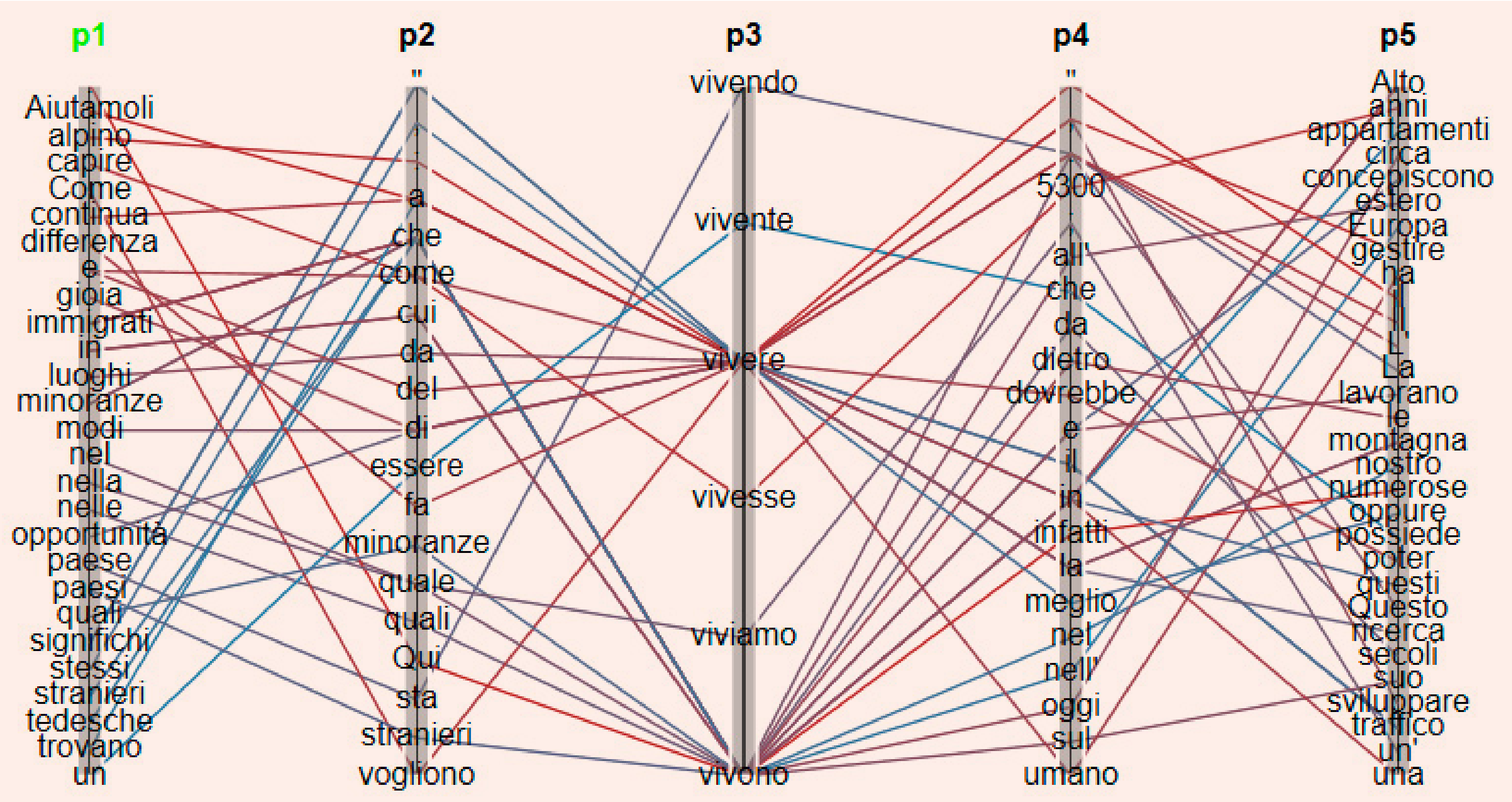
EURAC research

EURALEX

EURALEX 2014



Proceedings of the XVI EURALEX International Congress:

The User in Focus

15-19 July 2014, Bolzano/Bozen

Edited by Andrea Abel, Chiara Vettori, Natascia Ralli

Part 3

Index

Part 1

Plenary Lectures

Phraseology and Collocation

Unusual Phrases in English MLDs: Increasing User Friendliness

Stephen Coffey
Università di Pisa
coffey@cli.unipi.it

Abstract

This paper investigates the presentation of compositionally anomalous phrases in English monolingual learners' dictionaries (MLDS). In particular, it argues that it would be pedagogically useful to explain to the dictionary user, where possible, the reason why certain types of anomaly exist. Two types of phrase are discussed: firstly, idiomatic expressions in which the relationship between phrasal meaning and original meaning may not be clear to the learner (e.g. *run the gauntlet*); secondly, phrases which include particularly unusual word forms or word senses. These include lexical fossils (as in *the whys and WHEREFORES*) and phrases partially motivated by phonological characteristics (as in *bits and BOBS*). In order to form an impression of how anomalous phrases are currently treated in MLDS, samples of items were looked for in both print and online editions. It was found that, overall, little attention is paid to the motivation of phrasal composition, and it is suggested that more should be done in this direction. This would involve integrating current description, almost entirely synchronic in nature, with historical data, at least in the case of some types of phrasal unit.

Keywords: phraseology; learners' dictionaries; idiomatic expressions; lexical fossils; alliteration; rhyme; etymology

1 Introduction

The description of phraseological units has always been an important feature of the English monolingual learner's dictionary (MLD), ever since the publication of Hornby et al's ground-breaking *Idiomatic and Syntactic English Dictionary* in 1942. In the last few decades, lexicographical description of phraseology has reaped enormous benefits from advances made in the field of information technology: specifically, 1) the availability of large language corpora has allowed the description of phraseology to become more complete and more precise, and 2) the arrival of MLDS in digital form, first as CD-ROMS and later through the internet, has meant that learner's dictionaries are now in a much better position to deal with the composite, and sometimes complex, nature of phraseology.[1]

[1] For an overview of the treatment of phraseology in successive editions of MLDS, from the beginnings till the late 1990s, see Cowie 1999 (52-81, and *passim*).

However, although lexicographical description of phraseological phenomena is now generally of a high standard, there are still aspects which could be improved. In this paper I focus specifically on one, pedagogically defined, sub-category of phrases, those which might appear to the learner to be unusual with respect to their lexico-semantic composition.

2 Perceived usualness and unusualness in phrasal composition

2.1 Usualness

There are a vast number of phrases in the English lexicon, and in many cases there is no noticeable clash, for the foreign language learner, between form and meaning. Even if the learner has never before come across (or never noticed) a particular item, the meaning may still be relatively clear (and, indeed, the learner may be unaware of the phrase's status as a lexical unit). Examples of such phrases are *bank manager*, *blood pressure*, *reading glasses*, and *road sign*. Some phrases may seem a little more unusual from a lexico-semantic perspective (e.g. *plastic money*, *money laundering*, *bottle bank*, *blood orange*, *full house*, *front office*, *fuel rod*), but the relationship between form and meaning should be relatively clear once a dictionary has been consulted. The phrase *blood orange*, for example, is defined in OALD 8 as "a type of orange with red flesh", and *bottle bank* (British English) is defined in LDOCE 5 as: "a container in the street that you put empty bottles into, so that the glass can be used again".[2]

Even phrases consisting wholly of a figurative use of the component words will in many cases cause no problems, at least once a dictionary definition has been read. Cross-cultural metaphor may be involved, or else the relationship between physical and figurative meanings may be very evident. Consider, for example, the following phrases and their dictionary explanations:

(1) *playing with fire* MEDAL 2: doing something dangerous or risky that could cause lots of problems for you – "He knew he was playing with fire by encouraging her attentions."

(2) *Out of the frying pan into the fire* OALD 8: from a bad situation to one that is worse.

(3) *water under the bridge* CALD 4: problems that someone has had in the past that they do not worry about because they happened a long time ago and cannot now be changed – "Yes, we did have our disagreements but that's water under the bridge now."

(4) *to have a frog in your throat* CALD 4: to have difficulty in speaking because your throat feels dry and you want to cough.

(5) *not have a leg to stand on* OALD 8: to be in a position where you are unable to prove sth or explain why something is reasonable – "Without written evidence, we don't have a leg to stand on".

[2] Most of the dictionaries cited in the present study are referred to in initialized form (e.g. OALD 8); full bibliographic details may be found in the References.

In cases such as these, it should not be difficult for most learners to connect the literal and lexicalized metaphorical meaning (though with some figurative expressions, recognizing this connection may depend on the linguistic and cultural background of the individual learner).

2.2 Unusualness

Alongside phrases such as those mentioned in Section 2.1, there are also phrases which may give rise, *even after* dictionary consultation, to a perception of discrepancy between form and meaning. Some lexicalized figurative phrases are of this sort, for example:

(6) *throw in the towel* COB 5: If you **throw in the towel**, you stop trying to do something because you realize that you cannot succeed – "It seemed as if the police had thrown in the towel and were abandoning the investigation".

The learner who can appreciate why we say *play with fire* and *water under the bridge*, may well be confused by the phrase *throw in the towel*.[3]

The majority of phrasal verbs could also be described as being compositionally anomalous, especially those in which the verb itself is highly delexicalized, for example *put up with*, *get round sb*, and *take off* (in the sense of "imitate"). A further set of phrases which may appear to the learner as unusual in their form are those which have been called "cranberry collocations" (Moon 1998: 21). This set of phrases is very disparate in nature as regards both form and meaning, but may be grouped together by virtue of the fact that they all include "items that are unique to the string and not found in other collocations" (*ibid*). Many of these items, as Moon points out, "are rare fossil words, or have been borrowed from other languages or varieties" (*ibid*: 78); some of the author's examples are *run AMOK*, *to and FRO*, and *SLEIGHT of hand*. Another set of anomalous items described in Moon's study are those which, for one reason or another, are *grammatically* ill-formed. Examples are *be seeing you*, *by and large*, *in brief*, and *put pen to paper*. Moon also mentions phrases which are highly anomalous from a collocational point of view (e.g. *look daggers at SB*).

2.3 Which 'unusual' phrases to annotate in the MLD

It would be counter-productive to comment on every phrase in which the form-meaning relationship of its component words was notably distant from what would be expected to be a normal juxtaposition of single-word lexical items in modern English. There would be a very large number of items involved, and each explanation would take up space, on the page or on the screen, and perhaps distract from more important information. Furthermore, what is anomalous may go largely or wholly un-

[3] In the case of figurative phrases such as *throw in the towel*, learners will have an advantage if there is an analogous phrase in their own mother tongue. For examples and discussion of such cross-language phrasal pairs and sets, see Piirainen (2012) and, from the perspective of pedagogical lexicography, Coffey (2002).

noticed by the language learner, especially where relatively frequent words are concerned and intuitable meanings. There would be little point in drawing attention, for example, to the fact that, phrases such as *in brief*, *in general* and *at last* are, in effect, PREPOSITION + ADJECTIVE sequences, or that in the phrase *in question* there is no article before the noun. Nor would it be necessary to comment on phrases such as the above-mentioned *blood orange* and *playing with fire*.

Exactly which phrases (or which types of phrase) it would be useful to comment on from the point of view of their composition may be best ascertained through specifically designed dictionary-user studies. However, it is perhaps not unreasonable to suppose that the following two general types of phrase would be strong candidates. Firstly, idiomatic expressions, of one sort or another, in which the motivation behind the idiom's form is wholly or partially hidden from the learner. Examples are the already mentioned *throw in the towel* and the phrase *(as) mad as a hatter*. Secondly, phrases which include a word not normally used as a single-word lexical item in present-day English, for example the word form "sleight" in the phrase *sleight of hand*. In actual fact, some phraseological items would fit perfectly well into both categories, since idiomatic phrases sometimes include fossilized word forms or word meanings. An example is the phrase *(to buy) a pig in a poke* which is both a semantically obscure idiomatic phrase and includes the lexical fossil *poke*.[4]

With both types of phrase, some sort of explanation of unusual form will address the learner's curiosity (even if it will not always be possible to fully satisfy that curiosity). In addition, in the case of idiomatic phrases, an explanation of origin (and therefore of phrasal composition) may help the learner to remember a given expression. In the case of lexical fossils, explicit comments on unusual words will help ensure that learners realize that the words in question are (virtually) phrase-bound and cannot normally be used in other ways.

In order to obtain an overall picture of current practice in MLDS, the following dictionaries were examined: a) print dictionaries: CALD 4, COB 5, LDOCE 5, MEDAL 2, MW, OALD 8; b) online dictionaries: e-CALD, e-LDOCE, e-MEDAL, e-MW, e-OALD. In actual fact, virtually no differences in content were found between the print and online versions of the dictionaries examined, and below I will often cite from, or refer to, the print dictionaries.

3 Idiomatic expressions

The phraseology of English, as is well known, is very complex to describe, and specific categories that we identify (or create) can themselves be very varied in nature. The examples of "idioms" described in this section are no exception to this; there are differences in grammatical role, type of meaning, and the relationship between phrasal meaning and the meaning of individual parts.

[4] For the sake of precision, it should be noted that the word *poke*, which here has the sense of "bag", is still used in some regional varieties.

In order to see whether dictionaries offer the learner any extra guidance with regard to the composition of partially or wholly idiomatic phrases, a total of 37 items were looked for, consisting of two different (pedagogically speaking) groups. The first group was composed of phrases which, it was considered, would definitely benefit from some explicit comment. The second group consisted of items which would not necessarily need any explanation, assuming they were located at the right headword or the appropriate sense of a given headword. The actual composition of these two groups was adjusted slightly once dictionaries had been consulted and current lexicographical data observed.

With regard to the inclusion of the phrases in the dictionaries examined in the present study, 25 out of the 37 phrases were present in all MLDs, and 34 in all dictionaries but one. One dictionary (COB 5) had significantly fewer phrases than the others, with eleven of the phrases being absent.[5]

3.1 Opaque idiomatic phrases

The following are the phrases in the first group (27 items):

a red herring; *a feather in your cap*; *an ivory tower*; *a gravy train*; *a fifth column*; *Bob's your uncle*; *Coals to Newcastle*; *the penny dropped*; *the gloves are off*; *be the bee's knees*; *send sb to Coventry*; *kick the bucket*; *be grist to/for the mill*; *play gooseberry*; *pass the buck*; *face the music*; *run the gauntlet*; *throw down the gauntlet*; *give sb a wide berth*; *throw in the towel*; *pull someone's leg*; *know* etc *the ropes*; *draw/get the short straw*; *live the life of Riley*; *lock, stock and barrel*; *hook, line and sinker*; *as mad as a hatter*.

As regards the form-meaning relationship of these phrases, it was found that, overall, there was very little comment in the dictionaries examined, and that some MLDs had no comment at all. This comes as no surprise, since explanation of this sort would involve introducing the historical dimension to language and the latter has never been a priority in the MLD; indeed, it has usually been completely absent.

As far as I am aware, few writers dealing with pedagogical lexicography have discussed or commented on the absence of historical in data in MLDs. One exception is Ilson (1983), who points out the potential usefulness of providing at least some etymological information, and includes mention of its relevance to phraseology. Much more recently, Boers (2007) points out the usefulness, for comprehension, of being made aware of the origin of idioms. Exemplifying with the expression *show sb the ropes*, he points out that, "It would help if you knew that the expression was originally used in the context of sailing, where an experienced sailor had to show a novice how to handle the ropes on a boat". In the context of the present article, this quotation has added significance since it comes from a short artic-

[5] It might also be mentioned here that all items but two are recorded in the *Collins COBUILD Dictionary of Idioms* (1995) – the exceptions are *fifth column* and *Beauty is in the eye of the beholder*. The latter is not actually an archetypal "idiom", since it is fairly transparent in nature. However, I have included it because the word *beholder* will render the phrase partially opaque to many learners and also because of the relatively unusual phrase "in the eye of".

le entitled Understanding Idioms, which is part of the Language Awareness section of MEDAL 2 (pp. LA2-3).[6]

Of the dictionaries investigated in the present study, the only one which has at least a few explanations of idioms is OALD 8. Two examples are:[7]

(7) OALD 8 **COAL** [...] **carry, take, etc. coals to Newcastle** [UK] to take goods to a place where there are already plenty of them; to supply sth where it is not needed. ORIGIN: Newcastle-upon-Tyne, in the north of England, was once an important coal-mining centre.

(8) OALD 8 **RED HERRING** an unimportant fact, idea, event, etc, that takes people's attention away from the important ones. ORIGIN: From the custom of using the smell of a smoked, dried herring (which was red) to train dogs to hunt.

Examples of definitions with no explanation of phrasal composition are the following:

(9) CALD 4 **MAD** [...] *(as) mad as a hatter / March hare* extremely silly or stupid.

(10) COB 5 **GAUNTLET** [...] PHRASE If you **run the gauntlet**, you go through an unpleasant experience in which a lot of people criticize or attack you – "The trucks tried to drive to the British base, running the gauntlet of marauding bands of gunmen."

There are differing degrees of difficulty in understanding the connection between form and meaning. For example, whereas *run the gauntlet* will be highly obscure to the uninitiated learner, form and me-aning should be much more connectable in the case of the following phrase description, even though there is no explicit explanation of the type seen previously in example (7):

(11) CALD 4 **COAL** [...] **carry / take coals to Newcastle** [UK] to supply something to a place or person that already has a lot of that particular thing – "Exporting pine to Scandinavia seems a bit like carrying coals to Newcastle."

The phrase *ivory tower* may also be relatively clear after reading definition and example, and may not attract too much curiosity on the part of the dictionary user. However, since the phrase has a well do-cumented origin, it might be useful to include it.

Another situation worth commenting on is that wherein a phrase is located at the entry for a single-word headword, but the headword itself has no description. This happens in all dictionaries, for ex-ample, with the word grist, found in the phrase *be grist to/for the mill.* The following is one such entry:

[6] Data regarding phrase origin *is* found, by contrast, in some dictionaries devoted to idioms, notably ODCIE (1983) and LDEI (1979), (whereas the previously mentioned *Collins COBUILD Dictionary of Idioms* does not include explanatory data of this sort). Mainstream language teaching publications have also shown little interest in the historical dimension of phraseology, though there are some applied linguists who recognize the potential of etymological explanation; see, for example, Boers et al (2004) and Boers et al (2007).

[7] From this point of the text onwards, in numbered examples I will use **bold** SMALL CAPS to indicate the headwords. The square brackets which sometimes follow this ([...]) indicate that there is other lexical description before that of the phrase I am discussing.

(12) COB **GRIST** PHRASE If you say that something is **grist to the mill**, you mean that it is useful for a particular purpose or helps support somone's point of view.

The "problem" with *grist* is that it is not commonly found in modern English outside of this phrase. This does not mean, however, that it has become a lexical fossil (except, perhaps, from a language tea-ching point of view). A short explanation of its meaning (in relation to this phrase) would be useful, together with an indication that it is usually only found in this phrase and variants thereof.The same happens with *hook, line* and *sinker* in **COB** 5, which is listed at the headword *sinker*, which, however, has no single-word explanation.

The phrase *Bob's your uncle*, present in five **MLD**s, is a similar case, with the expression being recorded in all dictionaries at the unexplained headword *Bob* (with a capital letter, and thus distinguished from entries with the headword *bob*). Actually, since *Bob's your uncle* is the only phrase at the head-word *Bob* in the various dictionaries, it might be simpler and neater to have the saying itself as the headword, in the same way as many noun phrases and other multiword items regularly constitute headwords.

Another phrase worth commenting on is the *bee's knees*. The five dictionaries which record this phrase give no explanation for its form. If they did, regardless of whether or not they were in a positi-on to explain its meaning from the point of view of its composition, it would be useful to underline the rhyme in the phrase, which is almost certainly at least part of the motivation behind its form. It is worth noting in this respect that sound repetition might well help memorization, and the very act of pointing out the (in this case) rhyme, may be of added benefit to learners(For discussion of this to-pic, see Boers & Lindstromberg 2005 and Lindstromberg & Boers 2008).

If a dictionary does adopt the policy of commenting on phrases of the type being considered here, it will sometimes be necessary to say that the phrase is of "unknown" or "uncertain" origin. In the current state of our knowledge of phrasal origins, this would happen, for example, with *kick the bucket*, *Bob's your uncle*, *face the music*, and *pull someone's leg*. Where there are several contesting theories as to the origin of a phrase, it would probably make little sense to go into details, and be best to label the phrase as being "of uncertain origin". It could be argued that it is of little help to the learrner to read that the origin of a phrase is "unknown", but I think that this is more user-friendly than keeping silent.

3.2 Potentially comprehensible idiomatic phrases

The following are the phrases in the second group (10 items):
Beauty is in the eye of the beholder; *go against the grain*; *jump the gun*; *bury the hatchet*; *play second fiddle*; *pull out the stops*; *make a mountain out of a molehill*; *green about the gills*; *cheek by jowl*; *like a bolt from the blue*.[8]
In each of these phrases, there is a word which may present problems for an understanding of the motivation behind the form of the phrase; of course, whether or not this is actually a problem will depend on the individual learner. The words in question are: *beholder*, *grain*, *jump/gun*, *hatchet*, *fiddle*, *stops*, *molehill*, *gills*, *jowl*, and *bolt*. In the case of *beholder*, *hatchet*, *molehill* and *jowl*, dictionaries only record one sense for each of the words, and if the phrase was explained at the same point of the dictionary at which the single word is explained, then the learner should be in a position to appreciate the motivation of phrasal form. This is what happens, for example, in the case of the following definition:

(13) COB5 **MOLEHILL** (1) A **molehill** is a small pile of earth made by a mole digging a tunnel; (2) If you say that someone is **making a mountain out of a molehill**, you are critical of them for making an unimportant fact or difficulty seem like a serious one.

It is to be noted also that the entry for the word *mole* itself is very close by in the text. So, we have *making a molehill* defined very close to *molehill*, which is itself defined very close to *mole*. And the relationship between the physical and figurative meanings should also be fairly clear to learners. Whereas this may seem a fairly easy case, and the COBUILD treatment of the phrase a useful one for learners, only one other dictionary locates this phrase at the headword **molehill**, the other four placing it at **mountain**. A similar situation is found with the phrase *cheek by jowl*, present in two dictionaries at **jowl**, and in four at **cheek**. We may contrast the following two descriptions:

(14) MEDAL 2 **JOWL** The lower part of your cheek, especially if the skin hangs down and covers your jaw. PHRASE **cheek by jowl** If two or more people or things are cheek by jowl, they are very close to each other.

(15) LDOCE 5 **CHEEK** [...] **cheek by jowl (with sb/sth)** very close to someone or something else – "An expensive French restaurant cheek by jowl with a cheap clothes shop."

