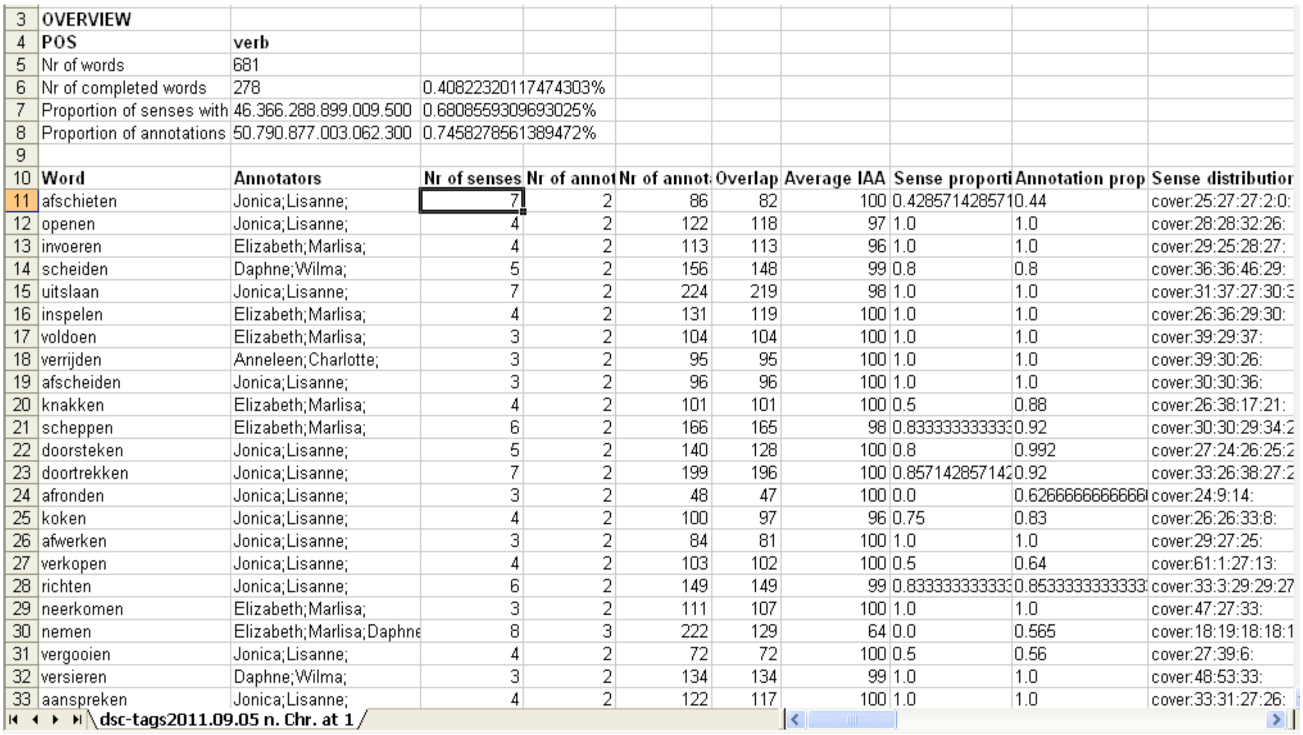
[4] http://ilk.uvt.nl/folia FoLiA is based on the D-Coi XML format, but introduces a universal paradigm allowing for various kinds of linguistic annotation; including lexical semantic sense annotation. FoLiA is also proposed as a CLARIN-NL standard in the context of the TTNWW project, and adopted in other projects as well.

[5] Tokens annotated by two annotators

[6] Inter-Annotator Agreement (also refered to as IAA)

[7] Proportion of senses with 25 or more annotations

[8] Proportion of annotations given 25 tokens per sense required



3	**OVERVIEW**									
4	**POS**	**verb**								
5	Nr of words	681								
6	Nr of completed words	278	0.40822320117474303%							
7	Proportion of senses with	46.366.288.899.009.500	0.6808559309693025%							
8	Proportion of annotations	50.790.877.003.062.300	0.7458278561389472%							
9										
10	**Word**	**Annotators**	**Nr of senses**	**Nr of annot**	**Nr of annot**	**Overlap**	**Average IAA**	**Sense proporti**	**Annotation prop**	**Sense distributior**
11	afschieten	Jonica;Lisanne;	7	2	86	82	100	0.428571428571	0.44	cover:25:27:27:2:0:
12	openen	Jonica;Lisanne;	4	2	122	118	97	1.0	1.0	cover:28:28:32:26:
13	invoeren	Elizabeth;Marlisa;	4	2	113	113	96	1.0	1.0	cover:29:25:28:27:
14	scheiden	Daphne;Wilma;	5	2	156	148	99	0.8	0.8	cover:36:36:46:29:
15	uitslaan	Jonica;Lisanne;	7	2	224	219	98	1.0	1.0	cover:31:37:27:30:3
16	inspelen	Elizabeth;Marlisa;	4	2	131	119	100	1.0	1.0	cover:26:36:29:30:
17	voldoen	Elizabeth;Marlisa;	3	2	104	104	100	1.0	1.0	cover:39:29:37:
18	verrijden	Anneleen;Charlotte;	3	2	95	95	100	1.0	1.0	cover:39:30:26:
19	afscheiden	Jonica;Lisanne;	3	2	96	96	100	1.0	1.0	cover:30:30:36:
20	knakken	Elizabeth;Marlisa;	4	2	101	101	100	0.5	0.88	cover:26:38:17:21:
21	scheppen	Elizabeth;Marlisa;	6	2	166	165	98	0.833333333333	0.92	cover:30:30:29:34:2
22	doorsteken	Jonica;Lisanne;	5	2	140	128	100	0.8	0.992	cover:27:24:26:25:2
23	doortrekken	Jonica;Lisanne;	7	2	199	196	100	0.857142857142	0.92	cover:33:26:38:27:2
24	afronden	Jonica;Lisanne;	3	2	48	47	100	0.0	0.6266666666666	cover:24:9:14:
25	koken	Jonica;Lisanne;	4	2	100	97	96	0.75	0.83	cover:26:26:33:8:
26	afwerken	Jonica;Lisanne;	3	2	84	81	100	1.0	1.0	cover:29:27:25:
27	verkopen	Jonica;Lisanne;	4	2	103	102	100	0.5	0.64	cover:61:1:27:13:
28	richten	Jonica;Lisanne;	6	2	149	149	99	0.833333333333	0.8533333333333	cover:33:3:29:29:27
29	neerkomen	Elizabeth;Marlisa;	3	2	111	107	100	1.0	1.0	cover:47:27:33:
30	nemen	Elizabeth;Marlisa;Daphne	8	3	222	129	64	0.0	0.565	cover:18:19:18:18:1
31	vergooien	Jonica;Lisanne;	4	2	72	72	100	0.5	0.56	cover:27:39:6:
32	versieren	Daphne;Wilma;	3	2	134	134	99	1.0	1.0	cover:48:53:33:
33	aanspreken	Jonica;Lisanne;	4	2	122	117	100	1.0	1.0	cover:33:31:27:26:

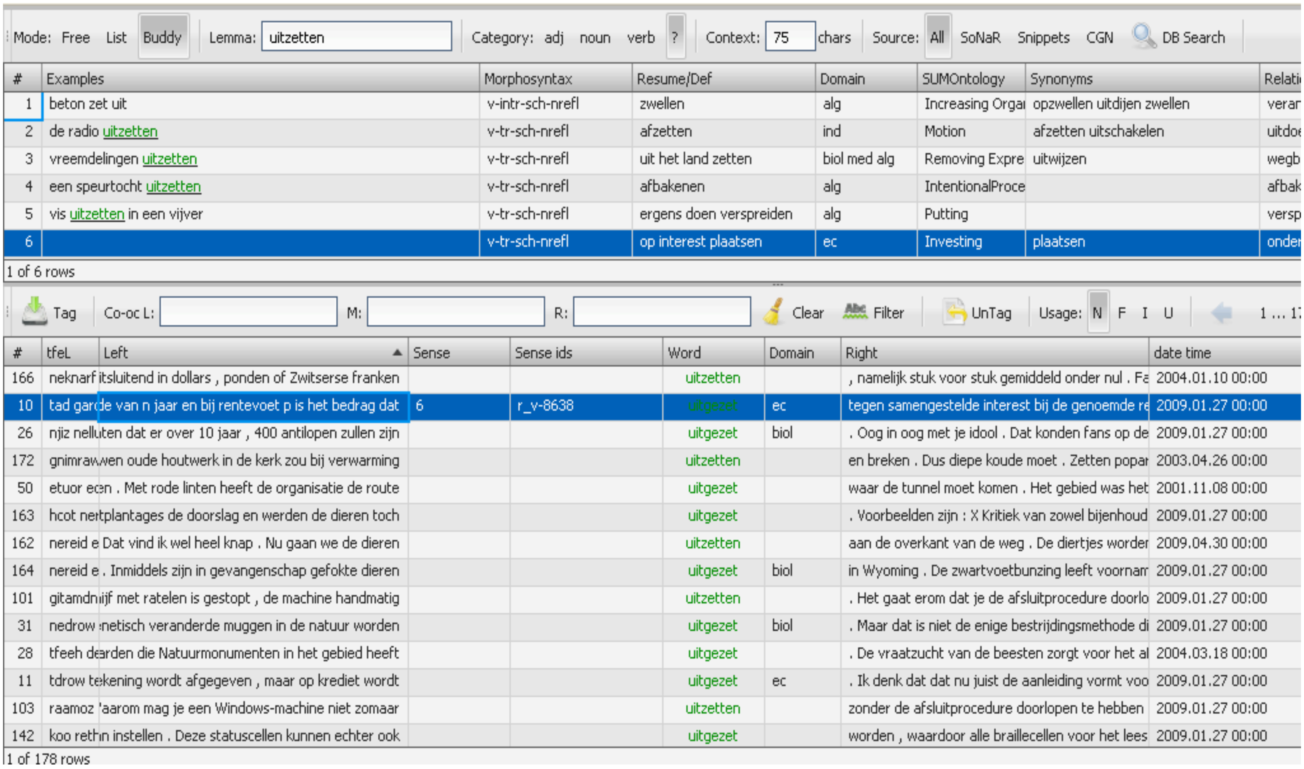
Figure 1: Logfile converted into feature table

4. Semantic Annotation Tool (SAT)

The SAT[9] is a web application for semantic tagging developed for DutchSemCor. The SAT user interface (see Figure 2) combines lexicographic information from the Cornetto database (in the top table) with corpus data from SoNaR (in the bottom table). For each lemma lexicographic and corpus data are retrieved. For each sense of the lemma the annotator selects the corpus lines that apply (the blue lines in the top and bottom tables in the screenshot). The combinations of word sense and applicable corpus lines are saved in a database, and the process is repeated until a sufficient number of instances in the corpus are annotated for each sense.

To ease the finding of required contexts, the SAT allows co-occurrence filtering of arbitrary words in the left and right context of the lemmas (Figure 3).

[9] The *Manual for Semantic Annotation* contains a more detailed description from the user perspective, together with accompanying screenshots. The SAT tool can be viewed at: http://cornetto.science.uva.nl:8080/dutchsemcor/



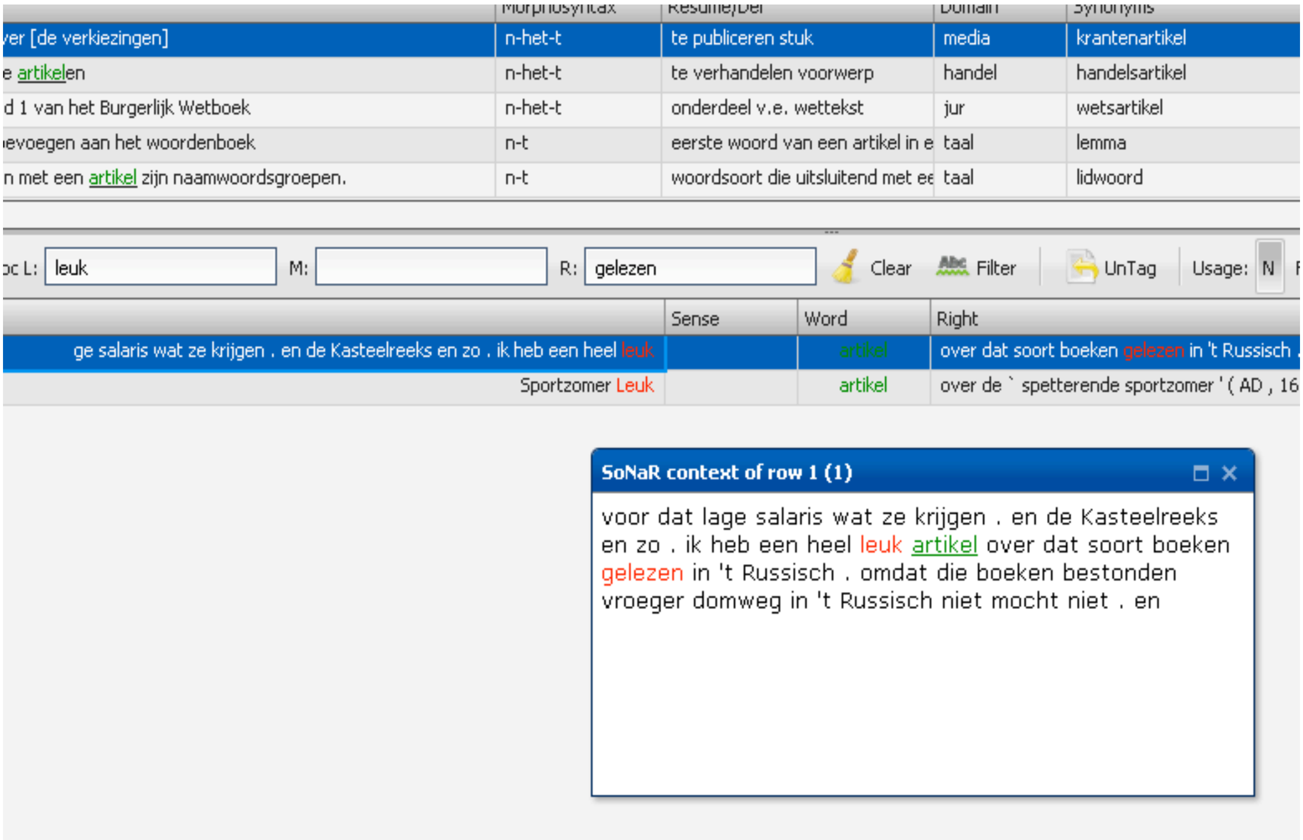
#	Examples	Morphosyntax	Resume/Def	Domain	SUMOntology	Synonyms	Relati
1	beton zet uit	v-intr-sch-nrefl	zwellen	alg	Increasing Orgai	opzwellen uitdijen zwellen	verar
2	de radio uitzetten	v-tr-sch-nrefl	afzetten	ind	Motion	afzetten uitschakelen	uitdo
3	vreemdelingen uitzetten	v-tr-sch-nrefl	uit het land zetten	biol med alg	Removing Expre	uitwijzen	wegb
4	een speurtocht uitzetten	v-tr-sch-nrefl	afbakenen	alg	IntentionalProce		afbak
5	vis uitzetten in een vijver	v-tr-sch-nrefl	ergens doen verspreiden	alg	Putting		versp
6		v-tr-sch-nrefl	op interest plaatsen	ec	Investing	plaatsen	onder

1 of 6 rows

#	tfeL	Left	Sense	Sense ids	Word	Domain	Right	date time
166	neknarf	itsluitend in dollars , ponden of Zwitserse franken			uitzetten		, namelijk stuk voor stuk gemiddeld onder nul . Fa	2004.01.10 00:00
10	tad gar	de van n jaar en bij rentevoet p is het bedrag dat	6	r_v-8638	uitgezet	ec	tegen samengestelde interest bij de genoemde re	2009.01.27 00:00
26	njiz nell	uten dat er over 10 jaar , 400 antilopen zullen zijn			uitgezet	biol	. Oog in oog met je idool . Dat konden fans op de	2009.01.27 00:00
172	gnimraw	wen oude houtwerk in de kerk zou bij verwarming			uitzetten		en breken . Dus diepe koude moet . Zetten popar	2003.04.26 00:00
50	etuor e	en . Met rode linten heeft de organisatie de route			uitgezet		waar de tunnel moet komen . Het gebied was het	2001.11.08 00:00
163	hcot ne	tplantages de doorslag en werden de dieren toch			uitgezet		. Voorbeelden zijn : X Kritiek van zowel bijenhoud	2009.01.27 00:00
162	nereid e	Dat vind ik wel heel knap . Nu gaan we de dieren			uitzetten		aan de overkant van de weg . De diertjes worder	2009.04.30 00:00
164	nereid e	. Inmiddels zijn in gevangenschap gefokte dieren			uitgezet	biol	in Wyoming . De zwartvoetbunzing leeft voornam	2009.01.27 00:00
101	gitamdn	ijf met ratelen is gestopt , de machine handmatig			uitzetten		. Het gaat erom dat je de afsluitprocedure doorlo	2009.01.27 00:00
31	nedrow	netisch veranderde muggen in de natuur worden			uitgezet	biol	. Maar dat is niet de enige bestrijdingsmethode di	2009.01.27 00:00
28	tfeeh d	arden die Natuurmonumenten in het gebied heeft			uitgezet		. De vraatzucht van de beesten zorgt voor het al	2004.03.18 00:00
11	tdrow t	ekening wordt afgegeven , maar op krediet wordt			uitgezet	ec	. Ik denk dat dat nu juist de aanleiding vormt voo	2009.01.27 00:00
103	raamoz	'aarom mag je een Windows-machine niet zomaar			uitzetten		zonder de afsluitprocedure doorlopen te hebben	2009.01.27 00:00
142	koo reth	n instellen . Deze statuscellen kunnen echter ook			uitgezet		worden , waardoor alle braillecellen voor het lees	2009.01.27 00:00

1 of 178 rows

Figure 2: SAT interface



	Morphosyntax	Resume/Def	Domain	Synonyms
ver [de verkiezingen]	n-het-t	te publiceren stuk	media	krantenartikel
e artikelen	n-het-t	te verhandelen voorwerp	handel	handelsartikel
d 1 van het Burgerlijk Wetboek	n-het-t	onderdeel v.e. wettekst	jur	wetsartikel
evoegen aan het woordenboek	n-t	eerste woord van een artikel in e	taal	lemma
n met een artikel zijn naamwoordsgroepen.	n-t	woordsoort die uitsluitend met ee	taal	lidwoord

	Sense	Word	Right
ge salaris wat ze krijgen . en de Kasteelreeks en zo . ik heb een heel leuk		artikel	over dat soort boeken gelezen in 't Russisch .
Sportzomer Leuk		artikel	over de ` spetterende sportzomer ' (AD , 16

SoNaR context of row 1 (1)

voor dat lage salaris wat ze krijgen . en de Kasteelreeks en zo . ik heb een heel leuk artikel over dat soort boeken gelezen in 't Russisch . omdat die boeken bestonden vroeger domweg in 't Russisch niet mocht niet . en

Figure 3: SAT co-occurrence filtering

5. WSD systems

Word sense disambiguation (WSD) is one of the target application areas of the DutchSemCor corpus, but it is also used for its creation. In the second phase of the project, we apply WSD methods to the corpus using the annotations that have been carried out in the first part. In fact, we apply a number of different methods:

- Knowledge-based WSD that employs the relations from the Cornetto database and in some cases from the English WordNet.
- Supervised machine learning-based WSD that creates word experts from annotated examples
- Named Entity recognition and Wikification

Named Entity recognition and Wikification are carried out independently of the Cornetto database and applied to the complete corpus. Each Named Entity will receive a link to the corresponding Wikipedia page if present. Besides representing a separate semantic annotation, the Named Entities can also be used as features for WSD.

In this paper, we focus on the first two approaches. For the knowledge-based WSD we use the UKB system that was developed by Agirre and Soroa (2009). UKB considers wordnet as a graph, where synsets are the nodes and the relations between synsets are edges. It applies a page-rank algorithm to calculate the weight for each synset (a node) in the graph. To disambiguate a new word, the personalized page-rank algorithm implemented in UKB activates the synset nodes of words that occur in the context of the focus word, and then propagates the weights from these activated synsets, resulting in a score for all the target word senses.

The supervised WSD system uses memory-based machine learning techniques implemented in TiMBL (Daelemans *et al.*, 2007) to build word experts, each responsible for disambiguating the senses of one of the designated target words in this project. The word experts base their decision on both local context, such as neighbouring words, and more global context, such as predictive words occurring in neighbouring sentences (Hoste *et al.*, 2002; Decadt *et al.*, 2004).

In the next sections, we describe both systems in more detail, and compare their performance.

5.1 UKB results

UKB requires a lexicon of lemmas with pointers to concepts and a data file with relations between concepts from which a graph is built. The Dutch lexicon contains about 84,000 lemmas that map to about 70,000 synsets. Table 1 shows the static semantic relations that have been used to build graphs for the UKB. The Dutch synset relations (DS:DS) are EuroWordNet relations (Vossen 1998). The synset-domain relations (DS:DO) originate from WordnetDomains and have been imported through the equivalence relations with the English WordNet to the Dutch synsets. We also included the domain hierarchy itself (DO:DO relations) as relations. Likewise, synsets for *tennis player* and *tennis ball* are related to the domain *tennis* but since the domain *tennis* is linked to the domain *sport*, the *tennis* synsets are indirectly related to synsets for *football player* and *football*, since the latter are related to the domain *soccer* which is also related to the domain *sport*. In case there is an equivalence relation between the Dutch synsets and the English WordNet, these are also presented as relations in the UKB (DS:ES). Finally, we have the relations from the English WordNet itself, both the direct relations (ES:ES) and the relations from a synset to the disambiguated glosses (ES:EG). In total, almost 1 million static semantic relations are available.

Type of relation	Relations
DS:DS, Dutch_synset/Dutch_synset	140,219
DO:DO, Domain/Domain	125
DS:DO, Dutch_synset/Domain	86,798
DS:ES, Dutch_synset/English_synset	73,935
ES:ES, English_synset/English_synset	252,392
ES:EG, English_synset/English_gloss_synset	419,387
	972,856

Table 1: Semantic relations used for the UKB

Many annotations of words in DutchSemCor occur in the same sentence. By assuming that these synsets are somehow semantically related, we can derive many new relations from the annotations. We extracted two sets: different polysemous words annotated in the same sentence and annotated polysemous words that co-occur with words that have a single meaning. This adds another 168K relations to the graph (see Table 2).

	Sentences	Relations	Overlap
Polysemous words	18,653	17,152	2,644
Monosemous words	189,411	151,598	3,471

Table 2: Semantic relations derived from the annotations

The co-occurrence relations hardly overlap with the relations already present in the Dutch wordnet: 2,644 polysemous word relations (15%) and 3,471 monosemous word relations (2%) were already present in the static relation set.

For determining the relevance of a relation, we cannot use the direct frequency since it is bound by the number of annotations per sense (25 on average).[10] Most relations occur only once. To still assign a weight to the extracted relations we calculated the average information value for each relation, where the information value I for a synset *s* is determined by the number of relations in which it occurs in the extracted set divided by the different synsets to which it is related:

$$I(s) = \frac{N(s)}{N(t)}$$

N(s) stands for the number of relations in which a synset *s* occurs and *N(t)* stands for the number of target synsets it is related to. We derive the average information value *AvgI* for a relation *r* as the sum of the information value of the two related synsets, divided by 2. The *AvgI* is added to the relations imported into the UKB graph. Inspection of the highest scoring relations showed many good conceptual relations. For example, we find among the polysemous words relations between *koning* (king), *koningin* (queen), *paard* (horse), *loper* (bishop), *toren* (tower), *stuk* (chess piece) and *slaan* (take a chess piece) all in their chess meaning.

We built 5 different graphs using the following relations:

- UKB1: DS:DS+DO:DO+DS:DO
- UKB2: UKB1+DS:ES
- UKB3: UKB2+ES:ES+ES:EG
- UKB4: UKB1+poly+mono
- UKB5: UKB3+poly+mono

5.2 Knowledge-based WSD results

Table 3 shows the results of evaluating the different graphs on a test set of 35,269 tokens (both nouns and verbs) extracted from the annotated data (see below for more details on the test set). The UKB has different methods for exploiting the graph. Our experiments so far showed that the personalized page rank considering each word separately (ppr_w2w setting) gave the best results.

UKB5 which uses all the relations has the best scores for both precision and recall. UKB4 comes very close, however, without using English (equivalence) relations. Actually, we see that adding the new relations derived from the annotations boosted the results with almost 9%. This suggests that the number of relations is, in fact, more important than the careful manual selection of relations. The fact that we find more syntagmatic relations in the annotations than paradigmatic relations from the wordnets is also very likely to play a role. Thus, when more data is annotated we can also increase the relations to be added and derive more statistical information on the strength of a relation (which is now limited by the maximum of 25 examples per sense).

	Precision	**Recall**	**F-measure**
UKB1	0.4557	0.4491	0.4523
UKB2	0.4557	0.4491	0.4524
UKB3	0.4560	0.4493	0.4526
UKB4	0.6360	0.6272	0.6316
UKB5	0.6411	0.6322	0.6366

Table 3: Evaluation results of the Knowledge-based WSD by UKB

Earlier versions of the Dutch UKB1 and UKB3 were evaluated in the SemEval2010 task on Domain Specific WSD (Agirre *et al.*, 2010). UKB3 performed best with a precision of 52,6%. For comparison, the English UKB scored a precision of 48,1% on the English task and ranked 10th among all participating systems. UKB3 performs 7% lower in our evaluation due to the fact that our test is more difficult: it is a sample-based evaluation for the most polysemous words only, whereas the SemEval2010 task was an all words task for a specific domain. In the latter case, there are more domain specific monosemous and low-polysemous tokens in comparison to our test.

5.3 Memory-based word experts

The supervised machine learning-based WSD system employs *k*-Nearest Neighbour classifiers (Aha et al, 1991) for word sense disambiguation. Our current approach follows previous research by Decadt et at (2004) and Hoste et al. (2002). Each classifier constitutes one word expert, and each word expert disambiguates between the senses of one of the target words selected for the project. We first illustrate the working of the system. Given the corpus and the annotated data gathered by the annotators, two datasets can be extracted: a training set and a test set. Recall that the project aims to manually annotate 25 examples per sense. Of these 25 examples, 10 are selected to be included in the test set while the remaining 15 (or more if more than 25 examples were annotated) are included in the training set. All examples that are included in either set have an IA above the predefined threshold of 80%, and the minimum number of examples per senses is satisfied for each sense of the word under consideration. Words that do not fit these requirements are ignored. These split sets are only used for evaluation purposes. When the system is run on the remainder of the corpus to automatically annotate previously unseen examples, the full 25+ examples per sense are used for training the system.

[10] Note that we will be able to extract these statistics when the complete corpus is tagged with sufficient precision by the WSD system.

To run the memory-based WSD system, training and test instances are extracted from the corpus for each word expert, each instance being one occurrence of the target word, sense annotated by a human annotator. These instances consist of a feature vector and a class label, the latter being the sense (i.e. lexical unit ID as defined in Cornetto). If no test set is used for evaluation, test instances are simply all previously unseen instances in the corpus, without associated class label, as there is no sense known prior to classification.

The feature vector consists of three components: a local context part including the word itself, a global context part and, optionally, a domain label if present. The local context part, in turn, consists of a certain number of words to the left of the word under consideration followed by the word itself and a certain number of words to the right. The context sizes, left and right, are adjustable parameters. In addition, the local context part of the feature vector may be enhanced with linguistic features; the corpus contains data on part-of-speech tags and lemmas that may be included in the feature vector for each word in context. These too are parameters to the system for which the optimal settings can only be found experimentally.