One of the reasons that we find dictionary explanations such as that in (15), is the fact that in some MLDs there is a standard rule whereby a phrase should be placed at the headword for the first content word in a phrase. I believe that there are pros and cons to having rules of this type. One important factor which is sometimes overlooked, is that while lexicographers and language teachers have a clear idea (most of the time) about what is, and what is not, a lexical phrase, the average language learner is not as linguistically sophisticated. A learner who reads, for example, of "An expensive French restaurant cheek by jowl with a cheap clothes shop" may view any eventual comprehension problem

[8] The verbal idiom *pull out (all) the stops* is also an example given by van der Meer (1996) in an article dealing with the MLD treatment of figurative meaning more generally (i.e. not just with reference to phraseology).

only in terms of not knowing the word *jowl* – everything else has a familiar look to it, so the problem must be with this word in particular.

Let us turn now to cases where the problem lies not with a possibly unknown word form, but with knowing which meaning of a word is involved (which may or may not be one that is known to the learner). Consider, for example, the phrase *go against the grain*. The relevant sense of the word *grain* is explained in CALD 4 in the following way:

(16) CALD 4 **GRAIN** [...] **the grain** the natural patterns of lines in the surface of wood or cloth – "to cut something along/against the grain".

It is interesting to note that the example phrase includes the phrase "against the grain". However, the meaning is the physical one, not that of the figurative idiom. From the point of view of understanding the motivation behind the idiomatic phrase, this would have been the ideal place to record the figurative expression *go against the grain*; however, it is not recorded at this point in the dictionary. In CALD 4, as in all the other dictionaries, the phrase *is* presented at the entry for *grain*, but as a phrase with no direct connection to any of the single-word senses of *grain*. A similar situation was found for the phrase *pull out all the stops*, (which comes from the idea of an organist pulling out all the organ stops in order to increase the volume). This use of the word *stop* is present in all the dictionaries, but none of them makes a direct association between word meaning and phrasal meaning. Also, three dictionaries place the phrase at the headword *stop* and three at *pull*.

An example of good dictionary treatment is the explanation of *like a bolt from the blue* in MW:

(17) MW **BOLT** a bright line of light that appears in the sky during a storm; a flash of lightning *a bolt of lightning = a lightning bolt* — often used figuratively in the phrases **a bolt from the blue** and **a bolt out of the blue** – "The news of his firing came as/like a bolt from the blue." [= like a bolt of lightning from the sky; it was surprising and unexpected]

The phrase *green about the gills* creates particular problems, since, in order to appreciate the form of the phrase, it may be necessary to see both a definition of *gills* and to understand which sense of *green* is involved. MW satisfies the second need very well:

(18) MW **GREEN** 5. [informal] having a pale or sick appearance – "Our flight hit some turbulence, and half the passengers started turning green." — often used in the phrase **green around/about the gills** – "The passengers were looking green around the gills."

Here, we not only find the phrase at the right sense of *green*, but also see the specification "often used in the phrase ...". The problem still remains, however, of the basic meaning of *gills*, which the learner would have to look up separately.

As has been argued, from the point of view of appreciating the original logic of the phrases, it would make sense (wherever phrases are explained at the entry for one of the component words, as opposed to having an entry of their own), for the phrase to be explained close by the less commonly known word. However, overall it was found that there were few dictionary entries in which the phrases examined were placed at an entry or subsense which would allow motivation of phrasal form to be understood (without explicit commentary). In all, 56 out of the 60 possible phrasal entries were present in the MLDs (10 entries x 6 dictionaries), but in only 12 cases were phrases explained at the appropriate point in the text. In the case of the e-dictionaries, and where a phrase is explained at the "wrong" entry, there is slightly less of a problem, since the reader can go quickly from one entry to another. But it is still a problem, in that the two definitions (of the single word and the phrase) do not appear on the screen together.

4 Other phrases which include unusual word forms or word senses

The second general phrase type for which I suggest it would be useful to have comment on phrasal composition are phrases which include a word not normally used on its own in modern English. In this case, there are two main types of information which the dictionary could provide. The first is, quite simply, the fact that the word in question (or that particular meaning, where homonymy is involved) is normally found just in the phrase indicated. The second data type is the explanation of the unusual word (what sort of word it is, and why it isn't used elsewhere).

There are a number of different reasons for the presence of phraseologically-bound word forms and meanings, and the specific reason will at least in part determine what the dictionary should say about the phrase. From an investigation of many different dictionary entries for lexical phrases, it would appear that the majority of such words are lexical fossils of one sort or another, and it is these that I will look at first.[9]

4.1 Lexical fossils

There are word forms which used to be freer lexical items but which are now found above all "fossilized" in lexical phrases. Some are found in the types of phrase discussed in Section 3. The word *poke* as in *(to buy) pig in a poke* has already been mentioned in this respect; other examples are the words *lurch* and *truck* found in, respectively, *leave sb in the lurch*, and *have no truck with sb/sth*. Examples in phrases which are less likely to be referred to as "idioms" are the already mentioned "fro" (*to and fro*) and "sleight" (*sleight of hand*), and further examples can be seen in the following phrases: *take UMBRAGE, the whys and WHEREFORES, a DAB hand, by DINT of*, and *in fine FETTLE*.

[9] The term "fossil", in a linguistic sense, is defined in the OED (3rd edn) as "A word or other linguistic form which has become obsolete except in isolated regions or in set phrases, idioms, or collocations". For a discussion of the notion of "lexical fossil", see Coffey 2013.

Some fossils are close in form to related words in modern English; an example is *afield*, used in phrases such as *far afield* and *farther afield*. Fossils may also have exactly the same form as a modern word, and thus be less noticeable. This applies to the already mentioned *poke* and *truck*. Other examples are the word forms "let" and "hue", as found in *without let or hindrance* and *hue and cry*. Grammatical word category may also be of relevance: the word *pale* is not usually found as a noun in modern English, but this usage (and relevant meaning) can still be seen in the phrase *beyond the pale*.

The fact that the words or word forms are not normally used in modern English as single-word lexical items has a number of consequences for language learners. Firstly, the learner may be puzzled as to the presence of a word in a phrase – why do we say *a pig in a poke*: what is a "poke" in this case? Secondly, the learner may feel a sense of frustration at not knowing anything about a word. The phrase *in high dudgeon*, for example, appears most frequently in MLDs under the headword **dudgeon**, implying, therefore, that the latter exists as a free-standing word, which, however, it doesn't. Thirdly, the learner may remember the unusual word (precisely because it is unusual), and later use it in an inappropriate way, for example by taking the word "umbrage" out of its usual phrasal environment (*take umbrage*).

Taken as a whole, the MLDs examined do not have much to say about the composition of items such as the above. The following are some examples of presentation:

(19) MEDAL 2 **POKE** a quick push with your finger or a pointed object. [...] **a pig in a poke** something that you have bought without seeing it first.

(20) OALD 4 **WHY** [...] *noun* **the whys and (the) wherefores** the reasons for sth – "I had no intention of going into the whys and the wherefores of the situation."

(21) E-LDOCE **FETTLE** *noun* **in fine/good fettle** [old-fashioned] healthy or working properly.

(22) CALD 4 **DINT** *noun* **by dint of sth** [formal] as a result of sth – "She got what she wanted by dint of pleading and threatening."

(23) COB 5 **WEND** PHRASE If you **wend** your **way** in a particular direction, you walk, especially slowly, casually, or carefully, in that direction [LITERARY] – "Sleepy-eyed commuters were wending their way to work."

(24) MW **PALE** *noun* **beyond the pale** offensive or unacceptable – "conduct that was beyond the pale".

With regard to the lexical fossils within phrases, the most useful feature I have found in the dictionaries examined is wording which is used sometimes in MW, and an example of which involves the word *umbrage*:

(25) MW **UMBRAGE** a feeling of being offended by what someone has said or done — usually used in the phrase **take umbrage** – "I imagine some people will take umbrage [= will be offended] when they hear the quote."

The important words here are "usually used in the phrase ...". This type of comment is not found for all fossils in MW (nor is it found only for fossils), but it is a step in the right direction.[10]
How much information should be given for a fossilized phrase will depend on the word in question. Often, it will be enough to indicate that it *is* a fossil and therefore probably only found in that particular phrase, but sometimes other information could be useful. For example, in the case of *WEND one's way*, it may be of interest to the learner to know that the verb *wend* is historically related to the word form *went*, now considered to be part of the verb *go*; and in the case of *without further/more ADO*, reference could be made to the play title *Much ado about nothing*.

4.2 Other unusual words in phrases

I will now briefly mention other types of phrase which may strike the learner because of one or more unusual word forms. First, I list a number of example descriptions in MLDS, and thereafter I comment on the features I wish to point out.

(26) COB 5 **BOB** [...] PHRASE **Bits and bobs** are small objects or parts of something [mainly British, informal] – "The microscope contains a few hundred dollars-worth of electronic bits and bobs."

(27) MW **ODDS AND SODS** [UK, informal] = **odds and ends** – "The store sells art supplies and other odds and sods."

(28) e-LDOCE **KITH AND KIN** *noun* [old-fashioned] family and friends.

(29) e-LDOCE **LOVEY-DOVEY** *adj* [informal] behaviour that is lovey-dovey is too romantic – "a lovey-dovey phone call".

(30) MEDAL 2 **CHIT-CHAT** *noun* [informal] friendly conversation about things that are not very important.

(31) CALD 4 **BUTCHER** *noun* [...] **have a butcher's** [UK, old-fashioned, slang] to look at something – "Let's have a butcher's at your present then."

In (26) can be seen an example of a word (*bob*, or rather *bobs*) which does not exist on its own with this type of meaning. Nor is it known to be a lexical fossil. The phrase is recorded in the oed (2nd edn) at the entry for *bit*, and there is no meaning of *bob* which might relate to this phrase. It may be presumed, therefore, that the phrase was coined for its alliterative effect. Example (27) is similar but involves rhyme rather than alliteration. The second phrase indicated, *odds and ends*, also involves sound repetition, since the two parts of the binomial are both monosyllabic, begin with a vowel, and end with the spelling/sound *-ds*. In (28), the phrase *kith and kin* combines alliteration with the presence of a lexical fossil (*kith*). Examples (29) and (30) also exhibit, respectively, rhyme and alliteration. However,

[10] Information about some lexical fossils may also be found in the brief etymological notes on individual words ("word origin") in e-LDOCE and the CD-ROM version of OALD 8; however, there is no explicit statement about the fact that the words in question are fossils nor that they are only usually found in the phrases indicated.

they are formally different from the preceding examples, in that they are reduplicatives, and are written, usually, as single, hyphenated words. In the case of *lovey-dovey*, both morphological parts are easily associated with other words (*love* and *dove*), though the word *lovey* also exists. The word *chit-chat* is different, in that only *chat* exists with a relevant meaning.

Examples (26) to (30), then, all involve sound repetition of some sort, in addition to the presence of unusual words or morphemes. Example (31) is a little different, in that there is no apparent sound repetition, just a word sense which seems unconnected to any of the various meanings of the headword *butcher*. However, rhyme is involved, indirectly, as can be seen in the explanation of the same phrase in MEDAL:

(32) MEDAL 2 **BUTCHER** *noun* [...] **have/take a butcher's** [UK, informal] to have a look at something – From *butcher's hook*, rhyming slang for 'look'.

Generally speaking, there is little comment in current MLDs on the types of phrase mentioned in this section. And whereas there are no more than a handful of word usages dependent on rhyming slang, there are many more items in which sound repetition has a fundamental role to play. It would not be difficult to point this out in the dictionary, and may help to satisfy the reader's curiosity as to the origin of the phrase.

4.3 Grouping phrases which share certain features

As part of the process of improving learners' general knowledge of the English lexicon, it would be useful if phrases which share a certain feature or features were brought together. This would be much easier in the e-dictionary, where space is not a problem and where the user could click on a link to find other examples of the phenomenon being looked at in a particular entry. Some of the phrasal types mentioned in 4.1 and 4.2 could be brought together in this way. In the case of fossils, since there are quite a large number involved, those with a further common characteristic could be brought together, for example those in coordinated phrases involving sound repetition (e.g. *kith and kin*, *the whys and wherefores*). Some idiomatic phrases could also be interlinked, for example with reference to the area of "original" meaning of phrases (e.g. "SHIPS AND THE SEA") or through the form of the phrases (e.g. *as [mad] as a [hatter]*).

Using the terminology employed by Apresjan (1993: 80), interconnections of this type would allow the dictionary to enhance its information on "lexicographic types", while at the same time, the improvement of data regarding phrasal composition would also allow each individual "lexicographic portrait" (*ibid*: 86) to be enriched.

5 Conclusions

English lexical phrases come in many shapes and sizes, and some of these shapes and sizes are dependent on factors which are not obvious to many present-day native speakers, let alone learners of English as a foreign language. However, whereas lexico-phraseological oddities cause no problems for native speakers and also usually go largely unnoticed, the same cannot be said for language learners, who tend to be more aware of form and have to reconcile it with meaning. Whereas I believe it is generally a positive asset that learners' dictionaries do not dwell too much on the compositional nature of lexical phrases, I think that they should do so when appropriate. In the types of phrase described in the present paper, this also involves bringing in the historical dimension of language, which has perhaps been left out in the cold too long. Given, especially, the enormous potential of web-based dictionaries, this should be perfectly possible.

6 References

Print dictionaries

CALD 4. *Cambridge Advanced Learner's Dictionary*, 4th edn (2013). Cambridge: Cambridge University Press.

COB 5. *Collins Cobuild Advanced Dictionary of English*, 5th edn (2006). Glasgow: HarperCollins.

Collins Cobuild Dictionary of Idioms (1995). London: HarperCollins.

Hornby, A.S., Gatenby, E.V. & Wakefield, H. (1942). *Idiomatic and Syntactic English Dictionary*. Tokyo: Kaitakusha. [later, 1948, published by Oxford University Press as *A Learner's Dictionary of Current English*, and subsequently, 1952, retitled *The Advanced Learner's Dictionary of Current English*].

LDOCE 5. *Longman Dictionary of Contemporary English*, 5th edn (2009). Harlow: Longman.

LDEI. *Longman Dictionary of English Idioms* (1979). Harlow: Longman.

MEDAL2. *The Macmillan English Dictionary for Advanced Learners*, 2nd edn (2007). Oxford: Macmillan Education.

MW. *Merriam-Webster's Advanced Learner's Dictionary* (2008). Springfield, Massachusetts: Merriam-Webster Inc.

OALD 8. *Oxford Advanced Learner's Dictionary*, 8th edn (2010). Oxford: Oxford University Press.

ODCIE. Cowie, A.P., Mackin, R. & McCaig, I. R. (1983). Oxford Dictionary of Current Idiomatic English, Volume 2: Phrase, Clause & Sentence Idioms. Oxford: Oxford University Press.

On-line dictionaries

e-CALD. Accessed at: http://dictionary.cambridge.org/dictionary

e-LDOCE. Accessed at: http://ldoce.longmandictionariesonline.com

e-MEDAL. Accessed at: http://www.macmillandictionary.com

e-MW. Accessed at: http://www.learnersdictionary.com

e-OALD. Accessed at: http://oald8.oxfordlearnersdictionaries.com/dictionary, and at: http://www.oxfordlearnersdictionaries.com/

OED. *The Oxford English Dictionary*. Oxford: Oxford University Press. Accessed at: http://www.oed.com

All online dictionaries were accessed at various times during the period September 2013 – April 2014

Other literature

Apresjan, J. D. (1993). Systematic Lexicography as a Basis of Dictionary-making. In *Dictionaries*, 14 (1992/93), pp. 79-87.

Boers, F. (2007). Understanding Idioms. In the *Macmillan English Dictionary for Advanced Learners*, 2nd edn, pp. LA2-3. This article is also available in the *MED Magazine*, Issue 49, February 2008. Accessed at http://www.macmillandictionaries.com/MED-Magazine/February2008/49-LA-Idioms-Print.htm [02/04/2014]

Boers, F., Demecheleer, M. & Eyckmans, J. (2004). Etymological elaboration as a strategy for learning idioms. In P. Bogaards, B. Laufer (eds.) *Vocabulary in a Second Language: Selection, acquisition, and testing*. Amsterdam / Philadelphia: John Benjamins, pp. 54-78.

Boers, F. Eyckmans, J. & Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: more than mere mnemonics? In *Language Teaching Research*, 11 (1), pp. 43-62.

Boers, F. & Lindstromberg, S. (2005). Finding ways to make phrase-learning feasible: The mnemonic effect of alliteration. In *System*, 33, pp. 225-238.

Coffey, S. (2002). Interlingual Phrasal Friends as a Resource for Second Language Learning: Outline of a lexicographical project. In A. Braasch, C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress*. Copenhagen: Center for Sprogteknologi, pp. 315-323.

Coffey, S. (2013). Lexical Fossils in Present-Day English: Describing and Delimiting the Phenomenon. In R. W. McConchie, T. Juvonen, M. Kaunisto, M. Nevala & J. Tyrkkö (eds.) *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3).* Somerville MA: Cascadilla Proceedings Project, pp. 47-53.

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

Ilson, R. (1983). Etymological information: can it help our students? In *ELT Journal*, 37(1), pp. 76-82.

Lindstromberg, S. & Boers, F. (2008). Phonemic repetition and the learning of lexical chunks: The power of assonance. In *System*, 36, pp. 423-436.

Moon, R. (1998). Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford: Clarendon Press.

Piirainen, E. (2012). Widespread Idioms in Europe and Beyond: Toward a Lexicon of Common Figurative Units. New York: Peter Lang.

van der Meer, G. (1996). The Treatment of Figurative Meanings in the English Learner's Dictionary (OALD, LDOCE, CC and CIDE). In M. Gellerstam, J. Järborg, S-G. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (eds.) *Euralex '96 Proceedings*. Göteborg: Department of Swedish, Göteborg University, pp. 423-429.

Harvesting from One's Own Field: A Study in Collocational Resonance

Janet DeCesaris, Geoffrey Williams
Universitat Pompeu Fabra, Université de Bretagne-Sud
janet.decesaris@upf.edu, geoffrey.williams@univ-ubs.fr

Abstract

This paper presents an initial study in the collocational resonance of three words: *field*, *champ*, and *campo*. *Field* and *champ/campo* do not share the same etymology, yet *field* displays sense extension that in some cases is parallel to that displayed by *champ* and *campo*. Collocational resonance posits that meaning associated with one context of use may be activated by speakers in another context, even though the original meaning context may fall into disuse in the language. To determine collocational resonance, the study considers both the historical development of senses as represented in dictionaries and the current behaviour of these words as represented in dictionaries and corpora. It is suggested that an approach to word meaning including prototypes and resonance can improve the representation of polysemy in dictionaries.

Keywords: collocational resonance; polysemy, metaphor

1 Introduction

Both homonymy and polysemy have long posed problems for lexicography. Whilst the former can be relatively easily handled from a synchronic perspective, this is not so in diachrony. Thus, a historical dictionary may well have to lump senses together in a single entry when a dictionary of the contemporary language would split them into different entries, as historically, what is now justifiably analysed as a homonym, might well be a polyseme. Even in dictionaries of contemporary language, however, the representation of polysemy is problematic, as the requirement for discrete senses denies the obvious continuum between senses. Obviously, these issues are rendered even more complex as language moves from a so-called literal sense to a figurative one through metaphor. In such movements, to adopt the terminology of Hanks (2013), the exploitation of the norm becomes the norm itself, which will, in turn, inevitably be exploited. Four issues thus require a tool for their joint management: etymology, homonymy, polysemy and metaphor.

Hanks (2000) has already proposed a solution to polysemy by proposing lexicographical prototypes, a series of simple propositions that are activated in each individual sense. Hanks (2005) also proposed a means of handling metaphor by proposing collocational resonance. At the same event, Williams (2008) independently also proposed collocational resonance, albeit approaching the data from a more inter-textual angle. Both researchers are heavily influenced by John Sinclair, so it is it not altogether surprising that, working independently, they should reach very similar conclusions and a selfsame term. Since these landmark presentations, work has gone on to combine lexicographical prototypes and collocational resonance in order to handle the variation of meaning potentials (Hanks 2013) across time and between languages (Williams 2012).

In this paper, we shall look at the word *field* and its equivalents in French, *champ*, and Spanish, *campo* as the beginnings of a longer study into the norms and exploitations of the agricultural metaphor in contemporary use and at the treatment of these words in dictionaries. In this study, we shall concentrate on the dictionary definitions that provide the initial prototype and the current usages of the word as found in corpora.

2 Collocational Resonance

The idea behind collocational resonance is that over time words have been attributed different meanings in different contexts. We define meaning as referring to a particular usage that can be defined through a series of propositions that in a stable generalised form provide what can be recognized as a dictionary sense. Meanings are created within a given textual environment within a given context of culture. Thus, meaning elements are shared within a society, with general agreement as to broad senses. As society and contexts of culture evolve, so do the meaning potentials of words. There can be a slight contextual variation, or a more radical one as an exploitation becomes a norm.

Collocational resonance posits that although the earlier senses attributed to a word may be lost, we live in a world of cumulated knowledge so that some meaning attributes may subsist, consciously or unconsciously, thereby colouring a user's use of a word. When there is a deliberate exploitation, for example through active metaphor, the user is quite conscious of the exploitation being made of meaning attributes, and the reader or hearer is expected to share this explicit knowledge. Collocational resonance eschews any so-called cognitive knowledge and posits that while a dead metaphor is simply dead, the knowledge of earlier usage may be found in the unconscious as this knowledge comes from an encounter with the word in a different meaning context, either in a text, a dictionary or through the educational process. The unconscious aspect of meaning attribute carry-over has been demonstrated by Williams (2012) in the case of biologists using rather Lamarkian terms when referring to neo-Darwinian concepts. This variation need not be diachronic; resonance can equally well show variations between general and specialised usage of language.

The aim of studies in collocational resonance, then, is to make meaning variation explicit over time or area of expertise and to show what meaning attributes may remain active. This has been shown for certain verbs used in the sciences, such as *probe*, and also for more general words such as *culture*. The word culture is an interesting case from a cross-linguistic perspective, as it has a common root in all Romance languages and has developed through metaphor in a similar way in all of them. Nevertheless, the earlier, literal meaning of *culture* ('the cultivation of soil; tillage'[1]) has remained active in some languages, but has virtually disappeared from the use of *culture* in English, although it is clearly present in the use of the derived word *agriculture*. Collocational resonance can be traced using lexicographical prototypes built using a mixture of sources such historical dictionaries as the *Oxford English Dictionary*, earlier dictionaries as the *Vocabolario* of the Accademia della Crusca, 17th century French dictionaries, 17th century Spanish dictionaries, the Spanish Royal Academy's many editions of the *Diccionario de la lengua castellana* (which would become the *Diccionario de la lengua española* starting with the 1925 edition) and diachronic corpora, when available. Collocational resonance may be demonstrated by showing variations in the collocational networks of lexical items: as the meanings of a word change, so do the collocational patterns associated with the word in question.