The global context part of the feature vector consists of binary bag-of-word features in which the presence or absence of important predictor words in the same sentence of the sample word is flagged. The global part refers to the fact that a certain word can be an important predictor for a given target lemma and sense and that it is computed globally over the corpus as a whole according to the method put forward by Ng et Lee (1996).

The machine learning algorithm used is implemented in the TiMBL software package (Daelemans et al., 2007) which is called by the supervised WSD system to train and test the word-expert classifiers. TiMBL is governed by several hyper parameters to tune the classifier performance. A key parameter for *k*-Nearest Neighbour classification is the value of *k*. Finding optimal parameters for a particular classifier is an experimental process in which ideally all interdependent parameter combinations are tested. In the supervised WSD system, we perform automated parameter optimisation for TiMBL on a per-classifier basis. Thus, for each word expert, prior to testing, optimal parameters are sought using a pseudo-exhaustive test of different hyper parameter setting combinations tested using leave-one-out cross-validation on the training data.

5.4 Supervised WSD-results

For the evaluation of the WSD system, we selected all words from the annotated part of the SoNaR corpus that had at least 25 agreed annotated instances per sense. We trained the word-expert for that word with all annotated instances, splitting the instances for each sense into 10 testing and at least 15 training examples.

Table 4 shows the performance in terms of token accuracy of the supervised WSD trained with different feature sets and evaluated over the test set. The training and test sets were generated from the annotated part of SoNaR at an early stage of the project, containing only 11,292 tokens. The size of the context window is shown for each type of feature as subscript. Two baselines are also included, one following a random heuristic and the other selecting the first sense based on Cornetto.

Feature set	Token accuracy
Chance Baseline	*0.2736*
First sense baseline	*0.2765*
Words$_1$	0.6287
Words$_1$ + Lemmas$_1$	0.6343
Words$_1$ + PoS$_1$	0.6307
Words$_1$ + Lemmas$_1$ + PoS$_1$	0.6333
Words$_2$	0.6511
Words$_2$ + Lemmas$_2$	0.6486
Words$_2$ + PoS$_2$	0.6393
Words$_2$ + Lemmas$_2$ + PoS$_2$	0.6409
Words$_3$	0.6606
Words$_3$ + Lemmas$_3$	0.6535
Words$_3$ + Bag-of-word	0.7212
Words$_4$	0.6551
Lemmas$_4$	0.6574
Words$_4$ + Lemmas$_4$	0.6475
Words$_5$	0.6503
Words$_6$	0.6467
Words$_7$	0.6438
Words$_8$	0.6425
Words$_9$	0.6395

Table 4: Performance of the supervised WSD

In general, the effect of considering a wider context does not have significant impact on the performance of the system. The same situation applies when enriching the set of features with part-of-speech tags and lemmas. The behaviour is not always as we would expect and the performance is not higher in all cases when a richer set of features is selected.

We also generated another training and test set in a more advanced stage of the annotation process. The number of tokens was 35,338. The polysemy distribution of the test set can be seen in figure 4. As said before, we do not consider monosemous words in our corpus. Most words in the test set have two or three senses.

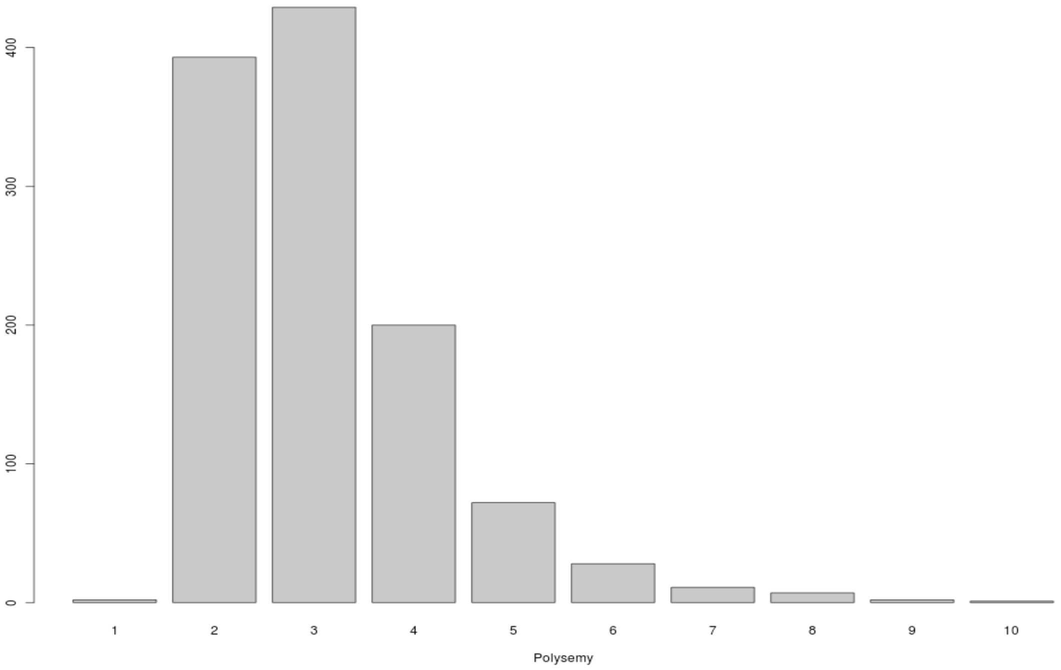


Figure 4: Polysemy of the test set

Table 5 again shows the performance of the system, using the newer data sets. In this experiment we assess the impact of the bag-of-word features and parameter optimisation.

Feature set	**Token accuracy**
Words$_1$	0.6462
Words$_1$ + Bag-of-words	0.7259
Words$_1$ + PoS$_1$ + Bag-of-words	0.7226
Words$_1$ + Bag-of-words + PS	*0.7931*

Table 5: Performance on newer data sets

As we can see, the use of bag-of-word sets leads to an important improvement of around 8% in token accuracy. On the contrary, the part-of-speech tag seems not to help the classification at all, not providing any advance. Last, we can see that using the parameter optimisation search (PS) for TiMBL, results can be improved with another 7%. The best performance 0.79 scores considerably higher than the knowledge-based UKB5 (0.64). This is also known from all earlier WSD evaluations in Senseval and Semeval. For future evaluations, we will create an all-words test set independent of the selected corpus to better test the systems independently of the corpus. We will also see how the two systems can complement each other.

Figure 5 shows the evolution of our system regarding the confidence assigned by the TiMBL engine to each token. In the standard evaluation, we considered for each test instance, the sense proposed by TiMBL, regardless the confidence assigned to it. We made an analysis of how good the confidence value was by filtering out instances with a confidence under a threshold. We expected the discarded instances to remain untagged, the recall to be lower and the precision to be higher.

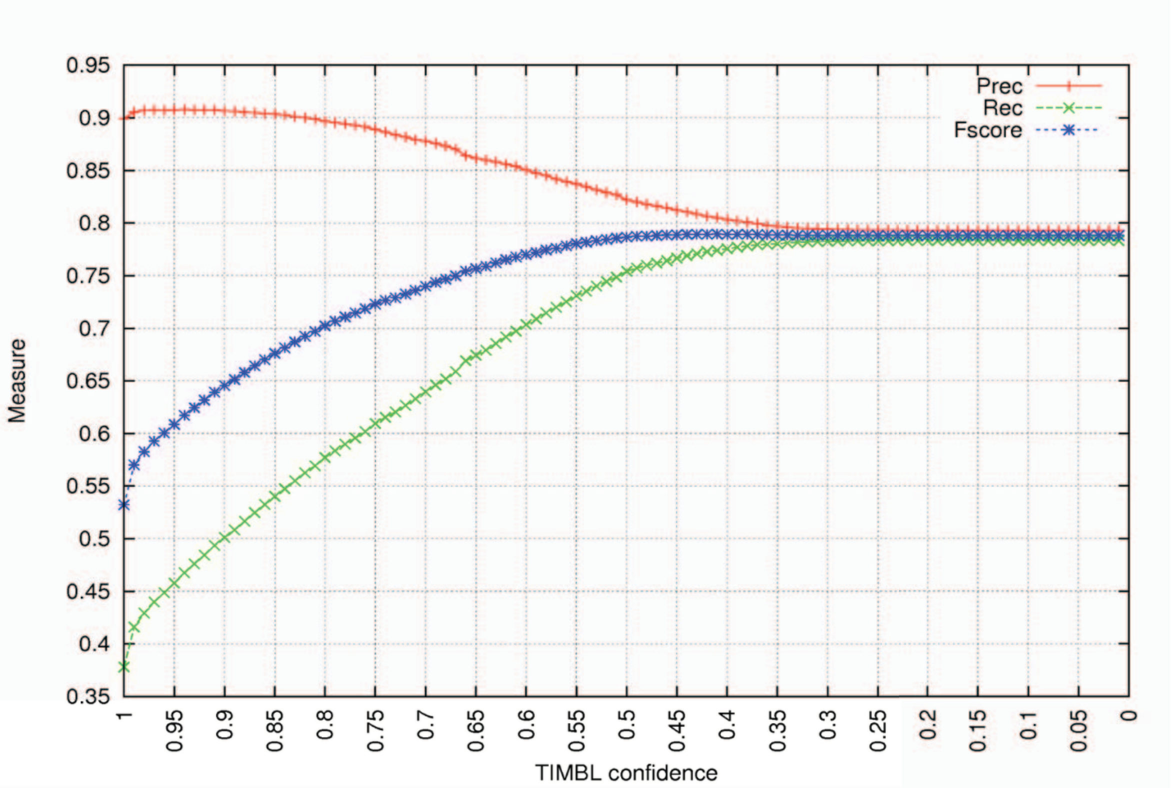


Figure 5: TiMBL confidence

The results were as expected. When we used a high TIMBL confidence value, the precision was indeed very high while the recall was hardly penalized. As we chose lower values for the confidence threshold, both values tended to be similar. It is worth mentioning that selecting a confidence of 0.55 for TIMBL results in the precision of 0.8370 (+0.439 compared with no filtering) and in an Fscore of 0.7804 (only -0.027 less than no filtering). We can filter the test instances according to the threshold of 0.55, improving the precision with 4 points and without losing too much recall.

In future experiments, it might be interesting to find out how the system performs for individual words in comparison with the its global accuracy. In spite of a good overall performance, it can be the case that the system works with high precision for certain words but reaching low results for other lemmas.

Figure 6 shows the number of words for which the system reaches a certain accuracy. Considering a quite high and reasonable minimum accuracy of 0.8, 65.54% of the nouns obtain a higher accuracy, and only the 31.21% of the verbs exceed this threshold. The manual annotation of verbs is still ongoing so these numbers are expected to increase.

5.5 Co-Training

The next phase in the project will consist of co-training. The procedure is as follows:

1. Train the WSD system with the current data (minus the test set) and determine the accuracy for each word and the F-measure for each word meaning.
2. Select which words perform with accuracy below 80% in the evaluation. This is the co-training word set *Wco.* Words that already perform well are ignored.
3. Apply the WSD systems to all occurrences of w_i element of *Wco* that have not been annotated yet.
4. We select the corpus sentences *S* in which the WSD assigned a sense *c* of w_i, such that *c* has an F-measure below 80% in the evaluation. Sentences with good performing meanings are ignored.
5. We determine a co-training score for each of sentence *s* in *S*.
6. We load the top-200 sentences into the annotation tool with the meaning assigned by the system as if it was an annotator.
7. The human annotators check the sentences assigned by the system and confirm or correct them.

8. After a week, we add the checked examples to the data to improve the WSD system and return to step 1.

The co-training score for each sentence is based on the confidence of the WSD system and the distance score of the TiMBL system. We select sentences with a high score and high distance. These are examples that are very different from the examples of the training set but for which the system nevertheless has strong evidence for the meaning. We want the students to find very different sentences for weak meanings but need to be sure that the sentences are relevant to that meaning.

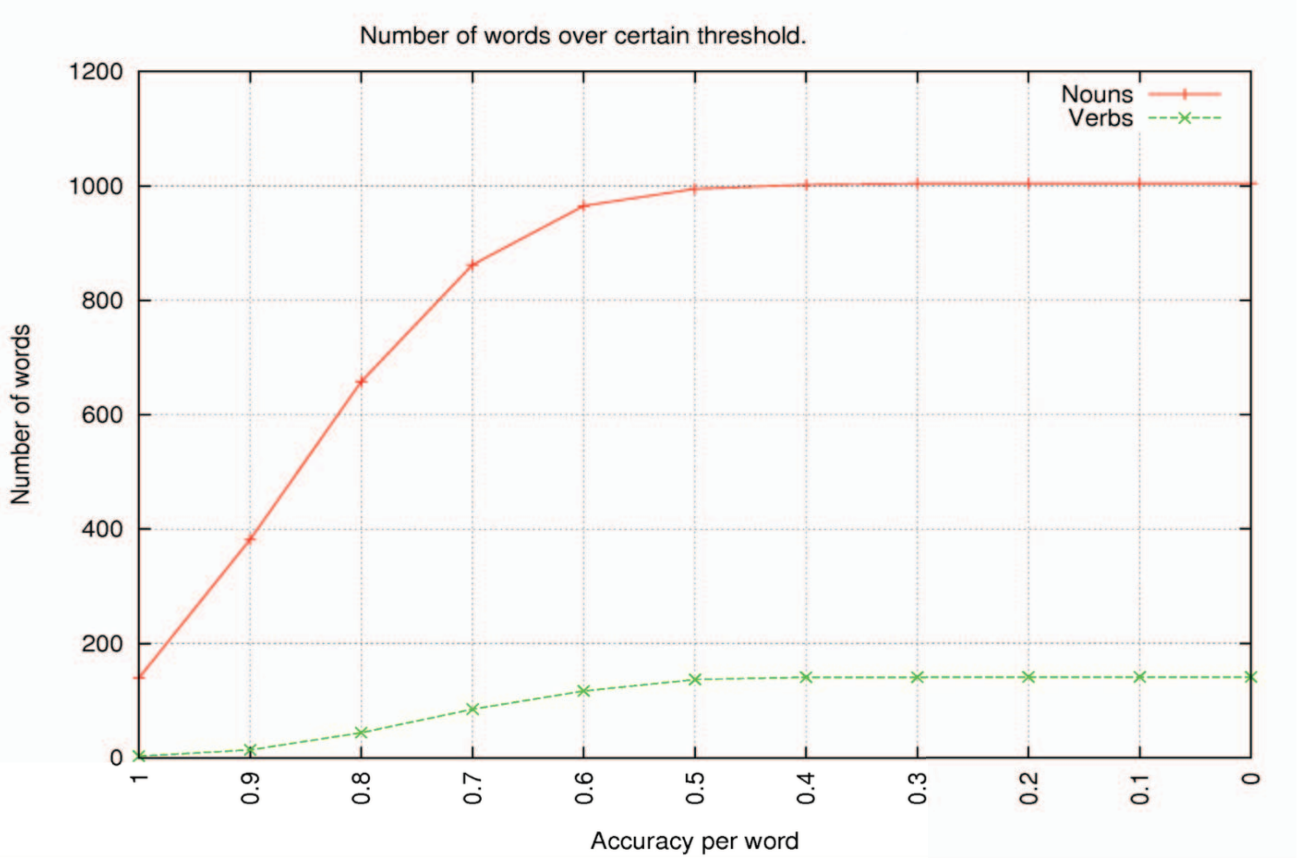


Figure 6: Number of words with a certain accuracy

We will repeat the cycles until we reach 80% accuracy for all the 3,000 words. When sufficient quality of the WSD is reached, we apply WSD to the whole corpus. The TiMBL system can only assign senses to the trained words (3,000). UKB can assign senses to all words in Cornetto. Monosemous words can simply be tagged. We will also experiment with combinations of WSD systems.