When prototypes are used to show variation of meaning potentials over time, contexts or languages, there is no privilege interlanguage acting as a translation hub. The prototypes are there to show meaning potentials and can be started in any language, the aim being to find translatable comparable units in other languages to see what potentials are activated. If here we are starting with the English, it is only because of the excellence of the *Oxford English Dictionary* as a starting point for looking at earlier usage. For French, we have used the *Trésor de la Langue Française informatisé* and the *Dictionnaire Historique de la Langue Française* from Le Robert and for Spanish the Spanish Royal Academy's *Nuevo tesoro lexicográfico de la lengua española*. In this study, we look at words with no etymological relation at all, *field* and *campo*/*champ*, but which are generally accepted as translation equivalents. In all cases, current usage is analysed making use of the Sketch Engine® tenten series of WaC corpora.

3 *Field*: From ploughing the soil to ploughing through data

Whether it be through culture or through units of land, agricultural metaphors are rife in that our modern urbanised societies could not exist without the organised production of food. The interest of the words under study lies in their totally different etymological origins, which signifies a potentially major difference in resonance. The etymology of *field* is unclear, but by the early Middle Ages it has taken on the notion of open land rather than woodland. In the Romance languages, it is easier to trace back to a Latin source that distinguishes plain and mountain. By the time of the *Vocabolario*, *campo* as being an agricultural area in which seeds are sown, but also a wide variety of other senses, including that of battlefield. Moving forward to the *Dictionnaire Universel* of Antoine Furetière (1690), *champ* has as it first sense an area of ploughed land. Thus, both Italian and French sources confirm an agricultural meaning of cultivated land as opposed to pasture land. Neither has any clear boundaries. Furetière also gives a large number of non-agricultural terminological uses, from heraldry to comb making.

[1] Paraphrase of sense I.1.a. of culture in the Oxford English Dictionary and also sense 4 of culture in The American Heritage Dictionary of the English Language.

In Spanish, the situation is somewhat different. Covarrubias (1611) lists the idea of flat land capable of being cultivated or an enclosed area used for farm animals as the first sense of *campo*. The entry for *campo* in the *Diccionario de autoridades* (1729, for letter C) is quite long, comprising several subsenses for the word in addition to several lexicalised phrases (such as *campo de batalla* 'battlefield' or *hombre de campo* 'man who works in fields'). The first sense defines *campo* as a wide, open plain that is outside a populated area; the notion of lying outside a populated area is quite prevalent throughout the twenty-two editions of the Spanish Royal Academy's dictionary and is still present in the first sense given for the word in the current dictionary ("*Terreno extenso fuera de poblado*" 'Large piece of land outside a populated area') . The second sense of *campo* in the *Diccionario de autoridades* is described as metaphorical and defines *campo* as the space or period including the whole of something, and the third sense of *campo* includes the idea of cultivated area (thus, the examples *están buenos los campos* 'the fields are fine', *los campos están perdidos* 'the fields are lost', and *buen año para los campos* 'good year for the fields'. Those three senses would be maintained with little variation as the first three senses listed until the Academy's dictionary of 1837, in which the idea of plain as opposed to mountain is listed as the second sense. Other subsenses listed in the *Diccionario de autoridades* for *campo* refer to *campo* as an army, *campo* used in textiles and in heraldry, and *campo* as the location chosen for a duel.

What emerges from this brief historical examination of data from three languages is that a wide variety of senses from two root etymologies have arisen and have merged in certain areas over time. Thus, *field*, *champ* and *campo* as agricultural units of land provide a common starting place for an exploration that will lead to areas of scientific endeavour, fields of study.

An initial prototype is drawn starting from dictionaries, notably the Oxford English Dictionary, Trésor de la Langue Française informatisé and early editions of the *Diccionario de la lengua española* (see Table 1) and consists of extracts from the initial entries for *field*, *champ*, and *campo*. As Table 1 shows, there is no direct correspondence across entries: in English and French, the idea of a piece of land that is delimited arises early on, whereas in Spanish the idea that the land lies outside of, and thus contrasts with, a populated settlement is important.

English	**French**	**Spanish**
a piece of ground	espace d'une certaine étendue	terreno extenso fuera de poblado
open land as opposed to woodland		en contraposición a sierra o monte, *campiña*
land or a piece of land appropriated to pasture or tillage	étendue plate de terre arable	tierra laborable
usually parted off by hedges, fences, boundary stones, etc.	plus ou moins nettement délimité	
	étendue plate de terre arable caractérisée par l'absence de clôture	
		sembrados, árboles y demás cultivos.

Table 1. Some prototypes for *field, champ*, and *campo.*

3.1 Field

Collocational networks have been drawn up for the first 5 occurrences in each category of four areas of Sketch Engine® output: 'object_of,' 'subject_of,' 'modifier' and 'modified.' An initial sweep of the object collocates brings forth four large meaning classes: [agriculture], [disciplines], [opportunities], and [sport]. Widening to the three other areas provided by the Sketch Engine® data, we can add [computing and mathematics], [physics], [in vivo situations] and [rural].[2] This done, it is now possible to see what aspects of the prototypes are activated in each case. In the text that follows, the concept areas are square bracketed and the collocates are in italics.

Unsurprisingly, the agricultural and rural senses are closely linked as these are the oldest recorded senses. Both are areas: more specifically, enclosed areas of unwooded ground. The delimitation of the agricultural *field* may be explicit or implicit, as we are hindered by our limited view of clearances, commons and enclosures as acts of creating and appropriating open spaces. The important aspects underlined are tillage, *ploughing* and *sowing*, and food production, which includes *grazing*, although the latter does not require enclosure. This area is relatively level so as to permit tillage. [Rural] is similarly areas of land, and it is opposed to urban areas, which they *surround*. It is possible that woodlands are included, as this is simply an area found in proximity to another area, *villages*, or *towns*. What we term [in vivo situations] are partially related to these, such as the notion of *field trial*, as opposed to laboratory testing, which implies getting out of an enclosed environment into an open space, which when linked to farming is a field as a place for food production. The notion of getting out into the open also comes with *field trip* and *field recording*. The exploitation of the prototype is thus an area, which is delimited, used for agricultural production, is not woodland, lies outside of an urban area, and is a closed space. [Sport] makes use of this as well, but further limits the area. Note, however, the salience of the parameters of outdoors, level and treeless in [sport].

Agricultural metaphors as *cultivate* [the mind] and *culture* [the arts] are frequent. It could be considered surprising thus that *fields of study* largely pre-date the metaphor of *culture*. What is carried over is the notion of a defined area, but a more dynamic one than in agriculture as new *fields* can *emerge*, and as enclosed spaces, they can be *entered*. What is emerging implies new [opportunities] that can be *wide* or *narrow*. Taking the metaphor of delimited area further, we can find the notions defined in [physics] such as electromagnetism, which also hark back to the more or less delimited area as there are no clear barriers here, unlike in computing, where a *field* in a database is limited and needs to be filled in, and is possibly like a *playing field* in that it is rectangular.

[2] Although the Sketch Engine® is an extremely powerful and useful tool, the data provided in a Word Sketch must be carefully analysed and, in some cases, checked against the examples in because sometimes the results can be misleading. Errors in tagging can occur; for example, in the Word Sketch for campo under the category 'subject of,' the proper noun mauthausen is given as the verb with the highest MI index of 6.72. Other proper nouns, such as Valderrama, Covadonga, and Huelva, also erroneously appear in the 'subject of' column, so presumably the grammar used for the analysis needs to be revised. Past participles are considered verb forms, but in many contexts are used as adjectives and as such the results in the column 'sujet de' for the French corpus includes constructions in which the word champ is not a subject (for example, expressions like un champ cultivé or les champs fleuris). Nevertheless, we have found this tool to be useful for providing a quick, overall picture of a word's behaviour.

What the above shows is the subtle linkage of concept areas that in dictionary terms would be called senses. Rather than simply describing them, the prototype approach can be used to show linkage between senses through what is being activated. It also allows us to map change across time and between languages.

3.2 A *champ* is and is not a *field*

As was done for *field* and *champ*, collocational networks have been drawn up for the first 5 occurrences in each category of four areas of Sketch Engine® output: 'object_of,' 'subject_of,' 'modifier' and 'modified.' Although [agriculture] and [disciplines] are common object collocates with *champ*, the notion of extending something (which inherently must have a limit) is very salient in the 'object of' output (the verb *élargir* 'broaden', for example, shows a Mutual Information (MI) index of 9.3 and *étendre* 'extend', an MI index of 6.87, in the French tenten corpus). The salience of the notions of 'used for agricultural production,' especially of crops, and 'treeless' is clear, as *champ* often occurs in a context in which *fôret* and *pâturage* are also listed and contrast with the idea evoked by *champ*.
The concept area with the highest MI index for '*champ* + modifier' is [physics], with *magnétic* and *electromagnetic* displaying both high frequency and high MI indices. Perhaps surprisingly, the concept area of [language] is salient in the French corpus data: we have *champ lexique* (MI index, 9.07) and *champ semantique* (MI index, 7.7). The extension of the agricultural metaphor in present in French but has taken a somewhat different direction from that in English; for example, the verb *cultiver*, which is a very strong collocate for *champ*, prefers crops and plants as a direct object, although one can also *cultiver le paradoxe*, *cultiver l'ambigüité*, and *cultiver la nostalgie* (none of which are typically *cultivated* in English, according to the corpus data).

3.3 A *campo* is and is not a *field*

As was done for *field* and *champ*, collocational networks have been drawn up for the first 5 occurrences in each category of four areas of Sketch Engine® output: 'object_of,' 'subject_of,' 'modifier' and 'modified.' Although [agriculture] and [disciplines] are common object collocates with *campo*, as in English, [computing] is very salient in the 'object of' output (the verb *rellenar* 'fill in', for example, shows an MI index of 9.27 in the European Spanish tenten corpus). Interestingly, the notion of [enclosure] , which is not included in the dictionary definitions for the several senses of *campo*, does seem to underlie some usage in contemporary Spanish, as the verb *delimitar* 'delimit' shows a reasonably high MI index for 'object of' (6.66). The area of [confinement of people], to which the underlying notion of enclosure is inherent, is also very salient for *campo*: we find *deportar* 'deport' ('subject of'); *concentración* 'concentration,' *exterminio* 'extermination,' and *refugiado* 'refugee' (n_modifier). Of course, that subject area in English is not associated with *field* but rather with the etymological cognate of *campo*, *camp*.

The notion of getting out into the open, which gives rise to much phraseology and several lexicalized expressions in English (e.g. *field trip*, *field work*), is important to the collocational network for *campo* mainly in conjunction with the nouns *trabajo* 'work', *experimento* 'experiment' and *estudio* 'study'; this notion is much less salient in the network for *campo* than it is in the network for *field*.

An important difference between the behaviour of *field* and *campo* is related to the area [military]: one can *invadir* 'invade' *campos* and be the *mariscal de campo* ('field marshall') on a *campo de batalla* ('battlefield'; MI index of 9.54). The fact that English *battlefield* is a morphological compound surely explains the fact that this subject area is not as salient for *field* as it is for *campo*.

In English, the verb *cultivate* is a common collocate of *field*, and similarly in Spanish, *cultivar* is a common collocate of *campo*. The agricultural metaphor appears to have been extended further in English, however, as *cultivate* often takes abstract nouns as a direct object (*mindfulness*, *friendship*, *relationship*, *virtue*), whereas the corpus data show that Spanish *cultivar* overwhelmingly takes crops and plants as its direct object. In fact, the Word Sketch only lists one direct object in 25 that is not literally related to agriculture and that word is *amistad* 'friendship.'

In Spanish, the concept area [sport] is present mainly because of *campo* is used to refer to golf courses (Spanish, *campo de golf*) and football fields (*campo de fútbol*), which can be stepped on (*pisar el campo*) once it has been opened and inaugurated (*abrir/inaugurar el campo*). This concept area, however, appears to play a smaller role in Spanish than it does in English.

Interestingly, the prototype of *campo* as being located away from a populated centre is still prominent in Spanish. Under the category of 'and_or', and discarding proper nouns which produce errors, the two nouns that are in some sense complementary to *campo* are *bosque* 'forest' (MI index of 6.09) and *montaña* 'mountain' (5.85). Using the Sketch Engine to consult a different corpus (the Spanish web corpus), the noun that appears with the highest MI index in this category is *ciudad* 'city,' which clearly evokes the contrast with *campo*.

4 Dictionary representation of collocational resonance

Collocational resonance, as stated earlier, claims that meaning attributes can be carried over from one context to another. As such, it can, and we believe, should, be taken into account in dictionary representation because it can help to show that there is a relationship between senses that are depicted as discrete items on a list. Such listing practice, of course, may be unavoidable in dictionary representation, but as Fillmore (1975) and Hanks (2000; 2013) have argued, is not necessarily a proper approach to word meaning. Monolingual dictionaries could attempt to highlight the relationship between senses, perhaps by ordering senses to show derived meanings and or by stating that a meaning has developed as an extended sense and fits into a metaphor that is operative in the language. To date, few monolingual dictionaries have attempted to represent metaphor in their entries, although the *Macmillan English Dictionary* (both in the printed and online versions), with 'Metaphor Boxes' and a section on Metaphor in the body of the dictionary, stands out in this respect. For the purposes of this study, we shall consider the entries for the noun *field* in the *Macmillan English Dictionary Online* and in *The American Heritage Dictionary of the English Language*, shown in Figures 1 and 2, respectively.

1 [COUNTABLE] an area of land used for keeping animals or growing food

There were horses grazing in the next field.

a corn/wheat field

field of:

We drove past huge fields of barley and hay.

a. an area of land covered in grass and used for sport

The England striker left the field with a knee injury.

a sports/football field

take the field (=walk onto it in order to start playing):

The crowd gave Ripken a standing ovation when he took the field.

on/off the field:

He behaves badly both on and off the football field.

b. a large area of land or water where something is found

a gas field

c. a large area of land or water covered in a particular substance

an ice field

d. MAINLY LITERARY an area of land where people fight a battle

2 [COUNTABLE] a subject that you study, or a type of work that you do

field of:

a chemist working in the field of polymer research

a field of study/endeavour/enquiry:

She has the ability to succeed in any field of endeavour.

a specialist/expert in a field:

Professor Edwards is one of the main experts in his field.

3 [SINGULAR] all the people or animals taking part in a race or competition: can be followed by a singular or plural verb

Henderson will be competing against a very strong field today.

4 [COUNTABLE] COMPUTING a part of a database that contains information of a particular type

Type your name in the User field.

5 [COUNTABLE] PHYSICS an area where a particular force has an effect

a magnetic field

6 [COUNTABLE] an area that a person or piece of equipment can see at one time

7 **the field** the team in baseball, cricket etc that is throwing the ball and trying to catch it when the other team hits it: can be followed by a singular or plural verb

Figure 1: Macmillan English Dictionary Online.

1. a. A broad, level, open expanse of land.

b. A meadow: *cows grazing in a field.*

c. A cultivated expanse of land, especially one devoted to a particular crop: *a field of corn.*

d. A portion of land or a geologic formation containing a specified natural resource: *a copper field.*

e. A wide unbroken expanse, as of ice.

2. a. A battleground.

b. *Archaic* A battle.

c. The scene or an area of military operations or maneuvers: *officers in the field.*

3. a. A background area, as on a flag, painting, or coin: *a blue insignia on a field of red.*

b. *Heraldry* The background of a shield or one of the divisions of the background.

4. a. An area or setting of practical activity or application outside an office, school, factory, or laboratory: *biologists working in the field; a product tested in the field.*

b. An area or region where business activities are conducted: *sales representatives in the field.*

5. *Sports*

a. An area in which an athletic event takes place, especially the area inside or near to a running track, where field events are held.

b. In baseball, the positions on defense or the ability to play defense: *She excels in the field.*

c. In baseball, one of the three sections of the outfield: *He can hit to any field.*

6. A range, area, or subject of human activity, interest, or knowledge: *several fields of endeavor.*

7. a. The contestants or participants in a competition or athletic event, especially those other than the favorite or winner.

b. The body of riders following a pack of hounds in hunting.

c. The people running in an election for a political office: *The field has been reduced to three candidates.*

8. *Mathematics* A set of elements having two operations, designated addition and multiplication, satisfying the conditions that multiplication is distributive over addition, that the set is a group under addition, and that the elements with the exception of the additive identity form a group under multiplication.

9. *Physics* A region of space characterized by a physical property, such as gravitational or electromagnetic force or fluid pressure, having a determinable value at every point in the region.

10. The usually circular area in which the image is rendered by the lens system of an optical instrument. Also called *field of view.*

11. *Computers*

a. An element of a database record in which one piece of information is stored.

b. A space, as on an online form or request for information, that accepts the input of text: *an address field.*

Figure 2: field in The American Heritage Dictionary of the English Language.

In the Macmillan entry, the fact that sense (1a), that of a *field* used in the concept area [sports], is separated from sense (3), the people taking place in a sporting competition, makes it difficult to see the relationship between these two senses. Notice that the same problem occurs in the American Heritage Dictionary, in which changing the order of senses (5) and (6) might make things clearer. Although the notion of boundary is latent in the wording of the definitions (notice the frequent occurrence of the preposition *in*, as in 'An area *in which* an athletic event takes place', the idea of *field* as an enclosure is not really explicit in either entry.

Entries for *field* in even very good, large bilingual dictionaries do not address the subtle differences an analysis of collocational resonance can reveal. Let us look at the entry for *field* in the well-regarded *Collins Spanish Dictionary*.[3]

1. a. (*agriculture*) campo *m*
(= *meadow*) prado *m*
b. (*geology*) yacimiento *m*
2. (*sport*) campo *m*, terreno *m* de juego, cancha *f* (*LAm*)
(= *participants*) participantes *mpl*
(*for post*) opositores *mpl*, candidatos *mpl*
⇒ is there a strong field? ¿se ha presentado gente buena? ⇒ to lead the field (*sport*, *business*) llevar la delantera ⇒ to take the field (*sport*) salir al campo, saltar al terreno de juego
IDIOM: to play the field (*informal*) alternar con cualquiera
3. (= *sphere of activity*) campo *m*, esfera *f* ⇒ field of activity esfera *f* de actividades, campo *m* de acción ⇒ my particular field mi especialidad ⇒ it's not my field no es mi campo *or* especialidad,no es lo mío ⇒ what's your field? ¿qué especialidad tiene Vd? ⇒ in the field of painting en el campo *or* mundo de la pintura ⇒ to be the first in the field ser líder en su campo
4. (= *real environment*) ⇒ a year's trial in the field un año de prueba en el mercado ⇒ to study sth in the field estudiar algo sobre el terreno
5. (*computing*) campo *m*
6. (*military*) campo *m* ⇒ field of battle campo *m* de batalla ⇒ to die in the field morir en combate
7. (*electricity and electronics*) campo *m* ⇒ field of vision campo *m* visual
8. (*heraldry*) campo *m*

Figure 3: The noun field in the Collins Spanish Dictionary.

Although the dictionary does an admirable job of grouping some related senses together, notice that there is no indication that the word campo, which is clearly the main equivalent for field as it appears in all but one of the eight identified senses, evokes land outside where people live, and like field, is a clearing that contrasts with mountains. That is why field is used in expressions like copper field. Fa-

[3] Boldface and color typesetting have been removed from the original in this figure.

ced with this entry, which is typical of bilingual dictionaries in that there is little attempt to link sense developments to one another, the speaker of Spanish may be bewildered by a word that can mean a cultivated area where human intervention is required (campo, with the subject label 'agriculture'), as well as an open space prado 'meadow' and area where a mineral is found naturally underground (yacimiento), all of which are grouped together in sense (1).

5 Conclusion

There is much to do in the analysis of resonance and of agricultural metaphor. This text aims only to show the potential of collocational resonance as a means of showing language change by mapping variations by use of mono- and multilingual lexicographical prototypes and of collocational networks. How dictionaries should incorporate collocational resonance into their descriptions is an open question at this point. For monolingual dictionaries, information about resonance is essential to show linkage across senses and accounts for collocational patterns, yet most contemporary dictionaries do not provide enough information from resonance for users to grasp the linguistic consequences of the metaphor. From a multilingual perspective, usually uncontroversial equivalents such as *field*, *champ*, and *campo*, are shown to develop different patterns of resonance, which—in our view—should have consequences for their representation in bilingual dictionaries.

6 References

Collins Spanish Dictionary Online. Accessed at: http://www.collinsdictionary.com/dictionary/english-spanish [06/03/2014]

Dictionnaire Historique de la Langue Française. (2006). Paris: Dictionnaries Le Robert.

Fillmore, C.J. (1975) An alternative to checklist theories of meaning. In *Papers from the First Annual Meeting of the Berkeley Linguistics Society*, pp. 123–132.

Furetière, Antoine. (1690). *Dictionnaire Universel*. La Haye. Accessed at :http://gallica.bnf.fr/ark:/12148/bpt-6k50614b [09/03/2014]

Hanks, P. 2000. Do Word Meanings Exist. In *Computers and the Humanities* 34, pp.205-215.

Hanks, P. (unpublished 2005). Resonance and the Phraseology of Metaphors. Paper presented at Phraseology 2005: The Many faces of Phraseology Conference. Louvain-la-Neuve.

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

Macmillan English Dictionary Online. Accessed at: http://www.macmillandictionary.com/dictionary [02/04/2014]

Oxford English Dictionary Online. Accessed at: http://www.oed.com [03/03/2014]

Sketch Engine® tenten corpora. Accessed at: https://www.sketchengine.co.uk [10/02/2014]

The American Heritage® Dictionary of the English Language. (2011). Boston: Houghton Mifflin Harcourt.

Tresor de la Langue Française informatisé. http://atilf.atilf.fr [10/03/2014]

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. (2012). Bringing Data *and* Dictionary Together: Real Science in Real Dictionaries. In A. Bolton, S. Thomas, E. Rowley-Jolivet (eds.) *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, pp. 219-240.

Acknowledgements

We gratefully acknowledge that work on this paper was supported by research project FFI2012-38847 *Análisis léxico basado en corpus y su incidencia en los diccionarios* funded by the Spanish Ministry of Economy and Competitiveness (DeCesaris, principal researcher and Williams, team project member).

The Use of Corpora in Bilingual Phraseography

Dmitrij Dobrovol'skij
Russian Academy of Sciences, Russian Language Institute
Austrian Academy of Sciences, AAC-Austrian Academy Corpus
dobrovolskij@gmail.com

Abstract

The present paper discusses issues in the compilation of bilingual dictionaries of idioms based on an analysis of corpus data. The advantages of using corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in the additional possibilities that the corpus materials provide for compiling the idiom list and structuring entries. Thus the corpus allows us to determine the degree of frequency of an expression (at least in the written language). The relevant principles are illustrated by data taken from a new German-Russian dictionary of idioms that is being constructed by an international team of linguists and lexicographers. Fragments of this dictionary are available on the website of the German Language Institute in Mannheim: "Deutsch-russische Idiome online" <http://wvonline.ids-mannheim.de/idiome_russ/index.htm>. Relevant information is also made available via the Europhras homepage on the website <http://www.europhras.org>. All examples of idiom usage in this dictionary are taken from the text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel German-Russian texts from the Russian National Corpus (RNC) are also used.

Keywords: corpus; bilingual lexicography; phraseology; idiom; German; Russian

1 Preliminary Remarks

Bilingual lexicography widely acknowledges the role of phraseology; for a discussion of relevant theoretical issues see (Lubensky & McShane 2007). Considerable work has been done recently on the compilation of bilingual phraseological dictionaries in languages such as English, German, Russian, Czech, Spanish, French, Italian and Portuguese; cf., for instance, (Heřman et al. 2010), (Kraus & Baumgartner 2011) and a series of German bilingual idiom dictionaries initiated and co-compiled by Hans Schemann. The dictionary in this field that is especially remarkable and meets the highest lexicographic standards is (Lubensky 2013). This most complete Russian-English dictionary of idioms first came out in 1995 in New York. It was subsequently published twice in Moscow (in 1996 and in 2004), and now it has appeared in an enlarged and revised version that includes about 550 new entries. (Lubensky 2013) offers virtually the only lexicographic description of Russian phrasemes with their English counterparts that is based on contemporary notions of linguistically significant features of idioms.

Against this background, it seems especially surprising that modern bilingual phraseography scarcely makes use of text corpora. Though Lubensky (2013: vii) points out that the "availability of language corpora made it possible to check the idioms' register and usage in multiple contexts", none of the aforementioned dictionaries is really corpus-based. This fact makes it necessary to address the question as to how corpora can be used as a primary source for compiling a bilingual dictionary of idioms. Today, as lexicography is experiencing "the corpus revolution" (Hanks 2012), this is a question of vital importance. The various uses of corpora in bilingual phraseography will be discussed here on the basis of data taken from a new German-Russian dictionary of idioms that is now under construction.[1]

2 German-Russian Phraseography: State of the Art

The need for a new German-Russian phraseological dictionary is motivated by the fact that existing such dictionaries do not meet present requirements. Both the vocabulary and the examples in Binovič and Grišin's German-Russian phraseological dictionary (Бинович, Гришин 1975) are out of date, and the work fails to satisfy current needs with respect to a number of other parameters as well. Although Dobrovol'skij's *Немецко-русский словарь живых идиом* "German-Russian Dictionary of Current Idioms" (Добровольский 1997) is on the whole more up to date, it also has certain shortcomings. Its idiom list is rather limited, and illustrative examples are often arbitrary and unpersuasive, which may be because it was written back in the "pre-corpus era". Actually, one of the basic goals of our new lexicographical project is to eliminate all the shortcomings of this dictionary and to significantly expand its idiom list.

Yet another dictionary of this type has appeared recently: *Новый немецко-русский фразеологический словарь* "The New German-Russian Phraseological Dictionary" (Шекасюк 2010). Its phraseme list is fairly large and up to date, but the work is difficult to use, primarily because the illustrative examples are not translated into Russian, and the division of entries into meanings and selected equivalents often appears hasty and arbitrary.

Thus there is an unquestionable need for a new dictionary containing the most widely used contemporary German idioms together with carefully selected Russian equivalents, explanations facilitating the correct use of these idioms, and good, authentic examples translated into Russian. It is also important that such a dictionary exist not only in print, but also (at least in part) in an online version,

[1] „Moderne deutsch-russische Idiomatik: Ein Korpus-Wörterbuch", unter der Leitung von Dmitrij Dobrovol'skij. Wissenschaftliche Redaktion: Dmitrij Dobrovol'skij, Artem Šarandin, Irina Parina und Tat'jana Filipenko; erarbeitet von Elena Krotova, Dmitrij Dobrovol'skij, Tat'jana Filipenko, Artem Šarandin, Viktorija Kosteva, Irina Parina und Denis Zaxarov. Russische Akademie der Wissenschaften, Moskau / Österreichische Akademie der Wissenschaften, Wien.

which will not only provide easier access to the information but will also ensure continuous revision and improvement.[2]

3 Corpus-based Bilingual Phraseography and Cross-linguistic Equivalence

The lexicographic treatment of the notion of equivalent in dictionaries based on corpus data encounters certain problems. Not infrequently, the generally accepted equivalent of an idiom cannot always be used to translate authentic texts.

Let us take an example. The German idiom *sich (D) die Beine in den Bauch stehen* (literally ≈ "to stand one's legs into the stomach") has a "standard" equivalent in Russian, namely the expression *отстоять себе все ноги* (literally ≈ "to stand on one's feet as long as they fall off"), both meaning something like 'to stand out' or 'to stand through'*.* It would be somewhat odd to doubt that these expressions are basically equivalent, since they are identical with respect to both their lexicalized meaning and have similar image components. Nevertheless, it turns out that it is far from always possible to translate the idiom *sich (D) die Beine in den Bauch stehen* with the Russian expression *отстоять себе все ноги.* Numerous contexts with the idiom *sich (D) die Beine in den Bauch stehen* can be found in text corpora in which this idiom has to be translated into Russian either by the verb *простоять/простаивать* 'to stand for some time' or by the collocations *стоять в очереди* 'to queue up' and *выстраиваться в длинную очередь* 'to stand in a long queue'.

(1) Schon am Nachmittag *standen sich* die Fans *die Beine in den Bauch***,** um ein Autogramm Ullrichs zu bekommen. Fast 200 Meter lang war die Schlange bis zum Tisch, an dem der Radstar [...] Autogramme schrieb. (Mannheimer Morgen, 28.08.2004)

Уже во второй половине дня фанаты выстроились в длинную очередь, чтобы взять у Ульриха автограф. Очередь к столу, за которым звезда велогонок раздавал автографы, была почти 200 *метров*.

(2) Endlose Warteschlangen winden sich um das Moskauer Puschkin-Museum. Biedere russische Hausfrauen, Veteranen mit Orden am Sonntagsanzug [...], elegante Moskauerinnen – sie *stehen sich* stundenlang *die Beine in den Bauch* für ein paar Blicke auf den Schatz. (Zürcher Tagesanzeiger, 23.04.1996)

Бесконечная очередь вьётся вокруг московского Пушкинского музея. Простые российские домохозяйки, ветераны с орденами на груди, элегантные москвички – все они часами стоят в очереди, чтобы взглянуть на сокровища.

[2] It goes without saying that putting a dictionary online does not automatically mean an easy access to data for its permanent revision. However, an online dictionary provides a better opportunity to improve the entries.

Consequently, despite the intuitively felt equivalence of the expressions *sich (D) die Beine in den Bauch stehen* and *отстоять себе все ноги*, this equivalence cannot be considered complete. For the lexicographer interested in a maximally precise description of the material, such instances are problematical. Either we acknowledge that *sich (D) die Beine in den Bauch stehen* and *отстоять себе все ноги* are equivalent, in which case it is necessary to explain why the "standard" equivalent is unacceptable in a number of contexts, or we deny that a relationship of bilingual equivalence obtains between *sich (D) die Beine in den Bauch stehen* and *отстоять себе все ноги*, and focus exclusively on translating specific contexts. Such a solution, however, is counterintuitive.

There are at least two possibilities to solve this problem. Either we refrain from giving equivalents and replace them with an explanation, or we provide the given equivalents with a commentary indicating relevant limitations.

In our dictionary we have followed the second path. Thus for the German idiom *sich (D) die Beine in den Bauch stehen* we give the Russian equivalent *отстоять себе все ноги* and explain divergences in the use of the idioms in the commentary, where we point to the fact that the Russian idiom *отстоять себе все ноги* is a perfectiva tantum, i.e. it cannot normally be used in the imperfective aspect.

Another example. The German idiom *jmdn. an der Nase herumführen* (cf. English *to lead s.o. (around) by the nose*) is not fully equivalent to its seemingly ideal Russian counterpart *водить за нос кого-л.* because this Russian idiom is an imperfectiva tantum and can be used in the perfective aspect only in non-veridical contexts such as *а народ не дурак, за нос его так просто не проведешь* or *за нос такого провести нетрудно*, which are encountered quite rarely. For more detail see (Dobrovol'skij 2013).

Normally, when used in contexts focusing the result, the German idiom *jmdn. an der Nase herumführen* has to be translated into Russian either by the verbs *надуть* and *одурачить* or by the idiom *обвести вокруг пальца*.

(3) Die Aktionäre fühlen sich vom größten deutschen Industriekonzern *an der Nase herumgeführt*. (Mannheimer Morgen, 08.08.1995)

У акционеров такое чувство, что самый большой промышленный концерн Германии обвел их вокруг пальца.

(4) In Wahrheit hatte er [Wolfgang Schäuble] aber 100.000 Mark [...] bekommen [...]. Und das hat er im Deutschen Bundestag [...] verschwiegen und hat das erst später, vier Wochen später in einem Fernsehinterview aufgedeckt und da haben viele gesagt, [...] der hat den Deutschen Bundestag *an der Nase herumgeführt*. (www.stroebele-online.de/themen/spendenaffaere/29273.html)

На самом деле он [Вольфганг Шойбле] получил 100.000 *марок. Причем он скрыл это от бундестага и только позднее, спустя четыре недели, признался в этом во время телеинтервью. И многие сказали тогда: он просто одурачил немецкий парламент.*

A question that arises from the perspective of phraseological theory (especially its contrastive aspects) concerns the essence of cross-linguistic equivalence of idioms. It seems expedient to distinguish two different aspects of equivalence: (a) equivalence in translation; that is, the relationship between an idiom of language L1 and its translation into language L2 in a particular text, and (b) equivalence in the language system; that is, the relationship between the compared idioms of L1 and L2 on the systemic level.[3]

One of the most important differences between translational and systemic equivalence (besides the fact that the former has to do with a concrete text and the latter with the lexical system) consists in the circumstance that equivalence in translation is a unilateral relationship, whereas equivalence in the language system is defined as bilateral. In other words, if a phraseme of language L1 is equivalent to a phraseme in language L2 (in terms of (b)-equivalence), this means that the L2 phraseme is also equivalent to the corresponding L1 expression. With respect to equivalence in translation, all that is being said is that an expression in language L2 is being used in the translation of some specific text in language L1 in such a way that between the L1 phraseme from this particular text and the L2 expression there is a relationship of semantic correspondence. The fact that the translation of some L1 phraseme into language L2 is its equivalent (at least with respect to this particular context) does not, of course, mean that the relationship can be reversed. That is, the L1 phraseme should not be regarded as an equivalent of the expression used in the translation of this phraseme into language L2 (even if this expression is a phraseme, which is not at all obligatory). Obviously, the study of equivalence in translation broadens our notions about the possibilities of cross-linguistic paraphrasing and about the role of contextual conditions in the selection of adequate correspondences, and it contributes to the development of both translation theory and contrastive phraseology.

As for equivalence in the language system, its study has both theoretical and practical significance for phraseology. Deserving of special attention from the theoretical point of view is the question of why one and the same concept is expressed by means of an idiom in one language but not in another. Another (no less important) problem concerns the fact that between basically similar idioms in language L1 and language L2, there are practically always certain semantic, pragmatic, and collocational differences that must be discovered and described. This is especially important in cases where a traditional description postulates a relationship of "full equivalence" but ignores the absence of functional interchangeability between the idioms. The practical aspect of systemic equivalence is what is reflected in bilingual dictionaries, where the entry consists of a phraseme of language L1 (in the lemma) and its idiomatic (to the extent this is possible) correlates in L2. Can these correlates be regarded as equivalents of the L1 phraseme? Yes and no. On the one hand, they must be at least "partial equivalents", for otherwise they could not be placed in the corresponding dictionary entry. On the other, often they cannot be used in the translation of specific texts. The reason, as a rule, is that the phrasemes of L1 and L2 display certain differences in their semantic, pragmatic, and collocational features. They can be considered cross-linguistic equivalents only in a rather approximate comparison of the idioms of the given languages, and are the starting point of a thorough contrastive analysis that attempts to

[3] That (a) and (b) represent different aspects of the equivalence phenomenon has been noted in various theoretical contexts. For example, Zgusta distinguishes between *explanatory* or *descriptive* and *translational* or *insertable* equivalents, Hausmann between *prototypical* and *textual* equivalents, and Gouws between *semantic* and *communicative* ones. For more detail see (Adamska-Sałaciak 2010: 392-397).

discover the unique properties of each idiom and thereby improve the lexicological and lexicographical description of phraseology.

Obviously, aspects (a) and (b) are, as it were, two sides of the same phenomenon or two approaches to studying it. We assume that one of the principal goals of contrastive phraseology is to discover genuine equivalents – that is, those that are as close as possible with respect to their actual meanings and – ideally – with respect to the image basis of the expressions, and that function equally well in analogous types of situations, which does not at all imply an obligatory "phraseme – phraseme" relationship. What is important for cross-linguistic correspondence, after all, is not "phraseologicalness," but functional equivalence.[4] It is this type of equivalence that is most interesting from the perspective of bilingual lexicography. To find out functional equivalents we have to simultaneously go two ways: from text to language system and from language system to text. On the one hand, not all systemic equivalents can function as counterparts in authentic texts and, on the other, not all translational equivalents can be included in the dictionary as typical parallels suitable for using in neutral contexts.

In contrast to a conception that is wide-spread within traditional phraseology, I claim that lexical units of any kind (i.e., not only idioms) in L2 which have the identical meaning and, in the ideal case, near-identical metaphorical basis as the L1-idioms from the source text are excellent functional equivalents, so they have to be considered not only more or less appropriate translational solutions, but also real functional equivalents, i.e. parallels in the lexicons of L1 and L2, which have to be fixed lexicographically.

4 Parameters of the new Corpus-based Dictionary

The basic parameters upon which dictionaries can be described and compared are (i) the word list (in our case, the idiom list), (ii) the corpus of illustrative examples, (iii) the macrostructure, and (iv) the microstructure, that is, the structure of the entries. Each of these parameters is briefly described below.

[4] I consider functional equivalents to be a kind of compromise between the translational and systemic approaches, i.e. functional equivalents are lexical items that on the one hand, semantically resemble each other as closely as possible, i.e. are intuitively felt similar in a contextless, isolated presentation, and, on the other, mostly can be used in similar situations. Thus, my interpretation of functional equivalence differs from, for example, Zgusta's approach. Zgusta (1984: 151) points out that "a translation should convey to its reader the same message with the same aesthetic and other values which are conveyed by the original text. Since languages differ in all imaginable respects, the translator-lexicographer must sometimes use means quite different from those used in the original in order to obtain the same results. If the different means do produce the same effect, the texts are considered functionally equivalent".

4.1 The Idiom List

The idiom list of our new dictionary is based primarily on that of Dobrovol'skij's *Немецко-русский словарь живых идиом* "German-Russian Dictionary of Current Idioms" (Добровольский 1997), which contains in all about 1000 items. While working on the monograph (Dobrovol'skij 1997), I conducted a detailed survey in which informants were asked to take into account not only the units that they felt were widely used in contemporary speech, but also those that were judged to be generally known although not necessarily used. In other words, a distinction was drawn between passive and active command of the phraseology. Combining these two idiom lists resulted in a new, expanded idiom list that was supplemented in the course of working with the corpora. At present our idiom list contains some 2000 idioms with variants. There is reason to believe that it covers a majority of commonly used and most familiar idioms of the contemporary German literary language.

Vulgar expressions were deliberately excluded, since such idioms are ill suited for active use by non-native speakers of German. Since the dictionary aspires to a certain extent to be active, its idiom list focuses not so much on understanding as on use.

4.2 The Body of Illustrative Examples

The basic difference between the present dictionary and traditional ones is that all examples of idiom usage in it are taken from the text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel texts from the Russian National Corpus (RNC) are also used. These examples are especially valuable because they have been translated by professional translators rather than by the authors and editors of the dictionary. Since this part of the parallel corpus of the RNC is still rather modest in size, however, examples needed for the dictionary were rarely encountered.

The use of authentic examples based on text corpora is a new approach in bilingual lexicography. Traditional dictionaries were based on a limited body of generally randomly selected examples, and the use of the idioms was often not even exemplified. The advantages of using corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in the additional possibilities that the corpus materials provide for compiling the idiom list and structuring entries. Thus the corpus allows us to determine the degree of frequency of an expression (at least in the written language). For example, the expression *ich fresse einen Besen* occurred in DeReKo 60 times, *Blech reden* 128 times, *bei Adam und Eva anfangen [beginnen]* 236 times, *jmdm. um den Bart gehen* 41 times, *Gift und Galle spucken [speien...]* 312 times, and *bittere Pille* 2804 times. The lower occurrence threshold for an expression to be included in the idiom list can be set differently for different dictionaries. The important point is that together with surveys of informants, the lexicographer now has a supplemental resource for determining the frequency of each individual idiom.

Yet another advantage of using corpora is that it increases our ability to determine the peculiarities of the formal and semantic structure of idioms, particularly in the description of the ambiguity and variation of a form. Although an analysis of examples of use clearly indicates that polysemy in phraseology is an extremely widespread phenomenon (for further detail see Dobrovol'skij & Filipenko 2009), traditional dictionaries rarely distinguish the different meanings of idioms, and seldom reflect the full diversity of variants actually represented in texts. Dictionaries often register only a single "canonized" form of an idiom that in many cases proves to be not the most frequent one.

In a number of instances text corpora allow us not only to determine the form of a lemma and a selection of its most frequent variants, but also to establish whether a given expression belongs to the sphere of phraseology. For example, Duden 11 (2002) cites two synonymous idioms with the verb *abberufen* in the passive: *abberufen werden: in die Ewigkeit abberufen werden* and *aus dem Leben abberufen werden.* The following synonymous expressions with this verb form are given in DeReKo: *aus dem Leben abberufen werden, zur großen Armee abberufen werden, in die Ewigkeit abberufen werden, ins Jenseits abberufen werden, in die ewigen Jagdgründe abberufen werden, in die ewige Heimat abberufen werden, von/aus dieser Welt abberufen werden, aus diesem irdischen Leben abberufen werden, aus unseren Reihen [aus unserer Mitte] abberufen werden, zu den Scharen der Engel abberufen werden, in eine andere Welt abberufen werden, in den ewigen Frieden abberufen werden, in ein besseres Jenseits abberufen werden, für uns alle viel zu früh abberufen werden, vom Schöpfer abberufen werden, von Gott (dem Herrn) abberufen werden, vom Tod (ins Jenseits) abberufen werden, von einem gnädigen Tod abberufen werden.* There are also expressions close in meaning in which the verb *abberufen* is used in the active voice: *jmdn. will Gott abberufen, jmdn. hat der Tod abberufen.* These findings suggest that the sense of "calling/summoning s.o. from life" is simply a metaphorical meaning of the verb *abberufen*. Consequently, what we have to do with here is not an idiom but a series of relatively free collocations based on a metaphor.

Another example. Duden 11 (2002) cites four idioms with the noun *Mundwerk*: *jmds. Mundwerk steht nicht still* (ugs.) 'jmd. redet ununterbrochen'; *ein böses/lockeres/loses/freches* o.ä. *Mundwerk haben* (ugs.) 'gehässig/vorlaut/frech o.ä. reden'; *ein gutes/flinkes Mundwerk haben* (ugs.) 'sehr gewandt reden'; *ein großes Mundwerk haben* (ugs.) 'großsprecherisch reden'. Corpus analysis has shown that the noun *Mundwerk* has a much broader combinatorial profile. Compare, e.g., *flottes, vorlautes, geschliffenes Mundwerk*. This noun can also be used without any adjectives, combining with verbs of various meanings. Cf.

(5) Manchmal wäre es vielleicht sinnvoller, mein *Mundwerk* etwas zu zügeln, nach dem Motto „Reden ist Silber, Schweigen ist Gold". (St. Galler Tagblatt, 08.04.1999)

Hence, we are dealing here not with four idioms but with a free noun. It seems that the combinatorial profile of this noun is relatively restricted. The only way to describe the collocational constraints in question is the consistent analysis of corpus data.

4.3 Dictionary Macrostructure

The dictionary has two parts: the *body*, consisting of entries listed alphabetically by headword, and the *index*, which makes it possible to find an idiom from any of its constituents.

The idioms are arranged alphabetically by headword, selected according to the following hierarchy:

- nouns
- adjectives (including adjectivized participles)
- adverbs (including adjectives in adverbial position and adverbalized participles)
- numerals
- verbs
- particles (with the exception of the negative particle *nicht*)
- pronouns (with the exception of the reflexive pronoun *sich*)
- prepositions
- conjunctions
- interjections

The order of this hierarchy is motivated by the variation features of the lexical structure of the idiom. Thus the verb can often be replaced by a synonym (or more rarely by an antonym), whereas adjectives and adverbs are more stable elements of the structure, and it is this that accounts for their higher position in the hierarchy. Adjectives and adjectivized participles, in turn, are more stable than adverbs. For example, cf. the structurally and semantically similar idioms *es ist (nicht) gut bestellt* (*um jmdn., etw. A*) = *дела обстоят (не очень) хорошо (с чем-л. / у кого-л.) и es ist (nicht so) schlecht bestellt* (*um A*) = *дела обстоят (не так) плохо (с чем-л. / у кого-л.*). Alphabetizing them according to the adverbial constituent would necessitate entering them in different parts of the dictionary (under GUT and under SCHLECHT, respectively), which is counterintuitive. Alphabetization according to the constituent BESTELLT, which is an adjectivized participle, is much more convenient for the user.

An example of a group of idioms based on the headword under which they are arranged is given in the Appendix.

4.4 Dictionary Entry Structure

Dictionary entries open with the *headword*, i.e. the word on which alphabetization is based. This (word, if it is a noun) is always given in the nominative singular (e.g. KOPF), even if the idioms following the headword contain forms such as *Kopfes, Köpfe, Köpfen* etc. The headword is followed by the *lemma* – the idiom in traditional dictionary form (nominative for nominal expressions, the infinitive with valencies for verbal ones).