6. Conclusion

In this paper, we have described the different phases of the DutchSemCor project: from manual annotation to the use of different WSD systems. We discussed the selection and processing of the SoNaR corpus as well as the working methods and tools used throughout the manual annotation phase. We showed that the development of WSD systems is not only a goal of the project itself but is also necessary for providing correct annotations for the complete corpus. We have seen that, even though the initial results of the two WSD systems are promising, there is still ample space for fine tuning the software through experimentation. Finally, we have summarized our future plans for Co-Training, which will take place in the coming months.

7. Acknowledgements

The DutchSemCor project is an NWO[11] Humanities "Middelgroot" (Medium) Investment subsidy project with a subsidy period of September 2009 - August 2012. We would like to thank NWO for making it possible to carry out the project.

8. References

Aha, D.W., Kibler, D. & Albert, M.K. (1991). Instance-Based Learning AlTaking into account these and previous results, we select the feature set of the last experiment (Words1 + Bag-of-words + Parameter Search), as the configuration to build our first version

[11] http://www.nwo.nl

of the WSD system. Further experimentation will be carried out using this system.gorithms. *Journal of Machine Learning*, 1, pp. 37-66.

Agirre, E., Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of EACL-09*, pp. 33–41.

Agirre, E., Stevenson, M. (2006). Knowledge sources for WSD. In *Word Sense Disambiguation: Algorithms and Applications.* New York, NY: Springer, pp. 217-251.

Agirre E., Lopez de Lacalle, O., Fellbaum, C., Hsieh, S., Tesconi, M., Monachini, M., Vossen, P., & Segers, R. (2010). SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain. In K. Erk, C. Strapparava (eds.) *Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations on Kyoto's subtask WSD17: All-words Word Sense Disambiguation on a Specific Domain, workshop collocation: ACL2010*, July 11-16, 2010, Uppsala, Sweden: The Association for Computational Linguistics (ACL), pp. 75-80.

Daelemans, W., Zravel, J., van der Sloot, K. & van den Bosch, A (2007). TiMBL: Tilburg Memory Based Learner, version 6.1. Reference Guide. ILK Technical Report 07-07.

Decadt, B., Hoste, V., Daelemans, W. & van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. In R. Mihalcea, P. Edmonds (eds.) *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, Barcelona, Spain, July 2004, pp. 108-112.

Eerten, L. (2007). Over het Corpus Gesproken Nederlands. In *Nederlandse Taalkunde*, 12(3) pp. 194-215.

van Gompel M. UvT-WSD1: A cross-lingual word sense disambiguation system. In *SemEval'10: Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden, pp. 238-241.

Hoste, V., Hendrickx, I., Daelemans, W. & van Den Bosch, A. (2002). Parameter optimization for machine-learning of word sense disambiguation. *Nat. Lang. Eng.* 8(4), pp. 311-325.

Kilgarriff, A. (2006). Word senses. In *Word Sense Disambiguation: Algorithms and Applications*. New York, NY: Springer, pp. 29-46.

Magnini, B., Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In M. Gavrilidou, G. Crayannis, S. Markantonatu, S. Piperidis & G. Stainhaouer (eds.) *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*. Athens, Greece, 31 May - 2 June, 2000, pp. 1413-1418.

Mihalcea, R. (2002) Bootstrapping large sense tagged corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluations* (LREC 2002), Las Palmas, Spain, pp. 1407-1411.

Mihalcea, R. (2004). Co-training and self-training for word sense disambiguation. In *Proceedings of the 8th Conference on Computational Natural Language Learning* CoNLL, Boston, MA, pp. 33-40.

Navigli, R. (2009). Word Sense Disambiguation: a Survey. In *ACM Computing Surveys*, 41(2), ACM Press. pp. 1- 69.

Ng, H.T. (1997). Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, USA, pp. 1-7.

Ng, H.T., Lee, H.B. (1996). Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (ACL '96). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 40-47.

Oostdijk, N., Reynaert, M., Monachesi, P., van Noord, G., Ordelman, R., Schuurman, I. &. Vandeghinste, V. (2008). From D-Coi to SoNaR: A reference corpus for Dutch. In: *Proceedings on the sixth international Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Marocco.

Palmer, M., Ng, H.T. & Dang, H.T. (2006). Evaluation of WSD systems. In *Word Sense Disambiguation: Algorithms and Applications.* New York, NY: Springer, pp. 75-106.

Pianta, E., Bentivogli, L. (2003). Translation as Annotation. In *Proceedings of the AI*IA 2003 Workshop `Topics and Perspectives of Natural Language Processing in Italy`*. Pisa, Italy, pp. 40-48.

Sameer, S.P., Nianwen, X. (2009). OntoNotes: the 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*. Boulder, Colorado: Association for Computational Linguistics, pp. 57-60.

Vossen, P. (2006). Cornetto: Een lexicaal-semantische database voor taaltechnologie, *Dixit Special Issue*, Stevin.

Vossen, P., Hoffman, K., de Rijke, M., Tjong Kim Sang, E. & Deschacht, K. (2007). The Cornetto Database: Architecture and User-Scenarios. In *DIR*, pp. 89-96.

Vossen, P., Maks, I., Segers, R. & van der Vliet, H. (2008). Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In: *Proceedings on the sixth international Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco.

Correcting Errors in Digital Lexicographic Resources Using a Dictionary Manipulation Language

David Zajic*†, Michael Maxwell†, David Doermann*, Paul Rodrigues†, Michael Bloodgood†

†University of Maryland Center for Advanced Study of Language (CASL)
*University of Maryland Institute for Advanced Computer Studies (UMIACS)
College Park, MD
E-mail: dzajic@casl.umd.edu, mmaxwell@casl.umd.edu, doermann@umiacs.umd.edu, prr@umd.edu, meb@umd.edu

Abstract

We describe a paradigm for combining manual and automatic error correction of noisy structured lexicographic data. Modifications to the structure and underlying text of the lexicographic data are expressed in a simple, interpreted programming language. Dictionary Manipulation Language (DML) commands identify nodes by unique identifiers, and manipulations are performed using simple commands such as *create*, *move*, *set text*, etc. Corrected lexicons are produced by applying sequences of DML commands to the source version of the lexicon. DML commands can be written manually to repair one-off errors or generated automatically to correct recurring problems. We discuss advantages of the paradigm for the task of editing digital bilingual dictionaries.

Keywords: noisy structured data; error correction; digital lexicography

1. Introduction

Digital lexicographic resources are frequently derived from print dictionaries, either manually or automatically, or are adapted from publishers' files. Often the resulting digital lexicographic resources contain errors. Discovering and correcting errors in lexicographic data is a common task for teams dealing with digital lexicographic resources. We propose a paradigm for correcting errors and discuss its advantages for the task of editing digital bilingual dictionaries.

A digitized dictionary contains not only the underlying text, but also structural information. The underlying text is divided into meaningful spans and the spans are organized into a structure that denotes the relationships among them. Print dictionaries denote structural information with fonts, indentation, special symbols, and other visual clues. In a digitized version of a print lexicon, the structural information is made explicit. One goal for editing structured lexicographic data is to ensure that the structural information matches the semantics implicit in the layout of the source print lexicon.

The process of repairing a digital dictionary includes correcting text errors, such as typos and OCR errors, and structural errors. Structural errors happen when the underlying text is split into text spans incorrectly or when the relationships among text spans are incorrect. We refer to data containing these types of errors as noisy structured data, because our goal is to recover the true representation of the dictionary contents by correcting errors introduced by the noisy process of digitizing it.

For example, in Qureshi (1971), an Urdu to English dictionary, the translation of Urdu word "بیجو" is "goal in children's game called باؤری." In the source digitization this translation was split into "goal in children's game called" and a separate lexical entry "باؤری." The underlying text was split incorrectly into two distinct text spans.

In some cases the underlying text is divided into correct spans, but the role of the text spans is incorrect. The translation of Urdu "شاذ و نادر" is "rarely," but in the source digitization, "rarely" was identified as a usage note rather than a translation.

There are also cases in which the text is divided and tagged with the correct role, but its relationship to other text spans is incorrect. For example, in Qureshi (1971) lexical entries are organized into blocks of text containing a headword followed by collocations containing the headword. We observed that in the digitized version of the paragraph for "ٹانگ", "leg", the translation of the phrase "ٹانگ تلے سے نکلنا", "to yield, submit" was attached to the headword instead of to the phrase.

Another goal of editing lexical resources is to map resource- and language-specific structures into resource- and language-independent standards, such as Text Encoding Initiative (TEI) (Ide & Véronis, 1995) or Lexical Markup Framework (LMF) (Francopoulo et al., 2006).

2. Database and Version Control Solutions

A straightforward method of editing a structured digital resource is to store the information in a shared repository and allow experts to edit the repository contents. The repository could be a relational database or an XML document under a version control system. When the resource is in a database users modify the data through a transaction processing system. For a document under version control, users check out a working copy of the resource, edit the copy, and commit their edited copies to the repository.

```
CREATE TextElement tag text relation anchor
CREATE Element tag relation anchor
CREATE Clone source relation anchor
REMOVE Element target
REMOVE Text target
RETAG target tag
MOVE Element target relation anchor
SET Attribute target attribute value
SET Text target text
```

Figure 1: A partial list of DML commands.

Suppose a team wishes to undo a local change to the resource. In both approaches it is straightforward to restore the resource to its state at a specific time, but all changes made subsequent to that time are lost. We propose a paradigm that supports non-chronological rollback of local operations, which would allow for undo of a local edit while preserving subsequent human effort.

At some point during or after the lexicon repair process a team might wish to analyze the changes made to the resource. The transaction information stored by database or version control approaches would require substantial processing to clearly represent the changes from the initial to the final versions of the resource. Our paradigm creates an executable record of the modifications necessary to convert the source resource into the final version.

If the editing process is a *lossy* transformation of the source into a standard format, meaning that it is not possible to reconstruct the original data from the transformed data, it is desirable to preserve a copy of the original source. Our paradigm makes preservation of the original source an integral part of the editing process.

3. Dictionary Manipulation Language (DML) Paradigm

The key intuition of our paradigm for editing digital lexicographic resources is that the edits take the form of commands in DML rather than direct modifications to a shared resource. DML commands can be written manually by language experts or generated automatically by computer systems. The end-to-end process of generating a final lexical resource from the original source consists of reading the original source lexicon into computer memory, applying a sequence of DML command sets to it, and writing the result to a destination resource. The original source file is never edited directly. Instead the DML command sets are edited by language experts for unique problems based on examination of the source lexicon, or an interim state of the lexicon. DML command sets are also generated at run-time to correct repeated problems and then applied to the in-memory resource.

DML commands are applied to a lexicon by a DML interpreter program. The interpreter loads an XML lexicon into memory, reads the DML commands from one or more DML files, performs the operations denoted by the DML commands on the in-memory lexicon, and writes the result to an XML file.

```
ENTRY ID="351782">
   <FORM ID="351783">
     <ORTH ID="351784">طرفہ</ORTH>
     <PRON ID="351785">tūr'fah</PRON>
   </FORM>
   ...
   <SENSE N="3" ID="351794">
     <USG TYPE="time" ID="351795">rare</USG>
   </SENSE>
   ...
  </ENTRY>

# ABC 5/27/2011 sense tagged as usage, retag
CREATE element   TRANS   under    351794   T
RETAG   351795   TR
REMOVEattribute  351795   TIME
MOVE    element  351795   under    T

  <ENTRY ID="351782">
   <FORM ID="351783">
     <ORTH ID="351784">طرفہ</ORTH>
     <PRON ID="351785">tūr'fah</PRON>
   </FORM>
   ...
   <SENSE N="3" ID="351794">
     <TRANS ID="351794+1">
       <TR ID="351795">rare</TR>
     </TRANS>
   </SENSE>
   ...
  </ENTRY
```

Figure 2: XML excerpt from Urdu dictionary, DML commands to correct a structural error, and the result of applying the DML commands to the source XML.

The work of lexicon repair consists of writing the DML commands to correct unique problems and writing programs that search for repeated problems and generate DML commands to correct them.

3.1 DML Commands

DML commands operate on XML documents in which each element of the XML document has a unique identifier. Figure 1 shows a partial list of DML commands.

Figure 2 shows an excerpt from an XML document containing a structural error, some DML commands to correct the error, and the result of applying the DML commands to the source. In this instance, annotator ABC observed that the translation of the third sense of “طرفہ” should be “rare,” instead of the usage note that the third sense is a rare meaning. She wrote a comment about her observation and her intended solution, then solved the problem by creating a new TRANS element, changing the element tag of USG to TR, moving the TR inside the new TRANS element, and removing the TIME attribute from the TR element.

3.2 DML Processing

The end-to-end process by which a source lexicon is converted to a final lexicon consists of reading the source lexicon from an XML file into memory, applying a sequence of DML command sets to it and writing the output to an XML file. As a diagnostic option, the interim state of the lexicon can be written to an XML file after the application of any or all of the DML command sets. The architecture is shown in Figure 3.

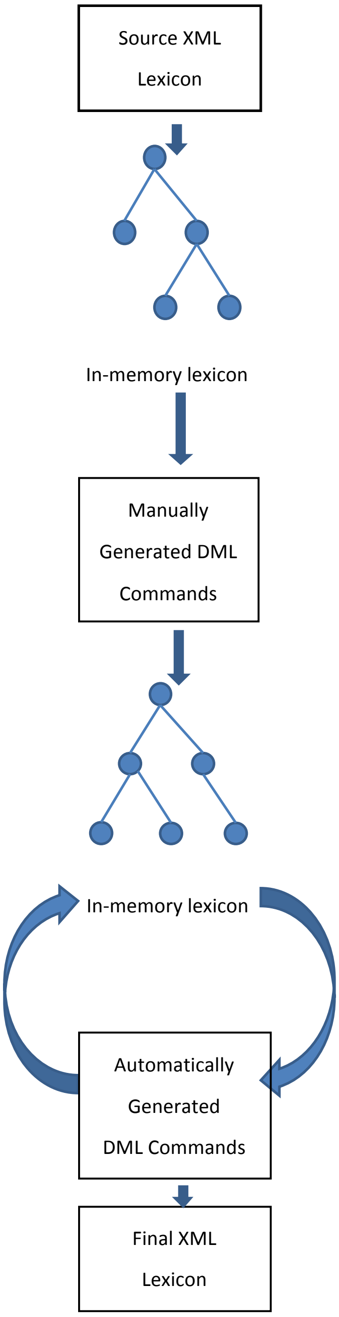


Figure 3: Architecture of DML processing.

A DML command set is a file containing DML commands that solve a collection of similar problems. A DML command set could convert the text contents of certain tags from non-standard legacy encodings to Unicode, or could relocate pieces of grammatical information that repeatedly and consistently appear in the wrong position.

Some DML command sets are written by language experts to correct unique problems while other DML command sets are generated at run time by applying patterns to the current state of the lexicon, and are then applied to the lexicon. For example, a module can search for all instances in which a word sense element has been incorrectly split from its lexical entry. Every time it finds one, the module generates DML commands to move the sense into its proper place. When the DML command set to solve this sort of problem is complete, the DML command set is applied to the in-memory lexicon. Thus we have a lexicon with the problem corrected, and we have a record of the specific changes that corrected the instances of the problem.

4. Advantages of the DML paradigm

This section will describe the advantages that motivate the use of DML for editing noisy structured lexicons.

4.1 Preservation of Source Data

The motivation for correcting errors in a lexicon is frequently to prepare the lexicon for use in a specific task. For example, a lexicon repair team might wish to load a dictionary’s contents into an enterprise-wide dictionary interface that requires a specific data format. A work paradigm that converts the source into the desired format by directly altering the source can cause the loss of valuable information. If it is later discovered that a significant error was made in generating the target lexicon, it is critical to have access the original source. Because the DML paradigm is based on application of DML commands to the source lexicon, it encourages preservation of the original source data and discourages direct editing of the source data.

4.2 Non-chronological rollback

Under the DML paradigm it is possible to undo a local change to the lexicon without affecting changes that were performed afterwards. Using a database or a version control system, it is possible to restore a resource to its state at a particular time. A change is undone by restoring the resource to a time before the change was made.

Under the DML paradigm, the source lexicon is never directly edited, so a change is undone by removing the relevant DML commands and rerunning the process to generate the final lexicon.

4.3 Effect of Application Order for DML Command Sets

DML command sets to solve repeated problems are generated at run time and are then applied to the in-memory lexicon. It is possible for application of a manually written DML command set to change some part of the lexicon so that it matches the pattern that triggers automatic DML generation. The automatically generated DML commands are not just created once and applied in that form in subsequent runs. They are recreated with every run, so they can take advantage of repairs to the lexicon made by earlier DML command sets in searching for their trigger patterns. In a similar manner, a repair to the lexicon can prevent the inappropriate application of automatically generated DML by changing a section of the lexicon so it does not match a trigger pattern. The effect of application order on DML command sets is similar to the ideas of feeding order and bleeding order of phonological rules.

The effect of ordering DML command sets can save effort for language experts. Finding a pattern in a lexicon can trigger the automatic generation of a large and complex group of DML commands. If a language expert can make a small local change in a lexicon so that so that it correctly matches the trigger pattern of a DML command generator, it is less work than manually writing the DML commands for the complex repair

4.4 DML as Documentation

The DML command sets serve as documentation of the changes that were made to the lexical resource. After an end-to-end run of a lexicon repair process, the manually and automatically generated DML command set files and the interim snapshots of the lexicon as XML files after the application of each DML command set remain as evidence of the changes that were made to the lexicon. One can examine all the changes to a local region of the lexicon by searching for element identifiers in the DML command files. Alternatively one can examine the effect of a DML command set by comparing the interim XML snapshots of the lexicon before and after the application of that set.

4.5 DML as Data

The DML command sets themselves can be analysed to better understand the process of discovering and correcting structural errors in digital lexicons. The DML command sets serve as training and evaluation data for research on machine learning systems. Our group is currently developing systems to automatically locate structural anomalies in digital lexicons (Rodrigues et al., 2011).

4.6 Support for Collaboration between Language Experts and Computer Scientists

We have found that language experts with no previous experience with computer scripting or XML data were able to learn how to write DML commands to make repairs to lexicons. This improved the workflow for finding and correcting structural errors since the same language expert who discovered an error could immediately correct it. This removed the bottleneck of creating a queue of errors and corrections to be implemented later by computer scientists. It also allowed the communication between the language experts and computer scientists on the lexicon repair team to focus on discovering and automatically correcting repeated error patterns.

5. Applications of DML

CASL's lexicon repair team has used the DML paradigm to perform structural repair on digital sources for three bilingual dictionaries: Iraqi Arabic to English (Woodhead & Beene 2003), Yemeni Arabic to English (Qafisheh 2000) and Urdu to English (Qureshi 1971). Table 1 shows the rough scope of these projects. These projects included restructuring the data to be compatible with LMF's resource- and language-independent schema for bilingual lexicons, and conversion of non-Latin text from legacy encodings to Unicode. The number of manual commands gives an idea of the scope of the human effort to correct unique textual and structural errors.

Lexicon	Entries	DML commands
Iraqi	13,719	Manual: 4759 Automatic: 1,594,688
Yemeni	16,069	Manual: 16,069 Automatic: 162,685
Urdu	44,237	Manual: 5,963 Automatic: 707,612

Table 1: Numbers of lexical entries, manually written DML commands and automatically generated DML commands for three lexicon repair projects.

6. Future Work

We have found that DML is easy to use by language experts, however it does require the overhead of using a programmer's editor and directly examining XML documents. We are developing a graphical interface that will allow language experts to view the lexicon as a tree and perform correction operations through the interface, which would generate the DML commands to implement the changes. The interface would also allow users to see the effect on the tree of applying specific DML commands and to view similar areas of the lexicon to determine if similar corrections should be applied.

7. Conclusion

We have described a paradigm for editing noisy structured lexicographic data using DML. This approach addresses the problems of non-chronological rollback and preservation of original source data. It also offers the advantages that the DML command sets serve as documentation of the corrections made to the lexicon, and be used as training and testing data in research in automatic detection of anomalies in structured lexicographic data.

8. Acknowledgements

The authors wish to thank the language experts who served as the testers and users of DML: Bridget Hirsch, Christian Hettick, Mohini Madgavkar, Valerie Novak, Rebecca McGowan, and Jennifer Boutz.

This material is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. Nothing in this report is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the authors of the product, process, or service that is the subject of this report. No one may use any information contained or based on this report in advertisements or promotional materials related to any company product, process, or service or in support of other commercial purposes.

9. References

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical Markup Framework (LMF). *International Conference on Language Resources and Evaluation – LREC 2006*, Gênes, Italy.

Ide, N., Véronis, J. (eds.) (1995). *The Text Encoding Initiative: Background and Context*. Kluwer Academic Publishers, Dordrecht (see also www.tei-c.org).

Qafisheh, H. (2000). *NTC's Yemeni Arabic – English Dictionary*. NTC Publishing Group, Chicago.

Qureshi, B.A. (1971). *Kitabistan's 20th Century Standard Dictionary Urdu into English*. Kitabistan Publishing Co., Lahore.

Rodrigues, P., Zajic, D. & Doermann, D. (2011). Detecting Structural Irregularity in Electronic Dictionaries Using Language Modeling. In I. Kosem, K. Kosem (eds.) *eLexicography in the 21st century: New applications for new users (Proceedings of eLEX2011)*, Bled, 10-12 November 2011. Ljubljana, Slovenia: Trojina.

Woodhead, D.R., Beene, W. (eds.) (2003). *A Dictionary of Iraqi Arabic: English – Arabic, Arabic – English*. Georgetown University Press, Washington, D.C.

Automatically extracted word formation products in an online dictionary

Sabina Ulsamer
Institute for German Language (IDS)
R5, 6-13; 68161 Mannheim
E-mail: ulsamer@ids-mannheim.de

Abstract

This paper presents a method of automatically extracting compounds and derivatives relating to particular base words from the headword list of the online dictionary *elexiko*, and a layout for displaying them in the entries of the base words. The aims of this project were to illustrate word formation relations and to provide a stronger connection between the headwords. The starting point for the automatic retrieval of word formation products was the morphological analysis of the headwords using a computational morphology. The analyses were stored in a database. Having obtained the base words, all compounds and derivatives relating to them were retrieved and displayed online in the entries of the base words.

Keywords: online dictionary; word formation; automation

1. Introduction

In metalexicographic literature (cf. Ulsamer (forthcoming) for an overview of the debate), the problem of illustrating word formation in dictionaries, i.e. morphological relations between dictionary entries, has been widely discussed. Mugdan (1984:274) states that compounds and derivatives form a complex net of connections between lexemes. Word formation products enrich the vocabulary by demonstrating relations between words. In a strictly alphabetical ordering of headwords (in a printed dictionary), these connections cannot emerge. With the advantage of space and hyperlinks, electronic dictionaries offer a variety of possibilities for illustrating morphological relations between headwords. Methods from computational linguistics and software tools increasingly complement work in lexicography (cf. Klosa, 2010) and provide new approaches to dictionary writing.

The German online dictionary *elexiko*[1], already combining lexicographically edited information with automatically generated information, aims to present all compounds and derivatives relating to a particular base in the entry of that base word. As part of the project 'User-adaptive access and cross-references in *elexiko* (BZV*elexiko*)[2]' at the Institute for German Language (IDS[3]), methods were developed to extract compounds and derivatives relating to a particular base word from a database of morphologically analysed headwords. The aim is to provide a better connection between the headwords. In this paper – after a brief description of the online dictionary *elexiko* and how word formation is handled there in section two – the underlying morphological analyses and retrieval methods are described in detail in section three. Section four focuses on the online display of the automatically retrieved word formation products. Section five discusses the problems and advantages of the presented method. After a summary in section six, implications for other languages and further research are considered in section seven.

2. The online dictionary elexiko

2.1 Background

elexiko is a corpus-based dictionary for contemporary German being compiled at the IDS and integrated into the lexicographic internet portal OWID[4]. *elexiko* is intended for native and non-native speakers of German, for linguists and non-linguists alike. The 300,000 headwords were extracted based on frequency from a dynamic corpus (currently consisting of 2.8 billion tokens) specifically compiled for *elexiko*. The corpus is comprised of daily and weekly newspapers and magazines from Germany, Austria, Switzerland and the former GDR. Instead of editing the entries in alphabetical order, modules of entries, defined by specific semantic, syntactic or morphological criteria, are chosen (cf. Storjohann, 2005:55-83).

2.2 Word formation in elexiko

The dictionary aims to explain the structure and word formation process of derived and compounded words in order to fulfil Barz's (2001:89) second objective, i.e. to reconstruct the recent word formation steps of a secondary word. In order to fulfil her first objective (Barz, 2001:88), that is to demonstrate the word formation activity of a primary word, all derivations and compounds containing the primary word in question should be listed in the entry of this headword.

The adjective *jugendlich* (*juvenile*) is described as an explicit derivation consisting of the base *Jugend* (*youth*), a noun, and the suffix *-lich*. In order to illustrate the word formation activity of the primary word *Jugend* *elexiko* aims to present all derivatives and compounds from the headword list in which *Jugend* appears: two derivations – *jugendlich* and *jugendhaft* – and about 400 compounds where *Jugend* is either the first constituent

[1] www.elexiko.de
[2] www.ids-mannheim.de/lexik/BZVelexiko/
[3] www.ids-mannheim.de
[4] www.owid.de

as in *Jugendbuch* ('book for adolescents') or the second constituent as in *Dorfjugend* ('young people of the village').