Idioms in *propositional* (or *personal*) *form* are indicated in cases where the subjective valency is filled in a non-trivial way or when the infinitive of the idiom translates poorly into Russian. Compare, e.g., **ei-**

nen dicken Kopf haben (*von etw. D*): (*jmd.*) hat (*von etw. D*) einen dicken Kopf = *(у кого-л.) голова болит (из-за чего-л.), (у кого-л.) голова раскалывается (от чего-л., после чего-л. – особенно с похмелья)*. The propositional form often helps to discriminate the senses of a polysemous idiom; cf. **jenseits von Gut und Böse sein: 1.** (*jmd.*) ist jenseits von Gut und Böse = (*кто-л.*) *чужд плотским удовольствиям (часто о пожилых людях)* **2.** (*jmd.*) ist jenseits von Gut und Böse = *(кто-л.) не от мира сего, (кто-л.) потерял связь с реальностью, (кто-л.) неадекватен (часто о людях, находящихся в состоянии сильного алкогольного опьянения)* **3.** (*jmd.*) ist jenseits von Gut und Böse = *(кто-л.) по ту сторону добра и зла* **4.** (*etw.*) ist jenseits von Gut und Böse = *(что-л.) выходит за привычные рамки, (что-л.) невероятно (что-л. очень хорошо либо очень плохо; часто о слишком высоких либо слишком низких ценах).*

The lemma or propositional form (if there is one) is followed by stylistic *labels*. The use of which follows the principles set forth in (Баранов, Добровольский 2008). Thus the label *разг.* (colloquial) is not used at all, since most idioms belong to the colloquial register. In other words, this label "works" by remaining silent. The following labels are used: *высок.* (high style) – for high style expressions, *книжн.* (literary) – for literary and bookish expressions, *офиц.* (formal) – for expressions in official language and business communication, *нейтр.* (neutral) – for expressions in the neutral register (that is, for idioms higher than colloquial expressions on the scale of stylistic registers), and *снижен.* (≈ very informal) – for idioms felt to be not entirely acceptable in the standard colloquial style (i.e. lower than *разг.*).

The *translation* of the idiom into Russian is generally oriented toward the system of the language, i.e. toward (b)-equivalents, rather than toward contextual conditions. Relevant functional and context-sensitive properties are additionally explained and illustrated in other parts of the entry, mostly in the commentary and illustrative field. That is, if in the examples of usage an idiom is translated in a non-standard manner, this does not mean that these – often unique – ways of translating it must be registered in the translation field. There it is often expedient to indicate several equivalents, first of all, those translations that with respect to their actual meaning and image basis maximally approximate the German idiom being described. The syntactic parallelism of suggested equivalents is also taken into account as far as possible. If an equivalent parallel to the lemma cannot be found or if it sounds strange, what is recorded in the field of the propositional form is the syntactic version of the German expression that would best correspond to the suggested Russian translation. The translation field can also contain explanatory commentaries that further indicate in which of the possible meanings the suggested Russian translation is equivalent to the German expression.

The *variant field* follows the translation field. Describing variants in a separate field makes it possible not only to reflect more completely the actual variation of the structure of the idiom, but also to avoid having to burden the notation of the lemma with a series of parentheses.

As for selecting *illustrative examples*, preference is given to modern examples, that is, to contexts with idioms dating from the past fifteen years. The basic source for illustrative examples is the corpus of the Mannheim Institute of the German Language (DeReKo). For more detail see section 4.2. In selecting illustrative examples, we have tried not to include examples peculiar to Austrian or Swiss usage, since these deviate from standard literary German (and due to their regional and cultural distinctiveness) do not fully satisfy the needs of a bilingual dictionary with educational goals.[5]

The search for contexts relies not only on the standard options but also on so called "co-occurrence analysis" (Kookkurrenzanalyse). This program helps to determine the lexical contexts in which a given idiom occurs especially often.

All contexts are given in the current (i.e. the "new") orthography. The peculiarities of Swiss orthography (for example, *ss* instead of the normative *ß*) are not preserved. Such deviations from prevailing standards are given in conformity with the spelling norms of the common German language. For the sake of convenience in using the dictionary, the authors have simplified extremely complex and verbose contexts. Deletions in abbreviated contexts are marked by [...]; cf. examples (1), (2), and (4) above. This indication is not repeated in the Russian translations. Contexts that are overloaded with specific information that is not relevant to conditions for using a given idiom are slightly modified. For example, unfamiliar proper names are replaced with neutral designations of the participants of a situation. In such cases the source is indicated (in parentheses immediately following the context) by *Nach:*. Compare examples in the Appendix.

The *commentary field* contains information significant for the correct use of the expressions if such information cannot be derived from the valency model and/or the semantic and syntactic features of the Russian equivalents. The commentary field indicates, for example, the syntactic and combinatorial properties of the idiom. Also reflected in the commentaries are features relating to the polarization (especially the negative polarity) of expressions, their aspectological peculiarities, possibilities of nominalization, characteristic metonymical shifts, etc., as well as any significant transformational properties of the idiom, especially if they do not coincide with the syntax of the Russian equivalent. Thus the idiom *Blech reden* (unlike its Russian equivalents *пустословить, нести чепуху, болтать языком)* can be passivized. The commentary field has no fixed position in the structure of the dictionary entry. Accordingly, it can be located in any part of the dictionary entry (depending on the nature of the information being provided).

[5] This does not mean that Austrian and Swiss sources of empirical data were excluded.

5 Conclusion

The use of corpora clearly expands the resources available to the lexicographer for creating the illustrative component of the dictionary entry, but it also offers a number of additional possibilities. Let us attempt to list the most obvious such advantages. Working with corpora makes it possible

- to determine the frequency of each idiom included in the dictionary;
- to determine whether a particular word group is an idiom;
- to determine the standard form of a lemma from the point of view of modern usage;
- to clarify the government models of relevant idioms;
- to determine the most significant variants of each idiom;
- to determine the polysemy structure of each idiom and refine the description of its concrete meanings;
- on the basis of corpus materials, to select the most adequate correspondences, including translations of concrete examples, for each meaning of an idiom;
- to describe the typical modifications of the structure of each idiom;
- to determine the typical environment of the idioms being described and the types of contexts in which they are perceived to be most natural.

Literature on the subject distinguishes two approaches to the use of corpora in lexicographical research: corpus-based and corpus-driven.[6] In the first approach, corpus data are used to confirm already existing hypotheses, while in the second it is the corpus itself that constitutes the data about linguistic structures, and it is only later that these data are interpreted by the linguist. It is clear that on the whole, lexicographers use corpora as the source of additional information about already given linguistic forms (that is, the corpus-based approach). As the material discussed in this paper shows, however, lexicographical work also presumes elements of the corpus-driven paradigm. In other words, in a number of cases corpus data provide lexicographers with knowledge about the structure and semantics of idioms to which they would not have had access even on the hypothetical level prior to consulting the corpus.

6 References

Adamska-Sałaciak, A. (2010). Examining equivalence. In *International Journal of Lexicography*, 21(4), pp. 387-409.

Dobrovol'skij, D. (1997). Idiome im mentalen Lexikon: Ziele und Methoden der kognitivbasierten Phraseologieforschung. Trier: WVT Wissenschaftlicher Verlag Trier.

[6] Cf., first of all (Tognini-Bonelli 2001), where this distinction is discussed. However, "good corpus research almost always uses both" (Kilgarriff 2013: 96).

Dobrovol'skij, D. (2013). German-Russian phraseography: On a new dictionary of modern idiomatics. In I. Gonzáles Rey (ed.) Phraseodidactic studies on German as a foreign language. Hamburg: Verlag Dr. Kovač, pp. 121-138.

Dobrovol'skij, D.O. & Filipenko, T.V. (2009). Polysemie in der Idiomatik. In C. Földes (ed.) Phraseologie disziplinär und interdisziplinär. Tübingen: Gunter Narr, pp. 109-115.

Duden 11 (2002) = *Duden – Redewendungen. Wörterbuch der deutschen Idiomatik.* 2., neu bearb. und aktualisierte Auflage. (=Der Duden, Band 11). Mannheim etc.

Hanks, P. (2012). The corpus revolution in lexicography. In *International Journal of Lexicography*, 25(4), pp. 398-436.

Heřman, K. et al. (2010). Deutsch-tschechisches Wörterbuch der Phraseologismen und festgeprägten Wendungen. Prag: C. H. Beck.

Kilgarriff, A. (2013). Review of Tony McEnery & Andrew Hardie. Corpus linguistics: Method, theory and practice. In *International Journal of Lexicography*, 22(1), pp. 95-97.

Kraus, R. & Baumgartner, P. (eds.) (2011). Phraseological Dictionary English-German: General Vocabulary in Technical and Scientific Texts. Berlin & Heidelberg: Springer.

Lubensky, S. (2013). Russian-English dictionary of idioms. Revised Edition. New Haven & London: Yale University Press.

Lubensky, S. & McShane, M. (2007). Bilingual phraseological dictionaries. In H. Burger, D. Dobrovol'skij, P. Kühn and N.R. Norrick (eds.) Phraseology: An international handbook of contemporary research. Vol. 2. Berlin & New York: Walter de Gruyter, pp. 919-928.

Schemann, H et al. (2013). *Idiomatik Deutsch-Spanisch.* Hamburg: Buske.

Schemann, H et al. (2012). *Idiomatik Deutsch-Portugiesisch.* 2., durchgesehene Auflage. Hamburg: Buske.

Schemann, H & Dias, I. (2013). *Idiomatik Portugiesisch-Deutsch.* 2., durchgesehene Auflage. Hamburg: Buske.

Schemann, H., Fenati, B. & Rovere, G. (2011). *Idiomatik Deutsch-Italienisch.* 2., durchgesehene Auflage. Hamburg: Buske.

Schemann, H. & Knight, P. (2011). *Idiomatik Deutsch-Englisch.* 2., durchgesehene Auflage. Hamburg: Buske.

Schemann, H. & Raymond, A. (2011). *Idiomatik Deutsch-Französisch.* 2., durchgesehene Auflage. Hamburg: Buske.

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

Zgusta, L. (1984). Translational equivalence and the bilingual dictionary. In R.R.K. Hartmann (ed.) LEXeter'83 Proceedings. Tübingen: Max Niemeyer, pp. 147-154.

Баранов, А.Н. & Добровольский, Д.О. [Baranov, A.N. & Dobrovol'skij, D.O.] (2008). *Аспекты теории фразеологии* [Aspects of the Theory of Phraseology]. *Москва: Знак.*

Бинович, Л.Э. & Гришин, Н.Н. [Binovič, L.E. & Grišin, N.N.] (1975). *Немецко-русский фразеологический словарь* [German-Russian Phraseological Dictionary]. *Москва: Русский язык.*

Добровольский, Д.О. [Dobrovol'skij, D.O.] (1997). *Немецко-русский словарь живых идиом* [German-Russian Dictionary of Current Idioms]. *Москва: Метатекст.*

Шекасюк, Б.П. [Šekasjuk, B.P.] (2010): *Новый немецко-русский фразеологический словарь* [The New German-Russian Phraseological Dictionary]. *Изд. 2-е, перераб. и доп. Москва: Либроком.*

7 Digital Resources

DeReKo – Das Deutsche Referenzkorpus des IDS Mannheim im Portal COSMAS II (Corpus Search, Management and Analysis System) <https://cosmas2.ids-mannheim.de/cosmas2-web>

DWDS – Corpora des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts <http://www.dwds.de>

Online dictionary "Deutsch-russische Idiome online" <http://wvonline.ids-mannheim.de/idiome_russ/index.htm>

RNC (НКРЯ) – Russian National Corpus (Национальный корпус русского языка) http://www.ruscorpora.ru

Acknowledgements

This paper is based on work supported by the RGNF under Grant 12-04-12041, by the RFFI under Grant 13-06-00403 and by the Basic Research Program "Corpus Linguistics" of the Presidium of the Russian Academy of Sciences.

Appendix

Licht

Licht bringen (*in etw. A*)

нейтр.

внести ясность *(во что-л.);* прояснить *(что-л.);* пролить свет *(на что-л.)*

🗐 Archäologen *haben* endlich *Licht* in einen Abschnitt der Geschichte Londons *gebracht*, der vom Abzug der Römer aus Britannien 410 bis ins Mittelalter reicht. (Berliner Morgenpost, 09.09.1999)

Археологам наконец-то удалось *пролить свет* на период истории Лондона с момента ухода римлян из Британии в 410 г. и до Средневековья.

🗐 Die Aussage eines 37 Jahre alten Beifahrers in einem anderen Auto *hatte Licht* in die zunächst rätselhafte Kollision *gebracht*. (Frankfurter Rundschau, 26.03.1999)

Показания 37-летнего пассажира, сидевшего рядом с водителем в другом автомобиле, *прояснили* это сперва казавшееся загадочным столкновение.

🗐 Auf jeden Fall sind die Funde aus Engers sehr wichtig, um *Licht* in das für Historiker „dunkle fünfte Jahrhundert" zu *bringen*. (Rhein-Zeitung, 25.05.2007)

В любом случае, находки в Энгерсе очень важны для историков, поскольку могут *пролить свет* на «тёмный пятый век».

das Licht der Welt erblicken: (*jmd., etw.*) erblickt das Licht der Welt

высок.

(*кто-л.*) появился на свет, (*кто-л.*) родился; (*что-л.*) родилось *(напр. об изобретениях); (что-л.)* увидело свет

🗐 In Deutschland *erblicken* pro Jahr rund 60 000 Säuglinge vor der 37. Schwangerschaftswoche *das Licht der Welt* – Tendenz steigend. (Rhein-Zeitung, 09.11.2011)

В Германии за год более 60 000 детей *появляются на свет* до 37-ой недели беременности, и эта тенденция растёт.

🗐 Die Anzahl seiner Geburtagspartys hält sich in überschaubaren Grenzen, denn er *erblickte* an einem 29. Februar *das Licht der Welt*. (Hamburger Morgenpost, 08.02.2009)

Ему не так часто доводилось устраивать вечеринки в честь своего дня рождения, потому что он *появился на свет* 29-го февраля.

🗐 1995 war auch das Geburtsjahr des Internet-Dienstes Yahoo und Ende 1998 *erblickte* die Suchmaschine Google in einer Garage *das Licht der Welt*. (Hamburger Morgenpost, 15.03.2009)

В 1995 году появился ещё и Интернет-сервис «Yahoo», а в конце 1998-го в гараже *была рождена* поисковая система «Гугл».

ein Licht geht auf (*jmdm.*)

глаза открылись *(у кого-л.); (кто-л.)* догадался

🕮 Идиома часто употребляется с наречиями *plötzlich, endlich, langsam* – перевод соответственно модифицируется.

🗐 Die italienische Schauspielerin Gina Lollobrigida kommt nach Nürnberg! *Langsam ging* den Verehrern des Filmstars *ein Licht auf*: April! April! (Nach: Nürnberger Zeitung, 29.04.2010)

Итальянская актриса Джина Лоллобриджида приезжает в Нюрнберг! *Постепенно* до поклонников кинозвезды дошло, что это была первоапрельская шутка.

🗐 Aber nachdem ich Hunderte von Interviews mit Trainern gelesen und gehört habe, *ist mir endlich ein Licht aufgegangen*: Ich bin eigentlich gar kein Arbeitnehmer oder Kunde, ich bin ein Helfer. (Nach: Rhein-Zeitung, 28.02.2002)

Но после того как я прочитал и прослушал сотни интервью с тренерами, я *наконец-то понял:* я никакой не наёмный рабочий и не клиент, я помощник.

🕮 Идиома может употребляться с валентностью *über etw. A* в значении ‘(*кто-л.*) осознал (*что-л.*)’.

🗐 Und *geht* der Mannschaft endlich *ein Licht auf* über den Ernst der Lage? (Hannoversche Allgemeine, 12.03.2009)

И *осознает* ли команда наконец всю серьёзность положения?

grünes Licht geben (*für etw. A*)

нейтр.

дать зелёный свет *(для чего-л.)*, дать разрешение *(на что-л.)*

🗐 Das Bauamt *hat grünes Licht* für einen Anbau ans Museum *gegeben*. Dort soll das Archiv untergebracht werden. (Rhein-Zeitung, 08.09.2011)

Строительное ведомство *дало разрешение* на постройку флигеля музея. Там будет располагаться архив.

🗐 Das Arbeitsgericht Frankfurt *hat* im Tarifstreit beim Flugzeugbauer Airbus *grünes Licht* für Warnstreiks *gegeben*. (Rhein-Zeitung, 01.10.2011)

Суд по трудовым спорам Франкфурта *дал зелёный свет* на предупредительную забастовку рабочих самолётостроительного концерна «Эйрбас» из-за тарифного конфликта.

🗐 Die Suche nach der seit Jahrzehnten verschwundenen Leiche von Lolita Brieger auf einer ehemaligen Mülldeponie in der Eifel kann beginnen. Experten *gaben grünes Licht*. (Rhein-Zeitung, 05.10.2011)

Поиски трупа Лолиты Бригер, исчезнувшей уже десятки лет назад, на бывшей мусорной свалке в Эйфеле, могут начаться. Эксперты *дали зелёный свет.*

🗐 Ben Bernanke (56) kann aufatmen. Er bekommt eine zweite Amtszeit als US-Notenbankchef. Der mächtige Bankenausschuss des Senats *gab* dafür *grünes Licht*. (Hamburger Morgenpost, 18.12.2009)

56-летний Бен Бернанке может вздохнуть с облегчением. Он во второй раз получил должность директора эмиссионного банка США. Вчера влиятельная банковская комиссия сената дала соответствующее *разрешение.*

🗐 Bis 2012 hat der Iran die Atombombe, glauben viele Beobachter. „Dann ist es zu spät“, lautet das Credo des israelischen Premiers Netanjahu. Er drängt auf einen Angriff. Von George Bush *gab* es dafür *grünes Licht*, von Obama nicht. (Hamburger Morgenpost, 14.04.2010)

К 2012 году у Ирана появится атомная бомба, полагают многие наблюдатели. «Тогда будет слишком поздно», – уверен премьер министр Израиля Нетаньяху. Он настаивает на нападении. Джорждж Буш *давал* ему на это *зелёный свет*, а Обама – нет.

🗐 Warschau. Ein Gericht in Polen *gab grünes Licht* für die Auslieferung eines mutmaßlichen Agenten des israelischen Geheimdienstes Mossad an Deutschland. (Hamburger Morgenpost, 08.07.2010)

Польский суд *дал зелёный свет* на экстрадицию в Германию предполагаемого агента израильской разведки «Моссад».

🕮 Возможна атрибутивная модификация.

🗐 Für entsprechende Projekte in Hammelburg, Freising, Kempten und Passau *gab* Wirtschaftsminister Otto Wiesheu jetzt *das lang ersehnte grüne Licht.* (Nürnberger Nachrichten, 10.08.1994)
Министр экономики Отто Визхой *наконец-то дал разрешение* на соответствующие проекты в Хаммелбурге, Фрайзинге, Кэмптене и Пассау.

in rosigem Licht
нейтр.
в розовом [радужном] свете
⚐ редко i**n rosa(rotem) Licht**
🗐 Frankfurt sieht nach zehn trüben Jahren plötzlich die eigene Zukunft *in rosigem Licht.* (Nach: Nürnberger Nachrichten, 20.02.2001)
После десяти мрачных лет будущее Франкфурта неожиданно предстаёт в *розовом свете.*
🕮 Прилагательное *rosig* может употребляться с неопределённым артиклем (*in einem rosigen Licht*). Возможны также сравнительная или превосходная степени прилагательного (*in einem rosigeren Licht, im rosigsten Licht*).
🗐 Ich war 13 oder 14 Jahre alt, als wir in der Schule einen Aufsatz über unsere Zukunftspläne und Lebensziele schreiben mussten. Aufsätze schrieb ich gern, und trotz aller Nachkriegseinschränkungen sah ich die Zukunft *in einem rosigen Licht.* (Nach: Mannheimer Morgen, 27.12.1997)
Мне было лет 13-14, когда нам в школе задали написать сочинение о планах на будущее и целях в жизни. Я любил писать сочинения, а будущее, несмотря на все тяготы послевоенного времени, видел в *розовом свете.*
🗐 Die Pannenserie der vergangenen Monate an Bord der Mir hat paradoxerweise dazu beigetragen, dass die Zukunft der russischen Raumfahrt heute wieder *in einem rosigeren Licht* erscheint. (Nürnberger Nachrichten, 25.08.1997)
Как ни парадоксально, но серия поломок на орбитальной станции «Мир» за последние месяцы способствовала тому, что сегодня будущее российской космонавтики снова кажется *более радужным.*
🕮 Прилагательное в составе идиомы может модифицироваться с помощью указательных местоимений, адвербиалов и пр.
🗐 Warum ist er nicht Lehrer geworden, wenn er die materiellen Vorteile des Lehrerberufes *in solch herrlich rosarotem Licht* sieht? (Frankfurter Rundschau, 23.07.1997)
Почему же он сам не стал учителем, если материальные выгоды этой профессии ему видятся *в столь прекрасном розовом свете?*
🕮 В контекстах с отрицанием идиома употребляется в форме *in keinem rosigen Licht* или *nicht in rosigem Licht.*
🗐 Beschwerden von Nachbarn über zu laute Musik und die anhaltende Ebbe im Stadtsäckel lassen die Zukunft des Jugendkulturhauses „Schillers“ *in keinem rosigen Licht* erscheinen. (Nach: Mannheimer Morgen, 19.03.2010)
Из-за жалоб соседей на слишком громкую музыку и постоянного недостатка в городской казне будущее молодёжного дома культуры «Шиллерс» видится отнюдь *не в розовом свете.*
🗐 Der scheidende Privatsekretär der Queen wollte zu Charles Entsetzen ein weiteres Diana-Buch genehmigen, das die Verstorbene *nicht* gerade *in rosigem Licht darstellt.* (Berliner Morgenpost, 02.11.1998)
Увольняющийся личный секретарь королевы хотел к ужасу Чарльза дать согласие на публикацию ещё одной книги о Диане, в которой покойная предстаёт отнюдь *не в розовом свете.*
🕮 Идиома может также употребляться в форме *(etw. A) in (ein) rosiges Licht tauchen [stellen, rücken ...].*

🗐 Bei den Neujahrsansprachen haben Politiker und Unternehmen die Wirtschaftswirklichkeit *in rosiges Licht* getaucht. Von drei Prozent Wachstum und der Wende am Arbeitsmarkt war da die Rede. (Rhein-Zeitung, 07.01.1998)
В своих новогодних поздравлениях политики и бизнесмены *представили* реальную экономическую ситуацию страны *в розовом свете.* Речь шла тогда о трёх процентах роста и переменах на рынке труда.