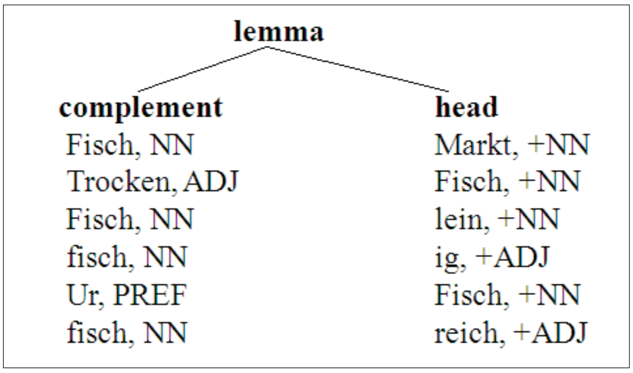
The aim of this approach is to enhance connections between words. Starting from the primary word, a net of word formations expands, offering the dictionary user a way of inferring new lexicological relations. Especially for non-edited entries, *elexiko* aims to provide more information by presenting automatically retrieved word formation products under a primary word.

3. Automatic retrieval of word formation products

3.1 Morphological analysis of headwords

As a prerequisite for the retrieval of word formation products the whole *elexiko*-headword list was morphologically analysed. The headwords were split into their constituents, and these specified morphologically. The analysis and segmentation were done with the computational morphology Morphisto.

Morphisto[5] is a computational morphology for German developed at the IDS within the TextGrid-project[6]. The tool is based on SMOR (Schmid et al., 2004). Morphisto is able to generate inflection paradigms as well as to analyse words. Additionally, Morphisto assigns a binary branching hierarchical structure to complex word formations (Zielinski et al., 2009). Morphisto was used to decompose every word of the *elexiko*-headword list into complement and head. The head denotes the grammatical and morphosyntactic category of the whole word. According to the Righthand Head Rule (Williams, 1981), the right-most constituent is the head of a German word. The complement is the accompanying constituent that specifies the head. In compounds, the second constituent is the head, while the first constituent denotes the complement. In derivatives with suffixes, the suffix functions as head, whereas the base is the complement. In prefixed words, the base is the head and the prefix the complement (cf. fig. 1).



complement	head
Fisch, NN	Markt, +NN
Trocken, ADJ	Fisch, +NN
Fisch, NN	lein, +NN
fisch, NN	ig, +ADJ
Ur, PREF	Fisch, +NN
fisch, NN	reich, +ADJ

Figure 1: Tree structure of various words

Furthermore, every complement and head was marked with a part-of-speech-tag, and the head POS-tag labelled with an additional plus sign. Suffixes are tagged with the part of speech they create. The diminutive suffix *-lein* (*-let*), which creates nouns, is labelled +NN. Prefixes were tagged PREF.

Moreover, Morphisto generated the word formation rule for every headword (cf. table 1). Word formation rules for compounds have the general form <+TAG>→<TAG> <+TAG>. A POS-tag surrounded by angle brackets is followed by an arrow which is itself followed by two further POS-tags, each in angle brackets. A nominal compound such as *Fischmarkt* (*fish market*) is described as <+NN>→<NN> <+NN>, which reads 'a noun is comprised of two nouns, where the second noun is the head of the whole construction'. In compounds, the second constituent always functions as the head which is marked by the plus sign before the second POS-tag after the arrow. The adjectival compound *fischreich* ('fish rich' – *rich in fish* [waters]) consists of a noun and an adjective: <+ADJ>→<NN> <+ADJ>. Word formation rules for derivatives can have two forms. Prefixed words are almost identical to compounds in their rule format with the general form being <+TAG>→<PREF> <+TAG>. A word like *Urfisch* (*ancient fish*) is therefore written as <+NN>→<PREF> <+NN>. Suffixed words, however, are designated with rules in the form of <+TAG>→<TAG> <SUFF>.

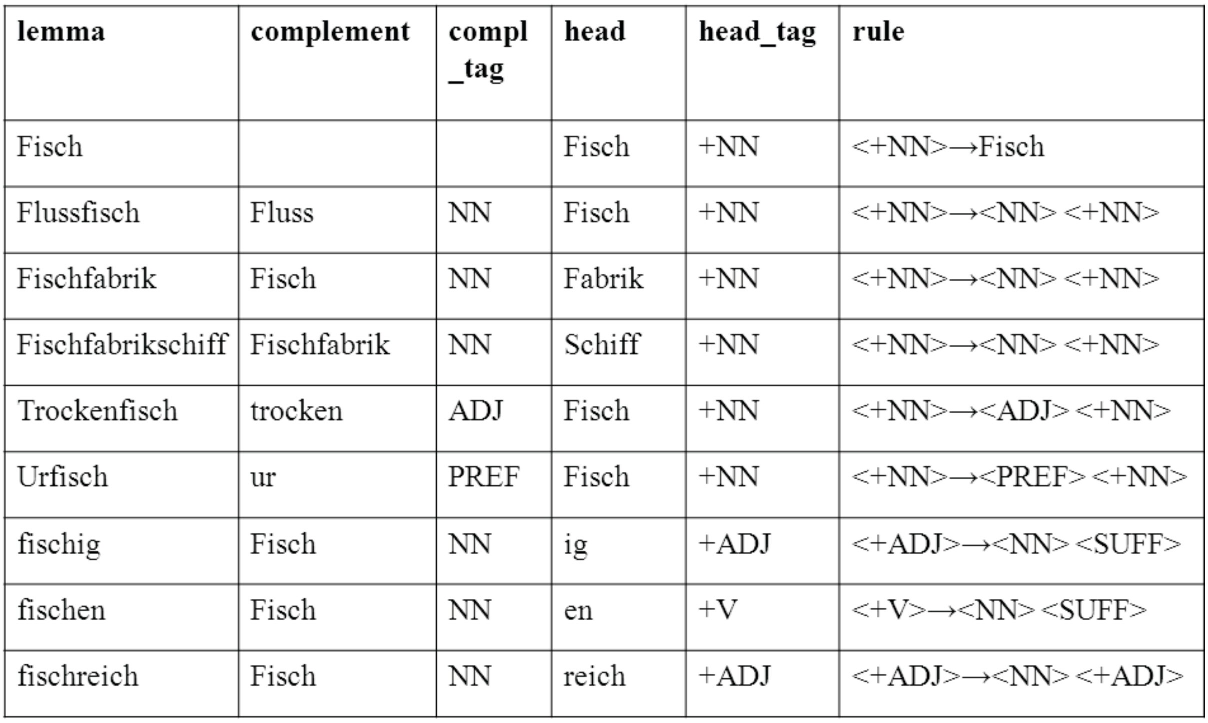
lemma	word formation rule
Fischmarkt	<+NN>→<NN> <+NN>
Trockenfisch	<+NN>→<ADJ> <+NN>
Fischlein	<+NN>→<NN> <SUFF>
fischig	<+ADJ>→<NN> <SUFF>
Urfisch	<+NN>→<PREF> <+NN>
fischreich	<+ADJ>→<NN> <+ADJ>

Table 1: A few words and their word formation rule

Simple words of any part of speech are annotated <+TAG>→lemma, e.g. <+NN>→*Fisch* (*fish*), <+ADJ>→*grün* (*green*), and <+V>→*schlafen* (*to sleep*).

The results of the morphological analysis for the entire *elexiko*-headword list were written into a relational database table (called the base table from now on) in the database management system Oracle. For each entry, the complement and the head, as well as their corresponding tags and their word formation rule, were inserted into different columns (cf. table 2). Storing the analyses in a database made it possible to search for specific patterns and retrieve the various word formation products accordingly.

[5] www.ids-mannheim.de/lexik/TextGrid/morphisto.html
[6] www.textgrid.de



lemma	**complement**	**compl _tag**	**head**	**head_tag**	**rule**
Fisch			Fisch	+NN	<+NN>→Fisch
Flussfisch	Fluss	NN	Fisch	+NN	<+NN>→<NN> <+NN>
Fischfabrik	Fisch	NN	Fabrik	+NN	<+NN>→<NN> <+NN>
Fischfabrikschiff	Fischfabrik	NN	Schiff	+NN	<+NN>→<NN> <+NN>
Trockenfisch	trocken	ADJ	Fisch	+NN	<+NN>→<ADJ> <+NN>
Urfisch	ur	PREF	Fisch	+NN	<+NN>→<PREF> <+NN>
fischig	Fisch	NN	ig	+ADJ	<+ADJ>→<NN> <SUFF>
fischen	Fisch	NN	en	+V	<+V>→<NN> <SUFF>
fischreich	Fisch	NN	reich	+ADJ	<+ADJ>→<NN> <+ADJ>

Table 2: Simplified diagram of the base table

3.2 Retrieval of the word formation products

Starting point for the retrieval of word formation products was the extraction of simple words for each part of speech. In order to retrieve compounds and derivatives containing simple nouns, these nouns had to be extracted beforehand. The aim was to create a database table in which every simple word is associated with its compounds and derivatives, so that *Fischmarkt* (*fish market*), *Flussfisch* (*river fish*), *fischen* (*to fish*), and *fischig* (*fishy*) are all related to *Fisch* (*fish*).

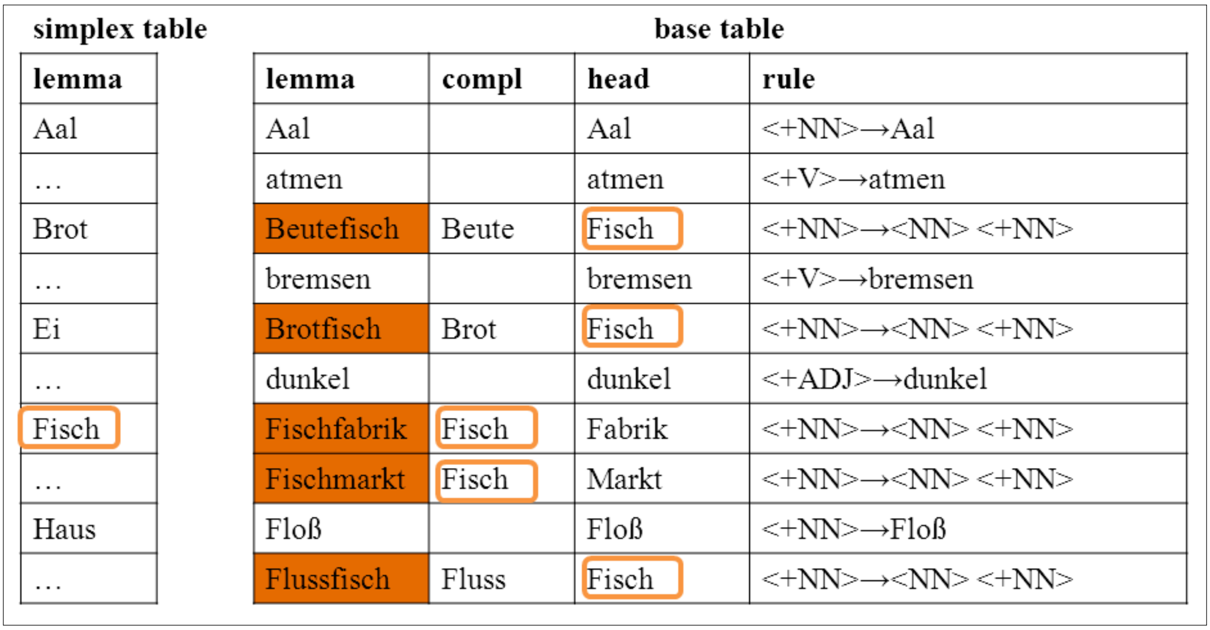
For nominal simplexes, a simplified definition of simplex was adopted, according to which all nouns are regarded as simple nouns if they cannot be segmented. Therefore, conversions such as *Tanz* (*dance* [noun]) or *Schlaf* (*sleep* [noun]) are simple nouns as well. In this paper, the retrieval of nominal simplexes and their compounds and derivatives is explained through the use of examples. The extraction of simple adjectives and verbs with their corresponding word formation products was done by analogy.

Simple words are characterised by the fact that they do not have a complement, i.e. this column is empty in the base table. Furthermore, the head equals the whole word and therefore the entries in the columns lemma and head in the base table are identical. The head POS-tag of simple nouns is +NN, and their word formation rule is always <+NN>→lemma. That is, the arrow is not followed by an angle bracket because that would indicate a derived or compounded word. The base table was queried for these conditions, and every line that fulfilled the conditions was extracted. The query resulted in about 5,600 simple nouns which were saved in a separate table, the simplex table.

For the following extraction of noun compounds that contain these simple nouns, two cases had to be considered. The simplex is either the first constituent or the second constituent. In the first case, the simple nouns from the simplex table are in the complement-column of the base table, as in the word *Fischmarkt* pertaining to the simplex *Fisch*. In the second case, the simple nouns from the simplex table are in the head column of the base table, as in *Flussfisch* to *Fisch* (cf. fig. 2).

To retrieve the compounds from the base table, a query was run in which the complement-column of the base table was compared to the lemma-column in the simplex table, and the head-column of the base table was compared to the lemma in the simplex table. Identity of the columns resulted in compounds with the simple nouns either as complement or as head. In the case of head compounds, it further had to be ensured that the entries in the columns head and lemma of the base table were not identical, because this would have resulted in simple nouns as in the query mentioned above. Additionally, the word formation rule had to be <+NN>→<NN> <+NN>.

The queries led to roughly 88,000 noun-noun compounds with one of the simple nouns as complement and about 106,000 compounds with the respective simplexes as head. Again, the results were stored in separate tables.



simplex table

lemma
Aal
…
Brot
…
Ei
…
Fisch
…
Haus
…

base table

lemma	compl	head	rule
Aal		Aal	<+NN>→Aal
atmen		atmen	<+V>→atmen
Beutefisch	Beute	Fisch	<+NN>→<NN> <+NN>
bremsen		bremsen	<+V>→bremsen
Brotfisch	Brot	Fisch	<+NN>→<NN> <+NN>
dunkel		dunkel	<+ADJ>→dunkel
Fischfabrik	Fisch	Fabrik	<+NN>→<NN> <+NN>
Fischmarkt	Fisch	Markt	<+NN>→<NN> <+NN>
Floß		Floß	<+NN>→Floß
Flussfisch	Fluss	Fisch	<+NN>→<NN> <+NN>

Figure 2: Comparison between the simplex table and the base table for retrieving the compounds relating to *Fisch*

Noun-adjective compounds (*fischreich*), adjective-noun compounds (*Trockenfisch*; 'dry fish'), verb-noun compounds (*Backfisch*; 'fried fish in batter') and the various derivatives were extracted following the same pattern. Each time, the simplex table and the base table were compared in terms of whether the simple noun was head or complement, and whether a specific word formation rule applied. Word formation products with simple adjectives or verbs were obtained likewise.

Furthermore, the corpus frequency for every product was retrieved and implemented as a column in the separate tables. All compounds and derivatives were labelled with a letter combination denoting the type of word formation and the role the simplex plays in that word formation. The nominal compound *Fischmarkt* is assigned the label 'compound-c-nn' when associated with the simplex *Fisch* (*Fisch* is the **c**omplement) but 'compound-h-nn' when associated with the simplex *Markt* (*Markt* is the **h**ead). The adjective *fischig* derived from the noun *Fisch* and is therefore labelled 'deriv-c-nadj'.

Finally, data from the various tables containing the different word formation products were put together and inserted into an overall table (cf. table 3). In this table, every simple noun, adjective and verb is associated with the compounds and derivatives in which the simple lemma appears. In this way, the aim of establishing a table containing all of *elexiko*'s word formation products relating to a particular simple headword was achieved.

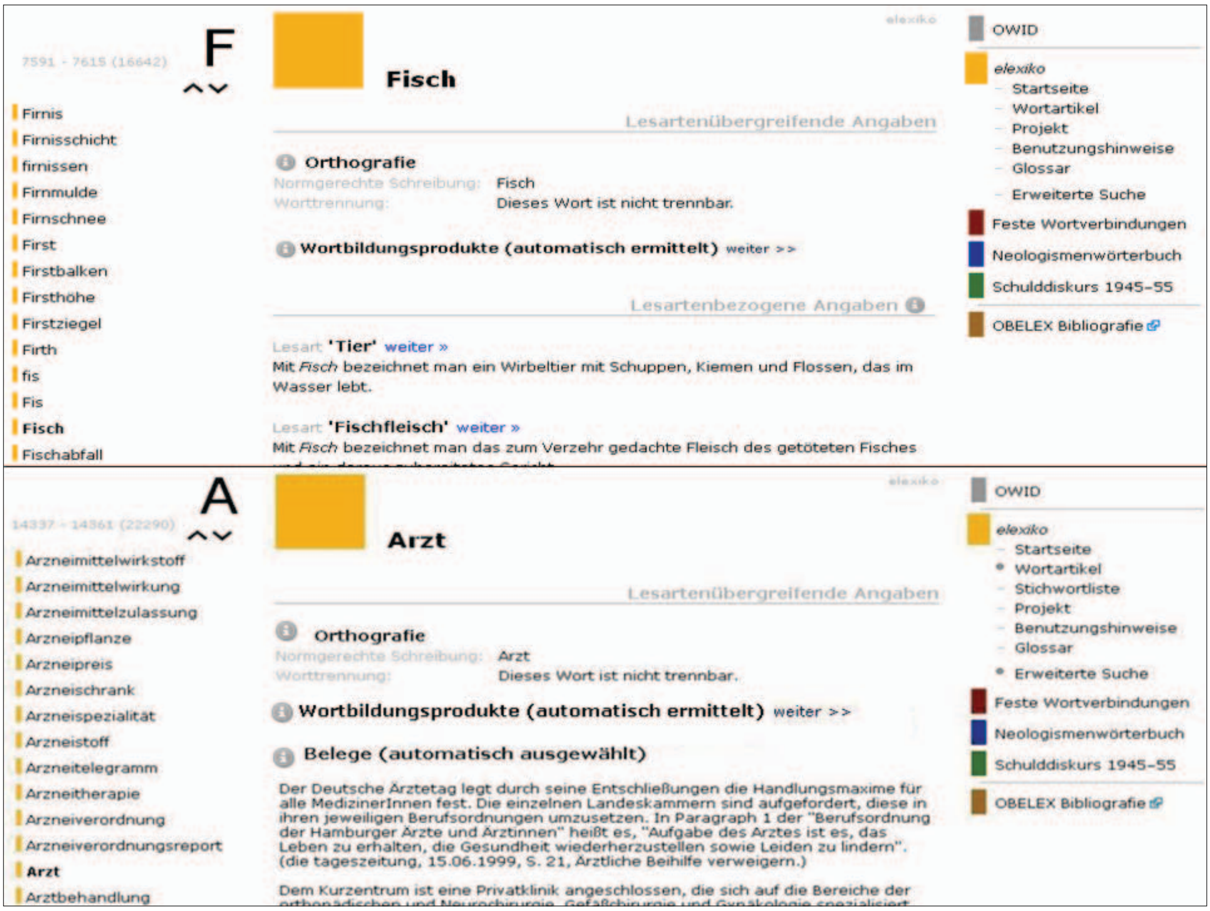
lemma	product	product_frequency	word formation role
Fisch	Anglerfisch	6	compound-h-nn
Fisch	Aquarienfisch	30	compound-h-nn
Fisch	Beutefisch	23	compound-h-nn
Fisch	…	…	…
Fisch	Fischabfall	32	compound-c-nn
Fisch	Fischadler	93	compound-c-nn
Fisch	…	…	…
Fisch	fischen	3670	deriv-c-nv
Fisch	fischig	49	deriv-c-nadj
Fisch	Fischindustrie	68	compound-c-nn
Fisch	…	…	…

Table 3: An extract from the overall word formations' table exhibiting the word formations associated with *Fisch*

4. Presenting the automatically extracted word formation products online

An *elexiko* article is divided into two parts: sense-independent information and sense-related information. The sense-independent information concerns the lexeme itself and focuses on information applying to the entire entry (Storjohann, 2005:62), i.e. details on spelling, spelling variation, and syllabification, among other things. Sense-related information provides information on a specific sense of the headword.

Due to the fact that automatic retrieval does not allow the association of the various word formation products with the specific sense of their base (e.g. the simple noun *Zug* has about nine senses), the compounds and derivatives have to be arranged under the sense-independent information. A link was placed reading 'Wortbildungsprodukte (automatisch ermittelt) weiter »' ('word formation products (automatically retrieved) more »') on the initial page of a headword under the sense-independent information (cf. fig. 3).



Fisch

Lesartenübergreifende Angaben

Orthografie

Normgerechte Schreibung: Fisch

Worttrennung: Dieses Wort ist nicht trennbar.

Wortbildungsprodukte (automatisch ermittelt) weiter >>

Lesartenbezogene Angaben

Lesart 'Tier' weiter »

Mit *Fisch* bezeichnet man ein Wirbeltier mit Schuppen, Kiemen und Flossen, das im Wasser lebt.

Lesart 'Fischfleisch' weiter »

Mit *Fisch* bezeichnet man das zum Verzehr gedachte Fleisch des getöteten Fisches

Arzt

Lesartenübergreifende Angaben

Orthografie

Normgerechte Schreibung: Arzt

Worttrennung: Dieses Wort ist nicht trennbar.

Wortbildungsprodukte (automatisch ermittelt) weiter >>

Belege (automatisch ausgewählt)

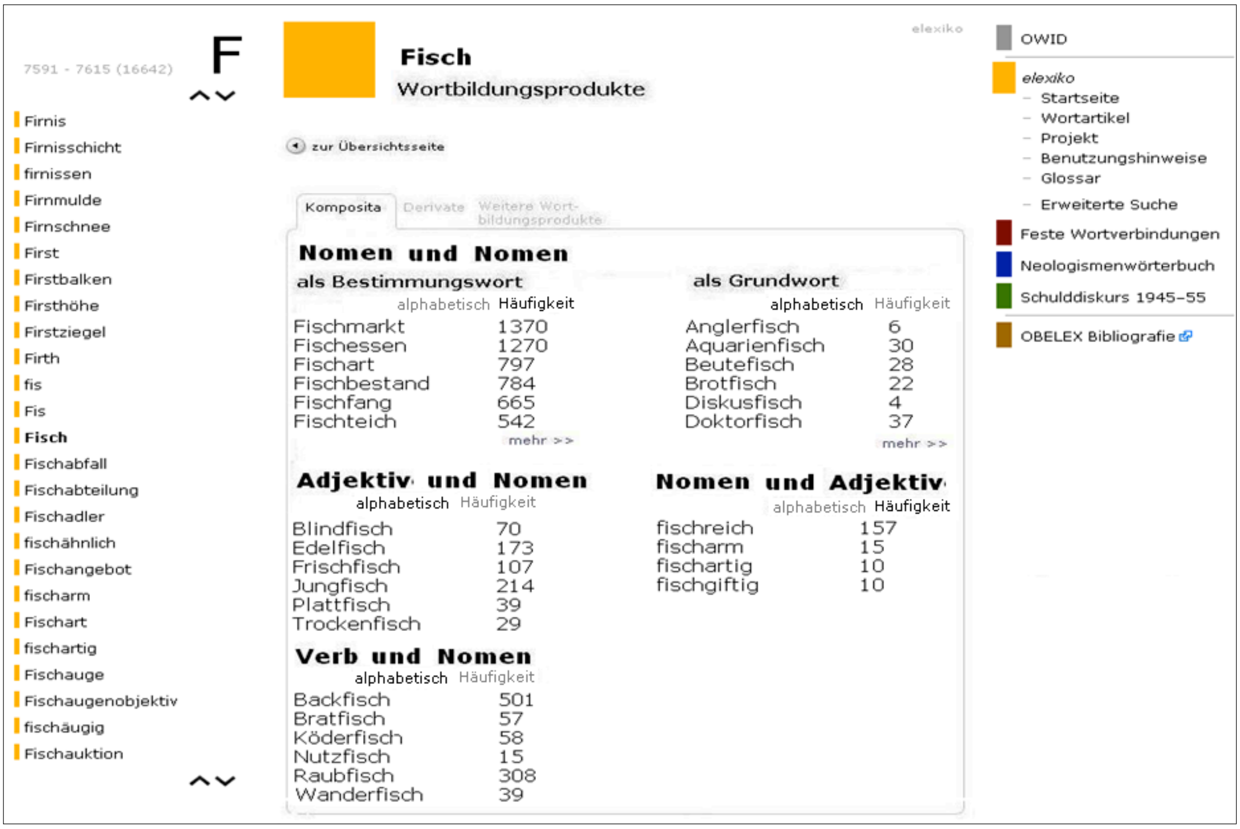
Der Deutsche Ärztetag legt durch seine Entschließungen die Handlungsmaxime für alle MedizinerInnen fest. Die einzelnen Landeskammern sind aufgefordert, diese in ihren jeweiligen Berufsordnungen umzusetzen. In Paragraph 1 der "Berufsordnung der Hamburger Ärzte und Ärztinnen" heißt es, "Aufgabe des Arztes ist es, das Leben zu erhalten, die Gesundheit wiederherzustellen sowie Leiden zu lindern". (die tageszeitung, 15.06.1999, S. 21, Ärztliche Beihilfe verweigern.)

Dem Kurzentrum ist eine Privatklinik angeschlossen, die sich auf die Bereiche der

Figure 3: Start pages of the edited entry *Fisch* (above) and of the unedited entry *Arzt* (*doctor*; below) with the link to the automatically retrieved word formation products

If the user clicks on 'more', a new view opens where the automatically extracted compounds and derivatives are presented in a tabbed pane with one tab for compounds, one for derivatives and one for further word formation products. Inside each tab, the word formation products are grouped into the resulting word class or according to the words they consist of. For simple nouns, the compounds are sorted into the categories noun-noun compounds, noun-adjective compounds, adjective-noun compounds and verb-noun compounds (cf. fig. 4). The noun-noun compounds are presented in two columns. In the left-hand column, the compounds with the simple noun as complement are listed, and in the right-hand column are the compounds with the simple noun as head.

In the derivations tab, the various derivatives are grouped according to their parts of speech, as illustrated in fig. 5. The nouns *Kinderei*, *Kindheit*, *Kindschaft* as nominal derivatives to the simple noun *Kind* (*child*) are classified under nouns. As with compounds, they are divided into prefixed and suffixed words. The same applies for adjectives deriving from adjectives (*ungut* – 'not good' from *gut* – *good*; *dümmlich* – 'slightly silly' from *dumm* – *silly*).