Tuch
ein rotes Tuch sein (*für jmdn.*)
действовать как красная тряпка на быка *(на кого-л.)*
⚐ **wie ein rotes Tuch wirken** (*auf jmdn.*)
🗐 Die Steuererklärung *ist* für fast jeden *ein rotes Tuch.* Viele sind schon in Hektik, denn der Termin rückt jetzt langsam näher: Spätestens am 31. Mai muss die Steuererklärung abgegeben werden. (Hamburger Morgenpost, 26.03.2010)
Налоговая декларация *действует* почти на каждого *как красная тряпка на быка.* Многие уже в панике, так как срок медленно, но верно приближается: декларация должна быть сдана не позднее 31 мая.
🗐 Als Außenminister Guido Westerwelle seinen Antrittsbesuch in Warschau machte, wusste er sehr genau, dass Frau Steinbach *ein rotes Tuch* in Polen *ist.* Die CDU-Abgeordnete hatte 1991 im Bundestag nicht für die Oder-Neiße-Grenze gestimmt. (Nach: Nürnberger Nachrichten, 12.02.2010)
Когда министр иностранных дел Гидо Вестервелле Германии совершил свой первый визит в этой должности в Варшаву, он знал очень хорошо, что госпожа Штайнбах *действует* на поляков *как красная тряпка на быка.* Будучи депутатом от партии ХДС в бундестаге, она не проголосовала в 1991 году за границу по Одеру-Нейсе.
🗐 Die Sicherheitskonferenz bleibt eine kitzlige Angelegenheit für die bayerische Polizei, zumal unter den Gästen auch einige sind, die auf die Gegner *wie ein rotes Tuch wirken.* (Nürnberger Zeitung, 11.02.2005)
Конференция по безопасности остается весьма щекотливым мероприятием для баварской полиции, тем более что среди гостей есть такие, *которые* действуют на своих оппонетов *как красная тряпка на быка.*

in trockene Tücher bringen (*etw. A*)
⚐ **in trockene Tücher bekommen** (*etw. A*)
окончательно оговорить [согласовать] *(что-л.)*
🗐 Der Wechsel des Holländers von Real Madrid nach München soll kurz vor dem Abschluss stehen! Die Bayern und Real wollen den Transfer heute *in trockene Tücher bringen.* Nur noch Details seien zu klären. (Hamburger Morgenpost, 27.08.2009)
Переход голландского игрока из мадридского «Реала» в мюнхенский клуб уже практически завершён! «Бавария» и «Реал» хотят *окончательно оговорить* трансфер сегодня. Осталось только согласовать детали.
🗐 Seehofer warnte vor einem Scheitern der Verhandlungen. Zur Not müsse die Ministerrunde die ganze Nacht zum Mittwoch verhandeln, um die Reform *in trockene Tücher* zu *bekommen.* Würde das Problem in das nächste Jahr verschoben, wäre eine Lösung noch schwieriger als jetzt. (dpa, 18.12.2007)
Господин Зеехофер предостерёг о возможном провале переговоров. По его словам, коллегии министров в крайнем случае придётся заседать всю ночь до утра среды, чтобы *окончательно согласовать* реформу. Если же отложить этот проблемный вопрос на следующий год, решить его потом будет ещё сложнее.
🕮 Ср. также идиомы *in trockenen Tüchern sein, in trockene Tücher kommen* и *in trockenen Tüchern haben (etw. A).*

in trockenen Tüchern haben (*etw. A*)
довести до конца *(что-л.), (окончательно)* уладить *(что-л.)*

🗐 Wir *haben* jetzt schon alles *in trockenen Tüchern* und wollten den neuen Trainer auch so schnell wie möglich der Mannschaft vorstellen. (Braunschweiger Zeitung, 26.01.2012)
Мы уже всё *уладили* и хотим как можно скорее представить нового тренера команде.

🗐 Vorrang vor allem anderen hat für Obama nach wie vor die Gesundheitsreform, die er möglichst noch in diesem Jahr *in trockenen Tüchern haben* will. (Mannheimer Morgen, 10.12.2009)
Предпочтение Обама по-прежнему отдаёт реформе здравоохранения, которую он, по возможности, планирует *довести до конца* в этом году.

🕮 Ср. также идиомы *in trockenen Tüchern sein, in trockene Tücher bringen (etw. A)* и *in trockene Tücher kommen.*

in trockene Tücher kommen
реализоваться, быть принятым

🗐 Während die Finanzmärkte in Asien und Europa zunächst positiv auf die Einigung in Washington reagierten, fiel der Enthusiasmus an der Wall Street gedämpft aus. Zudem warten die Märkte ab, ob die Vereinbarung zwischen Obama und den Kongressführern tatsächlich *in trockene Tücher kommt*. (Nach: St. Galler Tagblatt, 02.08.2011)
В то время как финансовые рынки Азии и Европы положительно отреагировали на достигнутую в Вашингтоне договорённость, на Уолл Стрит она была встречена с вялым энтузиазмом. Кроме того, рынки выжидают, будет ли *принято* соглашение между Обамой и конгрессменами.

🗐 Die Nervosität im Regierungslager steigt. Schließlich sollen innerhalb der nächsten fünf Wochen die ehrgeizigsten Reformen *in trockene Tücher kommen,* die sich die Koalition vorgenommen hat. (Nach: dpa, 02.06.2006)
В правительственном лагере растёт нервное напряжение. Ведь в течение следующих пяти недель предстоит *реализовать* самые смелые реформы, о которых заявила правящая коалиция.

🕮 Ср. также идиомы *in trockenen Tüchern sein, in trockene Tücher bringen (etw. A)* и *in trockenen Tüchern haben (etw. A).*

in trockenen Tüchern sein: (*etw.*) ist in trockenen Tüchern
(что-л.) (окончательно) решено, *(что-л.)* доведено до конца

🗐 Neben der Arbeit an Großveranstaltungen trifft der Coach auch Personalentscheidungen für seine Mannschaft. „Aber zu meiner Philosophie gehört es, erst über Namen zu sprechen, wenn alles *in trockenen Tüchern ist*", sagt er. (Nach: Braunschweiger Zeitung, 11.01.2012)
Тренер не только организует крупные мероприятия, но и принимает решения, касающиеся состава команды. «Таков мой принцип – называть имена, только когда всё *решено окончательно»,* – замечает он.

🗐 Allerdings *ist* die Finanzierung des Großprojekts noch nicht *in trockenen Tüchern.* Es steht noch nicht ausreichend Geld zur Verfügung. (Nach: Mannheimer Morgen, 13.01.2012)
Однако вопрос о финансировании этого крупномасштабного проекта ещё не *решён окончательно.* Пока не было выделено достаточно средств.

🕮 Ср. также идиомы *in trockene Tücher bringen (etw. A), in trockene Tücher kommen* и *in trockenen Tüchern haben (etw. A).*

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

Laura Giacomini
University of Heidelberg
laura.giacomini@iued.uni-heidelberg.de

Abstract

Today there is still a significant need for specific lexicographic resources in digital form, as they can remarkably improve access to data and actively assist the process of text production. This paper describes the way in which corpus data can be explored to retrieve suitable material for the representation of a particular class of culture-specific phraseologisms and similes in a bilingual dictionary for translators.

Keywords: phraseologisms; corpora; translation

1 Introduction

The characteristic formal stability and contentual figurativeness of phraseological expressions as a result of cultural encoding best reflect a society's deeply rooted patterns of world interpretation. Given the strong presence of phraseologisms in the lexicon of a language and the translatability issues they inevitably raise, it is necessary to support the practice of translation by means of specific and up-to-date lexicographic resources. This paper describes the way in which corpus data can be explored to retrieve suitable material for the representation of a particular class of culture-specific phraseologisms, similes, in a bilingual lexicographic resource for translators. It is based on a larger study carried out in 2013 at the Australian National University (Canberra) on the language pair Australian English (AuE) – Italian and recently published (Giacomini 2014). Translatability issues arise from semantic obscurity linked to the presence of culture-specific words and concepts.

General language dictionaries usually define similes and other multiword expressions through a paraphrase, without providing synonymic idioms even if these are available. This prevents translators from finding functionally equivalent phraseological data, which would be particularly useful for reproducing semantic and pragmatic features as well as the overall familiarity of the multiword expression in the target language. In the case of language pairs with phraseology reflecting distinctive cultural marks (Wierzbicka 2010), the exploration of corpus data can efficiently support the selection of adequate phraseological equivalents through reliable quantitative measures, thus forming a useful dictionary basis.

2 Object and sources

Similes are based on an explicit comparison between entities and are semantically related to metaphors, in which resemblance becomes implicit and one thing is understood and experienced in terms of another (Lakoff/Johnson 2003: 21 ff., Wikberg 2008: 128). In the usually regular syntactic structure of similes, the resemblance relation between the two compared entities is expressed by a connective, mostly *like/as* in English and *(così) come* in Italian. In addition, similes can reveal a different phraseological status: they can be defined either as collocations or as semi-idioms, according to the transparency of the comparison.

Two comparison patterns in AuE have been considered, the first being similes containing the phrase *full as* (e.g. *full as a tick*) meaning a) "having no empty space", b) "having eaten to one's limits or satisfaction", or c) "drunk", and the second involving a single culture-specific lexeme as the second compared entity, mostly a native animal. The semantic pivot is the contextual reading of the word that designates the shared property (*tertium comparationis*) and that determines the referential object of the multiword expression as a whole, both on the denotative and the connotative level. However, whereas similes belonging to the first pattern are made up of elements that are semantically transparent in their literal and figurative meanings, both for the English and the Italian native speaker, the others confront the translator with the presence of *realia* (cultural keywords) involving a culture- or environment-specific referent (Peters 2007: 249-251).

AuE similes were chosen on the basis of their relevance in a large-scale digital corpus of full-text Australian general news sources[1], major general language dictionaries of AuE, and selected dictionaries of idioms or colloquialisms. Italian monolingual general language dictionaries and a comparable newspaper corpus (articles published between 2000 and 2013 in major Italian newspapers, totalling around 980 million words) were also employed for this purpose. Due to their stable structure, similes can be split into bigrams. The closest Italian equivalents of the semantic bases (e.g. *full ≈ pieno, sazio, ubriaco*) can be used to query the corpus for their collocators. The absolute frequency of the extracted bigrams can be compared with their log likelihood value, which provides reliable information on the association strength of a certain bigram and thus on its suitability as a phraseological equivalent (cf. Dobrovol'skij 2009[2]). The results of data analysis are displayed in Table 1 according to an onomasiological procedure, which assigns Italian phraseological units to the concepts expressed by the AuE similes. Up to five equivalents for each concept are shown and arranged according to their absolute frequency F in the corpus and the log likelihood ratio LL.

[1] National and regional newspaper texts covering the period 1985 to 2012, the Australian Corpus of English (ACE), and the Trove database (National Library of Australia).

[2] Log likelihood has been chosen because of its reliability with sparse data, which is the case of the chosen words in the AuE corpus (for the topic of LL and normal distribution cf. McEnery et al. 2006, 53).

to be full as a goog/ state school (hat rack)/ catholic school/ fat lady's sock	
(a) "full/ overcrowded"	essere pieno zeppo (1799/>100), essere pieno come un uovo (194/>100), essere pieno da scoppiare (10/>100), essere pieno come una botte (1/9)
to be full as a goog/ tick/ boot/ fairy's phone book/ fat lady's sock	
(b) "full up/ satiated"	essere pieno come un uovo (194/>100), essere pieno da scoppiare (10/>100), essere pieno come un otre (1/17)
(c) "full/ drunk/ intoxicated"	essere ubriaco fradicio (567/>100); avere bevuto come una spugna (5/21); essere pieno come un otre (1/17), essere pieno come una botte (1/9)
to be mad as a cut snake, to be pissed as a parrot	
(d) "angry/ nervous"	essere (incavolato/…) nero (245/>100); essere arrabbiato/ incavolato/… come una bestia (27/>100)/ una belva (7/>100)/ una iena (6/>100)/ una biscia (5/>100)
to be pissed as a parrot: cf. meaning (c)	
to be mad as a cut snake, to be mad as a gum tree full of galahs	
(e) "crazy/ eccentric"	essere fuori di testa (1548/>100), essere tutto matto (343/>100), essere pazzo/ matto da legare (96/>100), essere pazzo/ matto come un cavallo (27/>100), essere fuori come un balcone (27/>100)
to be (as) game as Ned Kelly	
(h) "game/ brave/ bold"	avere coraggio da vendere (154/>100), avere un coraggio da leone/leoni (130/>100); avere il coraggio di un leone (7/59); essere coraggioso come un leone (5/83)
to be flat out like a lizard drinking	
(j) "fully extended"	essere/ stare lungo disteso (60/39)
(k) "with the utmost effort"	col massimo impegno (251/>100); impegnarsi al massimo (139/>100); col massimo sforzo (9/42); sforzarsi al massimo (2/7)
(l) "very busy"	essere pieno/ oberato di lavoro (404/>100), essere pieno di impegni (376/>100)
(m) "at full speed"	cf. (i)
to be miserable as a bandicoot	
(n) "wretchedly unhappy"	essere un povero diavolo (243/>100), essere un povero Cristo/cristo (238/>100)
(o) "contemptible"	-
(p) "needy"	essere povero in canna (139/>100), essere povero come Giobbe (2/31)

Table 1: Corpus data in the target language with F and LL values.

3 Translatability

AuE similes turn out to have close counterparts among phraseological Europeanisms, even though this may happen to varying degrees of equivalence. *Full* usually retains in the simile both its literal and figurative meanings, thus determining the total or partial compositionality of the multiword expression and contributing to its transparency.

Compositionality can be stated on the denotative (either literal or figurative) and connotative semantic level, but not always on the pragmatic level. The observations concerning compositionality of the multiword expression are also true of the second comparison pattern, in which a variable adjective (e.g. *mad, full, miserable…*) or a verb (e.g. *to shoot through*) designating shared property is combined with a culture-specific entity, used with a predicative or an adverbial function. In the case of similes belonging to this pattern, *realia* inevitably produce a lexical gap. Compatible data in the target language and culture cannot be sought for in terms of denotatively equivalent phraseological expressions, which is possible in the case of the *full*-pattern, but, at the most, in similes sharing the semantic pivot (*matto, veloce, coraggioso,* etc.) and with an equal degree of compositionality.

4 Lexicographic representation

The extracted clusters of equivalence candidates disclose the presence of alternative comparative patterns in Italian. For instance, we have prepositional phrases headed by *da* (*avere un coraggio da leone*) or *di* (*avere il coraggio di un leone*). In the Italian language, a significant part of the extracted lexical components in similes and other idioms stereotypically refer to animal behaviour and belong to a common European cultural heritage.

Dictionary data have been a useful resource to identify initial information on some widespread phraseologisms, but they fail to cover context-dependent phraseological variation and variability. The comparable newspaper corpus, instead, has revealed that, in concrete language usage, non-lexicalized and not conventionalised collocative or semi-idiomatic variants performing more specific message functions are constantly created on the basis of already existing patterns. The evaluation of corpus data can also disclose differences in phraseological distribution among languages (e.g. a strong tendency of the Italian languages towards metaphorical comparisons for purely descriptive purposes) and point out recent lexical formations which have not yet been recorded in dictionaries (cf. *essere fuori come un balcone*) but are already perceived by native speakers as familiar.

Corpus analysis in the target language supports the creation of a lexicographic basis for a bidirectional dictionary that is suitable for both translation directions. On the one hand, it activates passive knowledge in the native speaker of Italian by providing him/her with pragmatic tags in the source language and a wide choice of equivalents in the target language. On the other hand, it supports the AuE native speaker who is performing an active translation task by 1) allowing for a statistic evaluation of the word combinations, 2) tagging equivalents with pragmatic marks and, most of all, 3) categorizing phraseologisms with varying idiomatic range (*pieno zeppo*, *pieno da scoppiare*) and distinguishing them from non-phraseological material (cf. Wiegand 2002: 52-53). Among non-phraseological equivalents are often single lexical items, usually an emphasised adjective (*strapieno*, *affollatissimo*; *sbronzo*) that can be specifically sought for in syntagmatic or paradigmatic dictionaries and further tested for phraseological strength. Every time a semantically and pragmatically equivalent phraseologism is missing, dictionary users are provided with an open set of non-phraseologisms, which function as reproducible syntactic models (e.g. superlative adjectives) and are particularly helpful for non-native speakers of the target language.

In order to take full advantage of the rich corpus materials and its bifunctionality, the dictionary should be designed as a digital resource, which should allow the translator to access lexicographic data along different combinable criteria, grasp the semantic connections existing between phraseological expressions, and retrieve unabridged corpus examples for each of them, in both the source and the target language.

The first two goals, *data accessibility* and *the disclosure of semantic connections*, can be primarily achieved through a coherent onomasiological macrostructure, which should group phraseologisms together along a common denotative/connotative meaning, and a systematic mediostructure, the aim of which should be to link each meaning to the correspondent phraseologisms and vice versa. The entry examples below show how lexicographic data in the section AuE-Italian can be displayed in a functional microstructural frame, and arranged based on a specific kind of search input (Table 2 according to a specific concept, Table 3 according to a specific phraseologism)[3].

PHRASEOLOGISMS MATCHING THE CONCEPT IN THE SOURCE LANGUAGE	**EQUIVALENTS IN THE TARGET LANGUAGE**
to be full as a goog *coll.* ≈ **state school (hat rack)** *coll.* ≈ **catholic school** *coll.* ≈ **fat lady's sock** *coll.*	PHRAS: essere (pieno) zeppo, pieno come un uovo *coll.*, pieno da scoppiare *coll.*, pieno come una botte *coll.* ◆ essere pienissimo, affollatissimo, strapieno *coll.*, stracolmo

Table 2: Search input: the concept FULL/OVERCROWDED.

[3] ◆ marks the division between phraseological and non-phraseological equivalents, ≈ indicates similes referring to the same concepts.

CONCEPTS MATCHING THE PHRASEOLOGISM IN THE SOURCE LANGUAGE	**EQUIVALENT PHRASEOLOGISMS IN THE SOURCE LANGUAGE**	**EQUIVALENTS IN THE TARGET LANGUAGE**
FULL/OVERCROWDED	≈ **state school (hat rack)** *coll.* ≈ **catholic school** *coll.* ≈ **fat lady's sock** *coll.*	PHRAS: essere (pieno) zeppo, pieno come un uovo *coll.*, pieno da scoppiare *coll.*, pieno come una botte *coll.* ◆ essere pienissimo, affollatissimo, strapieno *coll.*, stracolmo
FULL UP/SATIATED	≈ **tick coll.** ≈ **boot coll.** ≈ **fairy's phone book coll.** ≈ **fat lady's sock coll.**	PHRAS: essere pieno come un uovo *coll.*, pieno da scoppiare *coll.*, pieno come un otre *coll.* ◆ essere pienissimo *coll.*, strapieno *coll.*, stracolmo *coll.*
FULL/DRUNK	≈ **tick coll.** ≈ **boot coll.** ≈ **fairy's phone book coll.** ≈ **fat lady's sock coll.**	PHRAS: essere ubriaco fradicio; avere bevuto come una spugna *coll.*; essere pieno come un otre *coll.*, una botte *coll.*

Table 3: Search input: the phraseologism to be full as a goog.

The modular microstructure includes the following items: concept, phraseologism in the source language, phraseologisms in the source language referred to the same concept, and equivalents in the target language (subdivided into phraseological and non-phraseological equivalents). Pragmatic tags are added to a phraseologism or equivalent whenever required.

According to the lexicographic corpus, these Italian similes do not have a marked level of usage. However, a glance at their concordances in the newspaper corpus reveals a frequent tendency towards a colloquial register. In comparison with the source text similes, a generally more neutral level of usage has to be clearly stressed. The equivalents are selected on the grounds of their statistical relevance in the corpus and are not meant to cover the whole spectrum of equivalence in the target language. For the professional translator, they constitute the starting point from which further translation proposals can be generated.

CONCEPT	PHRASEOLOGISM	CONCORDANCES
FULL/ OVERCROWDED	**PHRAS: to be full as a goog *coll.***	The carpark in a certain flat pack emporium, starting with I and ending with A, middle letters K and E, was **as full as a goog**. We drove about 4000km with five adults and a lot of luggage and although the car was **as full as a goog** it was a good performer. She took a chance and opened up Swansea cafe. **Full as a goog**. And she must be doing something right because her business is a finalist in the Cafe category. How weird, though, that Old Trafford can hold only - even when it's **as full as a goog** - 23,000? They reckon they could have sold 70,000 tickets for the last day.
	PHRAS: essere pieno come un uovo *coll.*	Non possiamo farvi entrare – gridano gli organizzatori ai tornelli - dentro è **pieno come un uovo** e non si respire. La platea è quella di lavoratori provenienti da tutta la regione, ieri mattina al PalaDozza (**pieno come un uovo**) E ci piacerebbe che il palasport fosse **pieno come un uovo** (i biglietti numerati sono già stati esauriti ieri in prevendita acclamato come una star nella sua tappa aretina del tour in camper. **Pieno come un uovo** l'auditorium del palaffari

Table 4: Links to corpus concordances.

In order to account for context-dependent variation in meaning and register, each equivalent needs to be hyperlinked to the correspondent corpus concordances both in the source and in the target language, which provide the translator with a large-scale database of real language examples (cf. Table 4 for corpus concordances of phraseologisms related to the concept FULL/OVERCROWDED).
This concept-based macrostructure could also constitute the architecture of a multilingual resource aimed at the representation of a core of cultural scripts shared by different languages (for recent research on Europeanisms cf. Piirainen 2012, Reichmann 2001).

5 Conclusions

Today there is still a significant need for specific lexicographic resources in digital form for translators, as they can remarkably improve access to data and actively assist the process of text production. In the best-case scenario, such resources could be integrated, together with other dictionaries and language tools, in multi-layer databases, allowing for advanced and customised search options.
This study shows that syntactic and semantic patterns can be effectively extracted from corpora and serve as lexicographic data in a digital resource which is specifically designed for supporting translation of culture-specific word combinations thanks to an onomasiological/conceptual macrostructure and a systematic mediostructure. Moreover, the study demonstrates that a corpus-based procedure is able to adequately account for phraseological variation and variability.

6 References

Dobrovol'skij, D. (2009), Zur lexikografischen Repräsentation der Phraseme (mit Schwerpunkt auf zweisprachigen Wörterbüchern). In Mellado Blanco, C. (ed.), *Theorie und Praxis der idiomatischen Wörterbücher*, Lexicographica Series Maior, pp. 149-168

Giacomini, L. (2014), Languages in Comparison(s): Using Corpora to Translate Culture-Specific Similes. In: SILTA Studi Italiani di Linguistica teorica e Applcata, Pacini Editore 3/2013.