Fisch

Wortbildungsprodukte

Nomen und Nomen

als Bestimmungswort

alphabetisch	Häufigkeit
Fischmarkt	1370
Fischessen	1270
Fischart	797
Fischbestand	784
Fischfang	665
Fischteich	542

als Grundwort

alphabetisch	Häufigkeit
Anglerfisch	6
Aquarienfisch	30
Beutefisch	28
Brotfisch	22
Diskusfisch	4
Doktorfisch	37

Adjektiv und Nomen

alphabetisch	Häufigkeit
Blindfisch	70
Edelfisch	173
Frischfisch	107
Jungfisch	214
Plattfisch	39
Trockenfisch	29

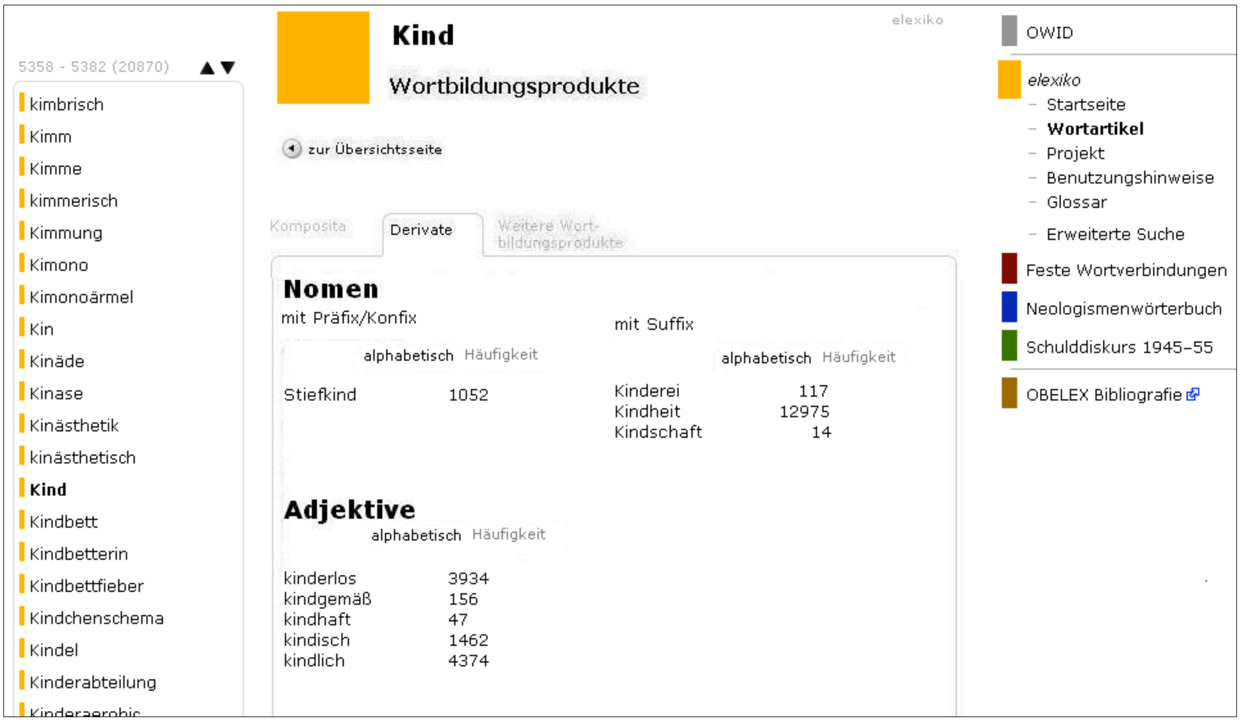
Nomen und Adjektiv

alphabetisch	Häufigkeit
fischreich	157
fischarm	15
fischartig	10
fischgiftig	10

Verb und Nomen

alphabetisch	Häufigkeit
Backfisch	501
Bratfisch	57
Köderfisch	58
Nutzfisch	15
Raubfisch	308
Wanderfisch	39

Figure 4: View of the tabbed pane for word formation products relating to the lemma *Fisch* with the tab ‘compounds’ opened



Kind

Wortbildungsprodukte

Nomen

mit Präfix/Konfix

alphabetisch	Häufigkeit
Stiefkind	1052

mit Suffix

alphabetisch	Häufigkeit
Kinderei	117
Kindheit	12975
Kindschaft	14

Adjektive

alphabetisch	Häufigkeit
kinderlos	3934
kindgemäß	156
kindhaft	47
kindisch	1462
kindlich	4374

Figure 5: View of the tabbed pane for word formation products relating to the lemma *Kind* with the tab ‘derivatives’ opened

The frequency of every word formation product in the *elexiko*-corpus is given. The compounds and derivatives can then be ordered either alphabetically or by frequency. The default presentation is alphabetical order because a usage study indicated a preference for this ordering (cf. Klosa/Koplenig/Töpel, forthcoming).

It must be pointed out that the listed compounds and derivatives are not 'nested' under the entry of the simple word. They do have their own unique address to which each of the word formation products presented in the simplex entry are linked and where further information is given. In the presentation chosen for *elexiko* and presented here, the hyperlinked word formation products are grouped according to morphological criteria. The compounds and derivatives are not nested, but, according to Engelberg & Lemnitzer (2009:150), are groups of reference lemmas organized by word formation.

5. Problems and Advantages of automatically retrieved word formation products

5.1 Problems

In order to get an idea of the percentage of errors, a sample of 88 simple nouns was taken and the number of their relating compounds counted. The simple nouns, ranging from very low to very high corpus frequency, led to 6,652 noun-noun compounds containing one of the simple nouns as complement or as head. On average a simple noun is either complement or head in 76 noun-noun compounds. About 792 of the retrieved 6,652 compounds were erroneous, i.e. almost 12% in total or 9 falsely analysed compounds per simple noun.

It became clear that several problems arise from false or missing morphological analyses. Particularly foreign words such as *Toxoplasmose* (*toxoplasmosis*) or very complex words such as *Staatssekretärsrunde* ('meeting of the secretaries of state') were not analysed by Morphisto and therefore not split into complement and head and assigned a word formation rule. Structurally ambiguous words led to false segmentations. A word like *Konzertsaal* (*concert hall*) is a compound consisting of the complement *Konzert* (*concert*) and the head *Saal* (*hall*). However, the word was decomposed into *Konzert*, linking element *-s* and head *Aal* (*eel*); from a structural point of view a possible analysis.

Wrong or even missing segmentations will lead to false results in the retrieval of word formation products and then later to problems in the presentation as well. If *Konzertsaal* is split into *Konzert*, *-s*, and *Aal*, it cannot be retrieved as a compound to *Saal* and therefore it will not be listed under the noun-noun compounds of the simplex *Saal*. Instead, it will falsely occur under *Aal*. In order to prevent false compounds and derivatives being presented in the entry of the simple headword, several corrections had to be performed. As far as possible, groups of words containing similar errors were extracted and their segmentations and word formation rules corrected manually over a period of 14 months by two student assistants. Only a small amount of words could be corrected by executing an overall update statement on the base table.

Almost 15% of 2,762 noun-adjective compounds were analysed wrongly, mainly due to false segmentations. After correction, the number of noun-adjective compounds was reduced to almost 2,550. From 5,866 adjective-noun compounds, 19% were decomposed into false constituents or provided with a false word formation rule. False segmentations mostly occurred with prefixes such as *haupt-* (*main*) or *vorder-* (*front*), incorrectly analysed as adjectives. Another source of errors was homographic constituents. After correction, approximately 4,700 adjective-noun compounds remained. 22% of 8,594 compounds of a verb and noun exhibited incorrect segmentations. Homographs were responsible for most of the errors here as well. Fortunately, almost 1,000 of those could have been corrected automatically. However, at the time of writing approximately 7,600 remained to be checked manually. With 38 false segmentations per 100 words, the 1,262 simple adjectives were the most erroneous class of words. The main source of errors was complex adjectives being falsely analysed as simple adjectives. The number of simple adjectives was reduced to 940 after being corrected and excluding participles. Still, a certain amount of errors remains.

A further problem concerns sense disambiguation. Due to the fact that Morphisto merely decomposes compounds and derivatives into their constituents regardless of their respective senses, it is not possible to relate compounds and derivatives to a specific sense of their base word. The highly polysemous noun *Zug* (*drag*, *draught*, *draw*, *flue*, *move*, *puff*, *stroke* [sports], *platoon*, *train*) appears in compounds such as *Zugbrücke* (*draw bridge*), *Schwimmzug* (*swim stroke*) or *Zugabteil* (*train compartment*). However, the compounds will all be listed under *Zug*, leaving the dictionary user with the task of relating them to the different senses.

5.2 Advantages

Despite the difficulties outlined above, the advantages of an automatic morphological analysis and an automatic retrieval of word formation products are clear. The tool Morphisto made it possible to have a list of 300,000 headwords morphologically analysed, a task that could not have been done manually in a reasonable amount of time, considering that it took two student assistants 14 months to correct only about 2,000 words by hand.

Storing all word formation products in a relational database table as explicated above (cf section 3.2) proved suitable for the extraction of the various compounds and derivatives. With a single query it is possible to retrieve all word formation products relating to a particular headword. Even lemmas that are subject to umlaut during word formation are associated with their respective bases, e.g. the adjective *ärztlich* (*medical*) to its base noun *Arzt* (*doctor*) from which it derives. A simple string-based query would not have found the umlauted case.

Presenting the compounds and derivatives that are associated with a particular simple headword in the entry of the simple word puts the simplex at the centre of a net of morphologically related words, illustrating the simple word's word formation activity. Corpus frequencies show how often the simple word forms the base of particular compounds and derivatives. In the case of compounds frequency information additionally offers an insight into the discourses in which the simple word mainly appears.

Links to the compounds' and derivatives' own entries provide the dictionary user with elaborate information on meaning, typical usage, paradigmatic relations and grammar (in the case of edited entries). On the one hand, the linkage serves to elucidate the morphological relations between the base word and its word formation products. On the other hand, the linkage ensures that the various word formation products presented in the entry of the base word are not left without explanation. Dictionary users, especially learners, are led directly to the answers to their specific questions regarding gender, inflection class and morphological make-up of words.

6. Conclusion

This paper gave an example of how automatic methods can accompany and complement lexicographic work and enhance dictionary writing.

In order to illustrate word formation relations in the online dictionary *elexiko*, each word from the headword list was analysed with a computational morphology. Every word was decomposed into its constituents, and its individual word formation rule given. The results of the analyses were stored in a relational database. The aim was to extract all compounds and derivatives relating to a particular (simple) base word, and to integrate the retrieved word formation products online into the entries of the simple word in question.

In spite of problems resulting from the automatic morphological analysis and automatic retrieval, the presented method and online presentation give the opportunity to interconnect the dictionary data. The display of all compounds and derivatives in which a particular lemma appears not only reflects the lemma's word formation potential, but also reveals relations between the lemma in question and other headwords that do not and cannot emerge in a printed dictionary. An online dictionary such as *elexiko* is able to connect the headwords on the basis of word formation so that a net of morphologically related words spans the dictionary.

7. Looking ahead

This final section aims to look ahead and to consider implications for other languages or suggest research ideas.

Although the proposed model is intended for dictionary users, researchers will find the number and variety of word formation products displayed along with the information on their corpus frequency helpful for studying the productivity of individual constituents. One might ask whether there is a correlation between frequency of base word and number of word formation products in total, or between polysemy of base word and number of word formation products. Further research questions might include to what extent a particular base word appears in compounds or derivatives, and in what word classes these compounds and derivatives are. Additionally, one might ask whether the base word has a propensity to combine with a certain word class, or whether it combines with words of a particular semantic field. The automatically retrieved word formation products offer initial answers to these questions.

The extraction of word formation products depends primarily on the morphological analysis with a computational morphology. The method presented in this paper was based on a computational morphology for German. Given a language-specific computational morphology the procedure outlined here should also work for languages in which compounds and/or derivatives not only exist but can also be decomposed into a binary complement-head-structure, e.g. Turkish *elma suyu* ('apple juice-suffix' *apple juice*). For languages that use different strategies to express determinatives such as *apple juice*, the method presented here cannot simply be adopted as it is. The whole model depends on the computational morphology which has to be language-specific. In order to extract the French *jus de pommes* ('juice of apples') as a kind of word formation related to either *jus* (*juice*) or *pomme* (*apple*), a computational morphology cannot simply deconstruct the phrase into a first constituent and a second constituent, which is not even the head. As a result of the structural difference, the database design might have to vary, and retrieval methods be adjusted.

8. Acknowledgements

This project is financed by the Gottfried Wilhelm Leibniz Scientific Community ('Joint Initiative for Research and Innovation').

9. References

Barz, I. (2001). Wortbildungsbeziehungen im einsprachigen Bedeutungswörterbuch. In J. Korhonen (ed.) *Von der mono- zur bilingualen Lexikografie für das Deutsche*. Frankfurt/M.: Peter Lang, pp. 85-100.

Engelberg, S., Lemnitzer, L. (2009). Lexikographie und Wörterbuchbenutzung. 4th revised and enlarged edition. Tübingen: Stauffenburg.

Klosa, A., Koplenig, A. & Töpel, A. (forthcoming). Benutzerwünsche und Benutzermeinungen zu dem monolingualen deutschen Onlinewörterbuch *elexiko*. In C. Müller-Spitzer (ed.) *Using Online Dictionaries*.

Klosa, A. (2010). On the combination of automated information and lexicographically interpreted information in two German online dictionaries. In S. Granger, M. Paquot (ed.) *eLexicography in the 21st century. New challenges, new applications. Proceedings of eLex 2009*. Louvain-la-Neuve, pp. 157-163.

Mugdan, J. (1984). Grammatik im Wörterbuch: Wortbildung. In H.E. Wiegand (ed.) *Studien zur neuhochdeutschen Lexikographie IV*. (Germanistische Linguistik 1-3/83). Hildesheim, New York: Georg Olms. pp. 237-308.

Schmid, H., Fitschen, A. & Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, pp. 1263-1266.

Storjohann, P. (2005). *elexiko*: A Corpus-Based Monolingual German Dictionary. *Journal of Linguistics* 34, pp. 55-82.

Ulsamer, S. (forthcoming). Wortbildung in Wörterbüchern – Zwischen Anspruch und Wirklichkeit. In A. Klosa (ed.) *Wortbildung im elektronischen Wörterbuch*. Tübingen: Narr.

Williams, E. (1981). On the Notions 'Lexically Related' and 'Head of a Word'. *Linguistic Inquiry* 12(2), pp. 245-274.

Zielinski, A., Simon, C. & Wittl, T. (2009). Morphisto: Service-oriented open source morphology for German. In C. Mahlow, M. Piotrowski (ed.) *State of the Art in Computational Morphology Workshop on Systems and Frameworks for Computational Morphology SFCM 2009, Zurich, Switzerland, September 4, 2009. Proceedings* Berlin, Heidelberg: Springer, pp. 64-75.

BZV*elexiko* – Benutzeradaptive Zugänge und Vernetzungen in *elexiko*. Accessed at: www.ids-mannheim.de/lexik/BZVelexiko/.

elexiko (2003 ff.). In Institut für Deutsche Sprache (ed.) OWID – Online-Wortschatz-Informationssystem Deutsch. Mannheim. Accessed at: www.elexiko.de.

OWID – Online-Wortschatz-Informationssystem Deutsch. Accessed at: www.owid.de.

Morphisto. Accessed at: www.ids-mannheim.de/lexik/TextGrid/morphisto.html.

TextGrid. Accessed at: www.textgrid.de.