Lakoff, G./Johnson, M. (2003), *Metaphors We Live By*, Chicago/London, University of Chicago Press.

McEnery, T. et al. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Milton Park/Abingdon/Oxon, Routledge.

Peters, P. (2007), Similes and other evaluative idioms in Australian English". In Skandera P. (ed.), *Phraseology and Culture in English*, Berlin, Mouton de Gruyter, pp. 235-256.

Piirainen, E. (2012), Widespread Idioms in Europe and Beyond: Towards a Lexicon of Common Figurative Units, Frankfurt am Main, Peter Lang.

Reichmann, O. (2001), Das nationale und das europäische Modell in der Sprachgeschichtsschreibung des Deutschen, Freiburg (Schweiz), Universitätsverlag.

San Vicente, F. (ed.), *Lessicografia bilingue e traduzione*, Milano, Polimetrica

Wiegand, H.E. (2002), Äquivalenz, Äquivalentdifferenzierung und Äquivalentpräsentation in zweisprachigen Wörterbüchern: Eine neue einheitliche Konzeption. In *Symposium on Lexicography XI: Proceedings oft he Eleventh International Symposium on Lexicography*, Copenhagen, pp. 17-57.

Wierzbicka, A. (2010), Experience, Evidence, and Sense: The Hidden Cultural Legacy of English, OUP.

Wikberg, K. (2008), Phrasal similes in the BNC. In Granger S., Meunier F. (eds.), *Phraseology: An Interdisciplinary Perspective*, Amsterdam/Philadelphia, John Benjamins, pp. 127-142.

Lexical Variation within Phraseological Units

Tarja Riitta Heinonen
Institute for the Languages of Finland
tarja.heinonen@kotus.fi

Abstract

This paper discusses lexical variation in phraseological units from theoretical and lexicographical perspectives. The starting point is the observation that the existence of lexical variation is sometimes disputed in principle. It has been argued that a change in a single word is sufficient to change the meaning of the whole, thus creating a new expression. Another argument is that after allowing variation in one word one cannot but allow it in multiple words, which quickly turns the original expression unrecognizable. In contrast to these views, this paper argues that lexical variation is not arbitrary but follows certain principles. When all contributing factors are taken into account, the variation in phraseological units is often item-specific, and yet it conforms to general patterns. Towards the end of the paper, Petrova's (2011) multi-level model will be introduced, offering a promising view for theoretical analysis. However, in dictionary work it is reasonable to adhere to generally accepted conventions and not to complicate the structure of the entries too much. An ideal entry gives a sound corpus-based description with representative examples of usage.

Keywords: phraseology; lexical variation; usage-based; idiomatic meaning; lexicographical practices

1 Introduction

Lexical variation within phraseological units raises theoretical and practical problems. One of the major questions is how phraseological units are learned and recognized if not with the help of the lexical items that constitute them. This paper starts with the question of whether lexical variation is acceptable in the first place (either in lexicography or in theory). Most dictionaries I have consulted allow variation but there are differences in "how much" and "what kind".

I will show that lexical variation is not radically different from other types of variation, grammatical and structural variation, and that their workings can be described in a common model.

2 Two Opposing Views

It is common knowledge today (Moon 1998a: 92; Atkins & Rundell 2008: 168 among others) that there is a considerable amount of variation within multiword expressions, or phraseological units as I call them in this paper. However, there are two opposing views on how to deal with expressions that are very similar to each other, except for single lexical choices as in, for instance, (1):

(1) rats desert/leave/quit/forsake a (sinking) ship (ODEI)

Some researchers exclude lexical variation from phraseological units in principle: substituting a word for another would automatically mean that the result no longer represents the same item as the original one (Wulff 2008: 76). In contrast, other researchers consider lexical substitutability as one type of variation alongside morphosyntactic variation (e.g. Sköldberg 2004).

The opposite views may be due to differences in theoretical frameworks, but they may also be related to the amount of real life data researchers are familiar with. In my data set, collected from newspapers and the Internet, there is a considerable amount of variation in which two or more (near)synonyms occur in one and the same context without difference in meaning, as in the case *rats desert/leave a sinking ship*. The sheer number of such cases requires attention.

If lexical substitutability is allowed, it leads to a question of how phraseological units are defined – and originally recognized – if not with the help of the words that constitute them. On the other hand, if all expressions that are not lexically identical represent different phraseological units, we are left with plenty of units that highly resemble each other and without any formal means to record this in the lexicon.

3 Variation Patterns

In order to get a more detailed idea of what kind of alternation patterns there are in phraseological units, I will briefly consider a few examples from earlier studies on variation. Here I will mostly rely on Moon's systematic work on fixed expressions in English (1998b) and my own studies on Finnish verb phrase idioms (Heinonen 2013, 2007), but I will cite examples from various sources.

3.1 Grammatical, Constructional and Lexical Variation

I will start by dividing the area of variation into three subfields: grammatical, constructional and lexical. These three phenomena are conceptually separable, but they co-occur in actual utterances. For instance, a constructional variant often occurs with specific lexical choices. By grammatical variation I mean variation in morphosyntactic features such as number, definiteness, voice and tense. Typically, the predicate verb inflects, albeit not quite freely (Moon 1998b: 94), but its nominal complements tend to be fixed in specific forms (2). However, in some cases also features in noun phrases may vary, especially if the idiomatic expression is metaphorically analyzable (3).

(2) she *gets* ~ *got* cold feet (! a cold foot)[1]

(3) (Swedish:) dra *sitt strå* (~ *sina strån*) till stacken
literally: drag one's straw (~ straws) to the stack
'contribute one's share to a common purpose' (Sköldberg 2004: 203, 311-312)

Inflectional restrictions are idiom-specific: for instance, some phraseological units passivize, some do not. Uttermost, the list of restrictions covers almost all the features: Čermák (2001: 15) cites a dictionary entry for the Czech idiom *tahat někoho za nohu* 'to pull someone's leg' which states that the predicate verb does not normally occur in the interrogative, negative, passive, conditional, imperative, future, or in the 1st person. It looks like these restrictions cannot be purely grammatical, but they describe the way the idiom is normally used.
For (2), the idiom dictionaries CCID and ODEI give also constructional variants with different predicate verbs (4a-b). This alternation pattern with verbs 'get', 'have' and 'give' is typical of possessive idioms.

(4a) *get* cold feet or *have* cold feet about something (CCID)
(4b) *get*, (begin to) *have*, or *give* somebody cold feet (ODEI)

The constructional variants also include the alternative expressions of causation, states and processes. An example of causative variant is given in (5). – Moon (1998b: 139ff) lists these and some more patterns under the label "systematic variation".

(5) *go* through the wringer – *put* someone through the wringer (Moon 1998b: 141)

Generally speaking, the variation in the above patterns forms a simple grid with three points: causation, change and state, see figure 1. All these points can have a possessive interpretation as well: one could represent that as parallel to the basic grid. Verbs like *give*, *put* and *throw* are among the verbs that occur in the causative pattern, verbs like *get* and *go* express change or movement, and the verbs *be* and *have* are the primary state verbs.

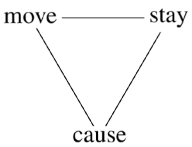


Figure 1: A basic constructional grid for causation, change or movement, and state.

[1] ! before the variant means that the idiomatic meaning gets lost.

Constructional and lexical variations also meet in extra modification (6).

(6) a *recent* tempest in a *publishing* teapot (Ernst 1981: 54)

Finally, the main focus of this paper, "pure" lexical variation occurs within a single construction. In examples (7) to (9) two or more near-synonymous or otherwise conceptually related words are in paradigmatic alternation. In (7) and (8), both choices are conventionalized, in (9), the first two.

(7) a *chink* ~ *crack* in one's armour (ODEI)
(8) *flog* ~ *beat* a dead horse (Moon 1998b: 133)
(9) (Finnish:) kahvihammasta *pakottaa* ~ *kolottaa* ~ *särkee* (and several other verbs meaning 'ache')
literally: a coffee tooth is aching
'someone has a craving for coffee' (Heinonen 2013: 123-124)

It is important to notice that *not* all lexical variation is like this. In example (10b), the substitute for *family*, "the EU", is not a semantically related word to *family*, but an expression whose referent can be seen as a kind of family. Here we are not operating inside the lexicon, but with categories or sets defined by common attributes (Philip 2008: 105-106).

(10a) black sheep of the *family*
(10b) black sheep of the *EU* (Philip 2008: 106)

Two or more lexical elements of one phraseological unit may also co-vary. Often they vary basically independently of each other (11), even though some combinations are more typical than others.

(11) shake ~ quake ~ quiver in one's boots ~ shoes (Moon 1998a: 95, Moon 1998b: 161)
As far as I know, covariance is not widely studied in phraseology. Moon (1998b: 161ff) refers to cases like (11) as "idiom schemas". The variants are listed and given some cover terms like 'oscillate in one's footwear'. In Moon (2008), she studies a similar phenomenon in similes of the structure 'thin' + *as* + 'something which is very narrow in comparison to its length'. Here, it is easy to see that the concrete realizations differ from each other in connotations even though the schema represents the variants in very much the same way as in (11). *Skinny as a rat* sounds less attractive than *thin as a whippet* (id. 9-11). There are also collocational preferences.
In some phraseological units, however, one variant is dependent on the other (12):

(12) (Finnish:) lähtee kuin *hauki rannasta* ~ *talonmies peltikatolta* ~ *faarao sarkofagista* (and many others)
literally: leaves like a pike from a shore ~ a janitor from a tin roof ~ a pharaoh from a sarcophagus
'leaves quickly'

In this simile schema, someone or something leaves a place where they either belong or that they are at least strongly associated with. An attempt at explaining the mechanisms behind covariance is made by Stefanowitsch and Gries (2005). Their answer is, unsurprisingly, that the varying elements cohere semantically. More exactly, they divide semantic coherence into (at least) three different kinds: coherence based on frame-semantic knowledge, coherence based on prototypes, and image-schematic coherence. To my understanding, the coherence in similes in (12) would represent "frame-semantic" coherence.

3.2 Lexical Variation in Dictionary Entries

The presentation of dictionary entries has traditionally been dense. This also shows in how phraseological variants may be placed side by side, for example, in ODEI and in Duden (13).

(13) einen klaren/kühlen Kopf bewahren/behalten (Duden s.v. *Kopf*).

The practices vary, but generally speaking dictionaries have specific means to indicate if a phraseological unit has more than one lexical choice (a slash in 13, the word *or* and parentheses in 14, a comma in 15). It follows that lexical variation in phrases is usually *de facto* accepted in lexicography. How systematically lexical variation is taken into account depends mostly on the type of the dictionary: phraseological dictionaries such as ODEI and CCID are systematic, general dictionaries, such as KS, are less so.

(14) throw (or pour) cold water on (NODE s.v. *cold*)
(15) (Finnish:) erottaa, seuloa jyvät akanoista (KS s.v. *jyvä*)
'separate, sift the wheat from the chaff'

Dictionary conventions actually work equally well for bundles of idiomatic expressions that have different, perhaps even opposite meanings, as in (16):

(16) say the right/wrong thing (ODEI)

Seen this way, a dictionary entry could also stand for partly formal idioms, following the terminology by Fillmore, Kay & O'Connor (1988). In their parlance, formal idioms are lexically open syntactic patterns or constructions, while what we have traditionally called idioms are substantive or lexically specified idioms. What I find interesting here is the area between these, ie. partially open idioms. An actual example of a partially open, partially specified (sub)entry is given in (17):

(17) play the –––– card e.g., *he saw an opportunity to play the peace card* (NODE s.v. *card*)

Two or more alternating slots predict multiplied combinations. In the case of (13), two adjectives (*klar, kühl*) and two verbs (*bewahren, behalten*) combine in four different ways. In this case, the generalization is valid: all predicted forms actually occur in texts. However, all combinations are not as common: corpus studies via DWDS reveal differences in frequencies. As far as I know, there are no lexicographic conventions that help the user pick the most idiomatic combination(s) in such situations.

Sometimes the variation is not limited to specific words but the alternation set is lexically open. Many Finnish verb phrase idioms allow plenty of variation in the predicate verb. I searched for verb variants in the idiom *heittää kapuloita rattaisiin* ('throw batons to (the wheels of) a carriage [in order to prevent something from succeeding]') in a newspaper corpus FTC and found about 20 different predicate verbs (some of them listed in 18a-e). Of these, six verbs are rather common (16–30 hits), three occur about five times and the rest are mostly hapaxes. In (18a-e), the attested verbs are divided into meaning groups (based on Heinonen 2007: 155 and Heinonen 2013:156-157), and the main six variants are in boldface.

(18a) 'throw': **heittää**, **heitellä**, viskoa, viskellä
(18b) 'put': **panna**, **laittaa**, asettaa, asetella
(18c) 'stick': **pistää**, pistellä, tuikkia
(18d) 'push': työntää, pökkiä
(18e) 'hit': **lyödä**, iskeä

The general dictionary KS mentions the verbs *heitellä* and *panna* in this idiom, and the phraseological dictionary SSIS lists all the six common ones. (SSIS is based on the same corpus as my search.)

One solution to the problem of long lists is to generalize over the choices as in (18a-e). However, it is not always clear how these sets should be labeled and interpreted. Also, some lexical items are preferred over others with similar meanings. Notice that in (17), the idiom contains an open, unspecified slot, but the explanation given – that it should refer to an "issue or idea" that can be exploited especially for "political advantage" – does not really limit the choice of appropriate fillers, and this is possibly the most we can say, besides giving attested examples. Another problem is that the sets (in whatever way they are defined) tend to leak. There are a few Finnish idioms with a lexical item that refers to a human head. Still, the appropriate sets for a 'head' are partly lexically specified: in one idiom you can refer to a head metaphorically as a 'cabbage', in another as a 'pin', but not the other way round (Heinonen 2013: 197-198). Jezek and Hanks (2010) make the same observation, saying that paradigmatic sets of words do not map neatly onto conceptual categories, and neither are there stable generalizations over different contexts.

4 One or More Units

There are two further points that speak for lexical variation in phraseological units: interpreation of regular derivational variants (4.1) and language learning as usage-based process (4.2). Issues of variation vs. modification and canonical forms are dealt with briefly in (4.3).

4.1 Difference in Meaning as a Criterion

A difference in meaning has been a crucial factor in separating phraseological units from each other. However, it should be kept in mind that a change in one element usually contributes to the full meaning quite predictably. In Finnish, certain derivative affixes can change the aspect of the verb. Substituting a verb with its derived counterpart expressing frequentative aspect counts as lexical substitution (cf. 19a and 19b); still, the resulting difference in meaning is quite straightforward: it is actually comparable to how inflectional affixes are interpreted (cf. 19b and 19c):

(19a) (Finnish:) *heittää* kapuloita rattaisiin
(literally:) throws batons to (the wheels of) a carriage
'places obstacles in order to prevent something from succeeding'
(19b) *heittelee* kapuloita rattaisiin
keeps throwing batons [...]
(19c) *heitti* kapuloita rattaisiin.
threw batons [...]

I would suggest that the predictability of a change in meaning also applies when a word is substituted with an unrelated word. As long as the phraseological unit is recognized, its meaning can be modified according to the contribution the substitute part carries along.

4.2 Where are the Limits of One Unit?

At which point does a phraseological unit change into another one? Čermák (2001: 7) raises this question of an idiom's identity if we allow lexical variation within them. I believe that it is not possible to define from outside how the units are organized in the mental lexicons of speakers. It is likely there is not just one way in which phraseological units and lexical items are connected to each other. The individual lexicons develop hand in hand with language use. When we learn an expression, we also learn, little by little, how it is used: inflection, meaning(s), suitable contexts.

Čermák's example idiom 'to pull someone's leg' referred to earlier is given a voluminous description in his and his colleagues dictionary (Čermák, Hronek & Machač 1994): not only does it cover the idiom's valency, inflectional restrictions, meaning, style and appropriate context of use, but also the related expressions and even equivalents in various other languages. Lacking knowledge of Czech, I cannot comment on this particular case but, in general, this sounds like a database and network of idioms as independent, lexically static units. Observations on language use may therefore lead to opposite conceptions on what phraseological units are like.

4.3 Canonical Forms and Variation

Talk about variation raises the question what the constants are. A canonical form of a phraseological unit appears to be a paradox. Philip (2008: 95, 103) refers to canonical forms as the most typical forms and, at the same time, as something that are generally outnumbered by corresponding non-canonical forms in language corpora. Canonical forms are often identified with dictionary citation forms, as if lexicographers would receive them by some divine announcement.

There is also a lexicographic tradition to keep established variation and temporary modification as separate phenomena (e.g. Burger, Buhofer & Sialm 1982). However, seen through corpora, variation and modification blend. They also derive from the same sources.

4.4 A Multi-Level Approach

If we accept lexical substitutability within one phraseological unit, we need to demonstrate how such a phraseological unit can be defined and identified. One applicable idea is the multi-level model by Petrova (2011), which is derived from Jackendoff's Conceptual Semantics and more directly from Nikanne's Tier-net model (Nikanne 2005: 191-210; Petrova 2011: 110ff). In Petrova's model, one unit can vary at several tiers (phonological, morphological, syntactic, conceptual) at the same time (see figure 2). The idea is that the lexical items, as well as some of the inflectional affixes, are chosen by default (this is represented by df-links in the figure), but it is possible to substitute or even leave out one or more of the default items. The predicate verb inflects freely in the example case. The conceptual structure is relevant when substituting lexical items, since any other words following the same syntactic pattern would not do. For instance, *to throw balls to boys* does not represent the same phraseological unit as *to throw pearls to pigs*.

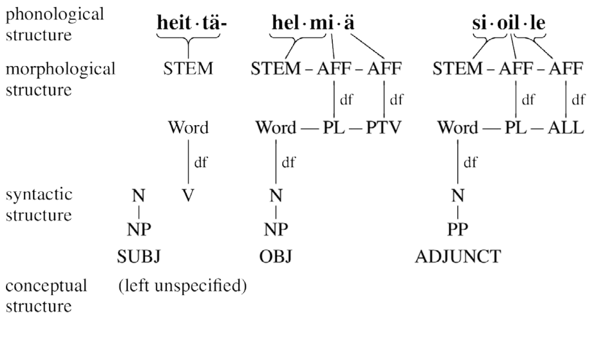


Figure 2: The structure of the idiom *X heittää helmiä sioille* 'X casts pearls before swine' (literally 'X throws pearls to pigs') according to Petrova (2011: 151). The figure is heavily simplified from the original by the present author. PTV (partitive) and ALL (allative) are case markers.

The model supports the view that lexical variation can be dealt with together with other varying elements and not as a separate issue. Petrova cites a good number of actual uses of the idiom *heittää helmiä sioille*, 'cast pearls before swine', which illustrate how the different sources of variation function together (20a-d). One typical pattern is a verbless construction (20a); in fact, it is the most common variant when all default features are taken into account simultaneously (ca. one quarter of all 480 tokens in Petrova's study) and far more common than a "canonical" citation form (Petrova 2011: 219-220). In quite a few examples the syntactic pattern, inflectional forms or the words are not the default ones (20b-d).

(20a) Ei helmiä sioille, kuten sanonta kuuluu.
'No pearls to pigs, as the saying goes.' (Petrova 2011: 278)
(20b) *possulle* heitetty helmi
'a pearl thrown to a piggy' (id. 284)
(20c) viisauden helmien *jakamiseen* myös *meille typerämmille yksilöille*
'distributing pearls of wisdom to us more stupid individuals' (id. 281)
(20d) *Menikö* taas helmet *sinne kaukaloon* [?]
'Did pearls go to that trough again [?]' (id. 279)

The conceptual structure (left unspecified in figure 2 for ease of representation) contains detailed information on verb semantics, argument structure, what sorts of objects nominal complements refer to etc. (id. 137-138). It is straightforward to substitute *pig* for its near-synonym *piggy* (20b), but to connect *a trough* with *pigs* as in (20d) requires a larger situational framework. Since the predicate verb 'throw' involves causation, it is predictable that the idiom participates in the causative alternation pattern (20d, cf. figure 1). Petrova points out that the thrown objects, *pearls*, are evaluated by the speaker as good, and the recipients, *pigs*, as inadequate in some way (id. 24, 138). The variant to 'pigs' in (20c), 'more stupid individuals' reflects this.

It could be argued that this particular idiom is not representative in allowing much more variation than is normally the case. As can be seen, the idiom functions even with just one default lexical choice, *helmi*, 'pearl'. One could claim accordingly that the noun *helmi* bears a metaphorical meaning that is available in any syntactic context. There are other similar cases of syntactically "free" idiomatic nouns, one often cited example is that of *carrot and stick* (Moon 1998a: 96). This is all true, but in examples (20b-d) the syntax is still pretty regular, and the lexical choices are bound to the default ones. I would claim that all idioms have specific ways of expressing themselves: *heittää helmiä sioille* and *heittää kapuloita rattaisiin* look similar on the front but differ in what options they offer for a language user.

The status of the lexicon in Petrova's model has probably made the model especially receptive to lexical variation. The lexicon is seen as an interface between phonological, syntactic and conceptual levels. As a drawback, this sometimes complicates discussion, as lexical items are referred to in terms of their phonological structure.

How the model keeps track of encountered usages of phraseological units is not, however, clear to me. All non-default choices are after all not as predictable. Sometimes, it may even be hard to pick the default among many potential ones (cf. 18a-e). Actually, the linking system emerges from usage, and is continually modified usage-based. Instead of one default link, there could be several, stronger or weaker links. In Petrova's model, non-default lexical choices are mostly licensed via conceptual structure (id. 317-344) or referential tier (id. 370-374). For instance, the door to constructional alternation opens through the conceptual structure of the predicate verb, in this case the verb *heittää* expressing caused motion. Cases like *black sheep of the EU* (10b above) fall into the referential tier, the substitution being partly based on extra-linguistic factors.

5 Practical Considerations

A phraseological unit is typically limited not only to the core lexical items but to a specific combination of restrictions and preferences with respect to inflectional features, grammatical patterns, contexts etc. All these factors should be taken into account when formulating a dictionary entry. However, the result should illustrate the most common patterns in usage, instead of overwhelming the user with unrelated details. I believe that one of the most successful ways to convey information on usage is to select representative examples (Fox 1987). It is also important to notice that phraseological units are not alike. For instance, the two *heittää* idioms cited in this paper behave differently in some respects even though they are almost identical morphosyntactically. For example, the verb *heittää* is not a clear default choice in the idiom *heittää kapuloita rattaisiin*, but it is in the idiom *heittää helmiä sioille*. The most common lexical variants are given in recent corpus-based phraseological dictionaries (CCID, SSIS).

6 Conclusion

Lexical variation within phraseological units raises theoretical and practical problems. The scope of lexical variation is quite wide overall and it is entangled with grammatical and contextual factors. One of the major questions is how phraseological units are learned and recognized in all their varying forms.

The view on phraseology and lexicon taken in this paper is usage-based. As I see it, all items are learned over and over again in contexts, and this leads to memorized items with a collection of information on different aspects, such as pronunciation, style, inflectional forms, conveyed meanings, and situational contexts. The memorized items are not stable, and there are numerous ways in which they can be modified. However, since these items are usage-based, earlier experiences on their modifiability guides the future variations.

Meanwhile practical considerations guide us on how to write articles in dictionaries. Dictionary users should not be overwhelmed with detailed information on something that does not help them to understand and use the current expression. Besides, language learners are not as dependent on dictionaries as they used to be: it is common practice today to search the net to check if a specific wording is in use or not. Moreover, it is not even possible to list all thinkable options that are available to a language user.

7 References

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

Burger, H., Buhofer, A. & Sialm, A. (1982). *Handbuch der Phraseologie*. Berlin: Walter de Gruyter.

CCID = *Collins COBUILD Idioms Dictionary*. 2nd edn. Glasgow: HarperCollins Publishers, 2002.

Čermák, F. (2001). Substance of idioms: perennial problems, lack of data or theory? In *International Journal of Lexicography*, 14(1), pp. 1-20.

Čermák, F., Hronek, J. & Machač, J. *Slovník české frazeologie a idiomatiky. Výrazy slovesné.* [‘Dictionary of phraseology and idiomatics, (part 3) verbal expressions’.] Praha: Academia, 1994.

Duden = *Duden Deutsches Universal Wörterbuch*. 6th edn. Mannheim: Dudenverlag, 2006.

DWDS = Das Projekt Digitales Wörterbuch der deutschen Sprache: DWDS-Kernkorpus. Accessed at: http://www.dwds.de [5/12/2011].

Ernst, T. (1981). Grist for the linguistic mill: idioms and “extra” adjectives. In *Journal of Linguistic Research*, 1, pp. 51-68.

Fillmore, C.J., Kay, P. & O’Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. In *Language*, 64, pp. 501-538.

Fox, G. (1987). The case for examples. In J.M.Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins Publishers, pp. 137-149.

FTC = Finnish Text Collection. Accessed at https://sui.csc.fi/group/sui/lemmie [2.4.2014]. Information on the contents at http://www.csc.fi/english/research/software/ftc.

Heinonen, T.R. (2013). Idiomien leksikaalinen kuvaus kielenkäytön ja vaihtelun näkökulmasta. [‘Idioms as lexical constructions: usage and variability’.] PhD thesis. University of Helsinki, Finland. Accessible at: http://urn.fi/URN:ISBN:978-952-10-8555-0 [17/03/2014].

Heinonen, T.R. (2007). Variation and flexibility within verb idioms in Finnish. In M. Nenonen, S. Niemi (eds.) Collocations and Idioms 1, Papers from the First Nordic Conference on Syntactic Freezes, Joensuu, 19-20 May, 2006. Joensuu: University of Joensuu, pp. 146-159.

Jezek, E., Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, 4, Corpus Linguistics and the Lexicon, pp. 7-22. Accessed at: http://lexis.univ-lyon3.fr/IMG/pdf/Lexis_4.pdf [17/03/2014].

KS = *Kielitoimiston sanakirja*. [‘Authoritative Dictionary of Contemporary Finnish’.] Helsinki: Institute for the Languages of Finland, 2012.

Moon, R. (2008). Conventionalized *as*-similes in English: a problem case. In *International Journal of Corpus Linguistics*, 13(1), pp. 3-37.

Moon, R. (1998a). Frequencies and forms of phrasal lexemes in English. In A.P. Cowie (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press, pp. 79-100.

Moon, R. (1998b). Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford: Clarendon Press.

Nikanne, U. (2005). Constructions in Conceptual Semantics. In J.-O. Östman, M. Fried (eds.) *Construction Grammars: Cognitive Grounding and Theoretical Extensions*. Amsterdam: John Benjamins, pp. 191-242.

NODE = *The New Oxford Dictionary of English*. Oxford: Oxford University Press, 2001.

ODEI = Cowie, A.P., Mackin, R. & McCaig, I.R. (1993) [1983]. *Oxford Dictionary of English Idioms*. [= *Oxford Dictionary of Current Idiomatic English, Volume 2*.] Oxford: Oxford University Press.

Petrova, O. (2011). Of Pearls and Pigs: A Conceptual-Semantic Tiernet Approach to Formal Representation of Structure and Variation of Phraseological Units. PhD thesis. Åbo: Åbo Akademis Förlag. Accessible also at: http://urn.fi/URN:ISBN:978-951-765-583-5 [17/03/2014].

Philip, G. (2008). Reassessing the canon: 'fixed' phrases in general reference corpora. In S. Granger, F. Meunier (eds.) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins, pp. 95-108.

Sköldberg, E. (2004). *Korten på bordet: Innehålls- och uttrycksmässig variation hos svenska idiom*. ['Cards on the table: Variations in content and expression in Swedish idioms.'] PhD thesis. Gothenburg: Meijerbergs institut för svensk etymologisk forskning, University of Gothenburg, Sweden.

SSIS = Muikku-Werner, P., Jantunen, J.H. & Kokko, O. (2008). *Suurella sydämellä ihan sikana. Suomen kielen kuvaileva fraasisanakirja*. ['Descriptive Dictionary of Finnish Phraseology'.] Jyväskylä: Gummerus.

Stefanowitsch, A., Gries, S.Th. (2005). Covarying collexemes. In *Corpus Linguistics and Linguistic Theory*, 1. pp. 1-43.

Wulff, S. (2008). Rethinking Idiomaticity: A Usage-Based Approach. London: Continuum.

Acknowledgements

I am grateful to Maria Vilkuna and three anonymous referees whose comments and suggestions helped me clarify my presentation on several points.

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

Chris Mulhall
Waterford Institute of Technology
cmulhall@wit.ie

Abstract

This paper undertakes an empirical investigation into the treatment of tripartite Italian idioms in selected monolingual Italian and bilingual Italian-English dictionaries. Tripartite idioms are phrasal constructs typically arranged into one of the three following syntactic forms: V+N+N, V+ADJ+N or V+N+ADJ. From an organisational viewpoint, these idioms are somewhat more problematic for lexicographers due to the presence of a third lexical constituent. Current lexicographical practice adopts a largely subjective approach to dealing with idioms, which for the most part are V+N forms, therefore those with a wider syntactic form require more considered decision making when determining their point(s) of entry in a dictionary. Certain tripartite idioms also collapse into binomial (N+N) or nominal-adjectival forms (ADJ+N/N+ADJ), thus placing a question mark over the necessity to record their verb element or not. Together, these issues contribute to making tripartite idioms one of the most acutely difficult phrasal categories for lexicographers. This paper examines the entry points of 100 tripartite Italian idioms in selected monolingual Italian and bilingual Italian-English dictionaries published within the last decade. Firstly, it discusses relevant theoretical viewpoints relating to the lemmatisation of idioms over the last 30 years. Thereafter, it focuses on the notion of an idiom as a lexicographical entry in consideration of its varied intrinsic features and how these are productive in selecting an appropriate entry point. The presentation of the empirical data in the following section is stratified according to the publishing house of the analysed dictionaries and gives a comparative discussion on their respective approaches to dealing with tripartite idioms. Finally, the last section unifies the theoretical arguments and practical approaches and proposes a theoretical framework model for lemmatising tripartite idioms to offer a more coherent and consistent platform for their organisation in monolingual Italian and bilingual Italian-English dictionaries.

Keywords: Italian Lexicography; Tripartite Idioms; Lemmatisation

1 The Lemmatisation of Idioms: Theoretical Viewpoints

Lexicographical theorising over the last 30 years has regularly put forward proposals to remedy the problematic issue of how best to position idioms within a dictionary text. In their entirety, these arguments aim to achieve a more considered and structured lexicographical coverage of idioms, but fail to reach a consensus on the most suitable approach. Certain theorists advocate embedding idioms within the microstructure but diverge significantly on the number and position of their assigned entry points, for example, there is little theoretical agreement on a single listing strategy: Petermann (1983) (notional point of entry); Burger (1989) (unspecified entry point); Lorentzen (1996) (noun entry point); Mulhall (2010) (lexico-semantic entry point). Contrastingly, Tomaszczyk (1986) suggests entering idioms under each of their constituent's lemmas. The unitary semantic function of idioms equates them to having a word-like function in the lexicon, which Al-Kasimi (1977), Gouws (1991) and Botha (1992) argue is a substantive rationale for their lemmatisation in a dictionary. This, as Gouws (1991:86) states, prevents an 'ambiguous reading' of an idiom's lexical status by dictionary users. Lemmatising idioms accurately portrays their semantic status in the lexicon, but its lexicographical practicality remains questionable and untried. Harras and Proost (2005) advance a bespoke entry model for idioms, configuring their point of entry in accordance with their semantic features; resulting in semantically opaque idioms being lemmatised and semantically interpretable idioms sub-lemmatised. Adopting this particular method would move dictionaries to a more semantic-based lemmatisation model for idioms but is potentially anomalous given its proposal of different organisational principles for the same category of phrases.

An overview of the proposed entry methods reveals that semantics motivates many decisions relating to the most appropriate entry point with Mulhall (2010) taking into account that any lemmatisation model must also incorporate the notion of lexical variability, which, according to Moon (1998) can occur in between 12 to 40 percent of idioms. Another characteristic feature of idioms that is less topical in lexicographical debates is that of syntactic form. Therefore, considering the different semantic, lexical and syntactic properties of idioms may offer a more robust decision-making platform for identifying their most suitable entry point in a dictionary.

The majority of Italian idioms fall into the standard syntactic category of Verb (V) + Noun (N) with potential syntactic expansion to V/V+N or V+N/N if lexical variation is permissible in either the verb or noun component. In such cases, a lexicographer only has two (or possibly three) available point of entry options. A more problematic subset is that of tripartite idioms, some of which may have two constituents of the same word class (V+N+N) or contain three distinct word classes with two different syntactic structures (V+ADJ+N or V+N+ADJ). Tripartite idioms are particularly challenging for lexicographers on a number of levels. Firstly, the subjective identification of the most important or prominent element becomes more complicated due to the presence of a third lexical component. In the case of V+N+N idioms, lexicographers favouring a noun-based listing model must, from the outset, decide whether the first or second noun element is the most appropriate entry point. Secondly, the

V+ADJ+N and V+N+ADJ categories contain a small number of expressions that retain their idiomatic identity in a nominalised form, for example, *a gonfie vele, duro d'orecchio, il nodo Gordiano, la pecora nera*, etc. Important issues arise from these syntactically truncated forms; such as the necessity to record their verb element(s) or not and the importance of the adjectival element in their syntactic binding and lexical identity. The entirety of these issues and the failure of lexicographical practitioners to adopt and integrate recent theoretical suggestions or propose alternative practical solutions contribute to the long-term status of idioms as arguably the most problematic dictionary entry.

2 Redefining Idioms for Lexicography

Theoretical linguistics offers a multitude of rich and varied definitions for an idiom, but these typically comprise singular, one dimensional features, such as 'fixed expression' or 'non-compositional' or describe it through vague terms of references, such as 'relatively fixed expression'. Idioms, by their nature, are a linguistic concept; therefore such definitions may not offer the requisite scope to give lexicographers a wider understanding of their form and behaviour to deal with them accurately in the context of a dictionary. Therefore, to ensure a more representative lexicographical treatment of idioms, it is important to factor in their three most salient characteristics; semantics, lexis and syntax. In consideration of these characteristics, Table 1 sets out three feature-specific maxims to reconsider idioms as both a lexicographic entry and linguistic unit.

Feature	**Definition**
Semantics	Idioms are a semantically complex and compositionally divergent subset of expressions. This often results in a clear semantic disconnect between the lemma as a stand-alone lexical unit and as an idiom constituent.
Lexis	Idioms display different layers of lexical fixity. Their potential variability necessitates a dictionary entry strategy that not only recognises variable expressions but also records them in a consistent and representative way.
Syntax	Idioms are syntactically heterogeneous. Therefore, the number and word class of idiom constituents within any given idiomatic frame may potentially influences the number and position of allocated entry points in a dictionary.

Table 1: Lexicographical Definitions of Idiom Features.

Redefining the notion of an idiom as a linguistic unit and a dictionary entry, the following operational definition is proposed to achieve a more holistic and tailored lexicographic treatment of idioms based on their inherent features:

Idioms are a category of multi-lexical, syntactically diverse expressions showing various degrees of semantic compositionality, some of which contain lexical constituents that can substituted for idiomatic equivalents.

This broader definition encapsulates the most salient characteristics of idioms; accurately portraying their status in the lexicon and describing features that are potentially influential in their lexicographical description and organisation.

3 Organisational Approaches to Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries

An analysis of monolingual Italian and bilingual Italian-English dictionaries from the eighteenth century onwards shows that idioms, as a lexicographic entry, failed to gain a centrality and an organisational foothold in the design and content of these reference works. This, in part, can be traced to their perceived linguistic impurity in eighteenth and nineteenth century Italian society, which resulted in their limited coverage in mainstream dictionaries. A change in this outlook came in the early twentieth century due to new linguistic models, burgeoning dictionary content and a reformatted microstructure. But in many dictionaries idioms still remained a peripheral entity; a notable exception to this was the *Sansoni-Harrap Standard Italian-English Dictionary* [SHSIED] (1970-1975), which made the singular attempt of this era to systematise the coverage of idioms, but its subjective,[1] rather than substantive, criteria failed to address this problem for dictionary users. Decisions about restructuring the organisation of idioms and providing this information to users remain relevant, but overlooked, in modern day lexicographical practice. This paper aims to provide further evidence identifying the need for a lexicographical reform, at least in an Italian context, in the approach to organising idioms. The research sample includes 100 tripartite Italian idioms; subdivided into the following syntactic categories: 50 V+N+N, 30 V+N+ADJ and 20 V+ADJ+N. Selecting and organising the empirical sample brought to light two recurring trends: the prominence of the V+N+N syntactic structures within the Italian tripartite subset and the presence of a high frequency verb[2] (HFV) element in 39 of 100 expressions. To gain a comparative insight into any converging or diverging entry strategies based on the expectations of Italian speaking users a monolingual Italian and bilingual Italian-English dictionary from three publishing houses formed part of the research corpus. Selecting different dictionaries from the same publishing houses allowed an interesting exploration into ascertaining whether or not certain publishing houses follow any systematic procedure when attempting to record tripartite idioms in their monolingual and bilingual reference works. The dictionaries used are as follows: *Il Sansoni Inglese* [ISI] (2006); *Il Sabatini-Coletti* [ISC] (2007); *Il Ragazzini* [ZIR] (2009); *Lo Zingarelli* [ZLZ] (2009); *Hoepli Dizionario Inglese* [HDIN] (2007) and the *Hoepli Dizionario Italiano* [HDIT] (2008).

[1] "The phrases, idiomatic expressions, proverbs, etc., that make up the phrase section are generally found under the first important word in the phrase" (SHSIED 1970: viii).

[2] The following Italian verbs occur with a high frequency across a number of different phrasal and idiomatic expressions: *andare, avere, dare, essere, fare, mettere, prendere, stare, tenere, venire*.

Dictionary Publishing House	Sansoni		Zanichelli		Hoepli	
Dictionary Name	Il Sansoni Inglese	Il Sabatini-Coletti	Il Ragazzini	Lo Zingarelli	Hoepli Inglese	Hoepli Italiano
V+N+N (N=50)						
Verb Entry	17	2	4	2	5	2
N1 Entry/N2 Entry	13/0	8/2	15/3	9/3	19/8	7/3
Double Noun Entry	0	12	7	11	2	13
Verb/Noun Combination Entries	12	19	14	20	4	20
Listed as a Nominalised Idiom	2	2	3	3	5	2
Showing a different Verb Element	1	3	2	0	2	2
Not Listed	5	2	2	2	5	1
V+ N+ADJ (N=30)						
Verb Entry	2	1	0	0	4	1
Noun Entry	15	1	0	7	11	5
Adjective Entry	0	5	11	0	3	1
Verb/Noun/Adjective Combination Entries	6	16	11	14	5	14
Listed as a Nominalised Idiom	1	3	7	4	2	6
Showing a different Verb Element	4	2	0	0	1	1
Not Listed	2	2	1	5	4	2
V+ADJ+N (N=20)						
Verb Entry	3	0	3	0	2	0
Adjective Entry	9	1	7	1	0	1
Noun Entry	0	4	0	5	9	3
Verb/Noun/Adjective Combination Entries	3	8	6	6	4	9
Listed as a Nominalised Idiom	1	0	1	2	0	1
Showing a different Verb Element	1	1	0	0	1	2
Not Listed	3	6	3	6	4	4

Table 2: Empirical Data on the Entry Points of Italian Tripartite Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries.

3.1 Il Sansoni Inglese (2006) and Il Sabatini-Coletti (2007)

An often lamented failing of dictionaries is their failure to provide any guidance to users about the exact location of idioms. A notable exception in this regard is ISI (2006), which in the preface clearly informs users where to locate such expressions with accompanying examples. In contrast, its monolingual equivalent, the ISC (2007), takes a different approach, instead exemplifying certain idioms without explicitly indicating their position in the dictionary (see Table 3).

Il Sansoni Inglese (2006)	Il Sabatini-Coletti (2007)
The phrases, idiomatic expressions and proverbs that make up the phraseology section are listed under the first key word contained in the expression (be it verb, noun or adjective). Therefore, for example, the proverb *he who pays the piper calls the tune* is found under the verb **pay** and the phrase *as hard as iron* is listed under the adjective **hard**. Likewise, the Italian proverb *le bugie hanno le gambe corte* is given under **bugia** and the phrase *cavalieri della Tavola Rotonda* under **cavaliere**. As an exception to this, certain extremely common verbs (*be, can, come, do, get, give, go, have, keep, let, make, must, put, take, will* in English and *andare, avere, dare, dovere, essere, fare, lasciare, mettere, potere, prendere, stare, tenere, venire, volere* in Italian) have been ignored in listing the phrases under the headwords. As a result, the phrase, *to get one's cards* is given under the headword **card**, and *prendere qcu. in castagna* is given under **castagna**. (*Il Sansoni Inglese* 2006:14)	All'interno delle parole piene che lo sviluppano, ma con grande evidenza, sono tratte anche **le unità polirematiche grammaticali**, ossia le locuzioni che hanno valore di preposizione, di congiunzione o di congiunzione testuale (*a conti fatti*, *a costo di*, *modo che*, *nella misura in cui*...). Ben diverso è il caso delle espressioni idiomatiche, tutte di senso figurato, che appartengono alla lingua comune e sono in genere ben familiari ai parlanti (*essere un pozzo di scienza*; *dare carta bianca*; *andare per le lunghe*; *tendere la mano*; *voltare pagina*; *cambiare registro*. Come appare evidente, queste fanno nesso fisso con un verbo). (*Il Sabatini-Coletti* 2007:16)

Table 3: Organisational Criteria for Idioms in Sansoni Publishing House Dictionaries.

Like the diversity of their organisational approaches to these entries, the empirical analysis also reveals disparities in the treatment of the same syntactic idiomatic categories in the ISI (2006) and the ISC (2007). On a general level, this divergence can be measured through the number of assigned entry points; for example, the ISI (2006) allocates a single entry to the 61/100 expressions in contrast to a multiple listing strategy favoured by the ISC (2007) for 64/100 expressions. A possible explanation for this different approach is the strict adherence by the ISI (2006) compilation team to inserting idioms under the first key word, but the application of this method is not entirely rigid. For example, data from the empirical sample reveal that 21/100 expressions are listed twice or more and 54/100 expressions are recorded directly in line with the information given in the preface. From an Italian speaking user perspective, locating tripartite idioms with a high frequency verb may prove more labourious in the ISI (2006) than in its monolingual equivalent, thus requiring the dictionary user to engage in the subjective assessment of whether the noun or adjectival element can be considered as the first key word.

3.2 Il Ragazzini (2009) and Lo Zanichelli (2009)

Listing patterns for tripartite idioms found in the two dictionaries from the Zanichelli publishing house reveal a largely unstructured arrangement, a problem exacerbated by the lack of any information detailing their location. This omission is problematic for users, but is, to a certain degree, offset by the multiplicity of idiom listings in both the ZIR (2009) and the ZLZ (2009). This pattern is apparent across all analysed tripartite syntactic groups in the monolingual version, in particular the V+N+N category with 33/50 expressions recorded twice or more. Similar patterns emerge in the ZIR (2009), but on a lesser scale, with a considerable number of V+N+N (23/50) and V+N+ADJ (15/30) expressions listed twice or more. Another overlapping feature of both empirical samples is the comparably higher number of tripartite idioms recorded in nominalised forms in Zanichelli dictionaries; accounting for 11/100 in the ZIR (2009) and 9/100 in the ZLZ (2009). The removal of the verb element is an inherent feature of certain tripartite Italian idioms, but it negates the objective of a dictionary, which is to record lexical items in their fullest and most descriptive forms. Furthermore, the ZLZ (2009) contains the equal lowest coverage of the analysed expressions with 13/100 not recorded.

3.3 Hoepli Inglese (2007) and Hoepli Italiano (2008)

Both Hoepli-published dictionaries follow individually contrasting, but somewhat consistent, approaches in their treatment of tripartite Italian tripartite idioms. The use of a noun-based entry strategy emerges clearly from the analysis of the three syntactic categories in the HDIN (2007), but its consistency is diluted by scattered recording of similar expressions under alternative entries. A noun entry strategy features most prominently in the V+N+N group (27/50), but the division of this into 19/27 under N1 and 8/27 under N2 is a microcosm of the internally inconsistent approach to their general recording. Keeping a consistency with the other monolingual dictionaries, the HDIT (2008) favours a multi-entry strategy for tripartite idioms, in particular V+N+ADJ (14/20) and V+ADJ+N (9/20) forms, but a similar incongruity to that found in the V+N+N group in the HDIN (2007) resurfaces in the HDIT (2008). In this case, 20/50 binomial idioms are found under a verb and noun element(s) with 13/50 inserted under both noun components. The rationale for recording under both noun elements may be explained by the presence of a HFV element, but this does not extend to the entire subset with *cercare un ago nel pagliaio*, *dire peste e corna di qualcuno*, *finire in una bolla di sapone* and *tirare sassi in piccionaia* not inserted in their verb entries.

4 Conclusion

Idioms appear to remain on the periphery of lexicographical importance, at least in an Italian context. Empirical data from monolingual Italian and bilingual Italian-English dictionaries reveal an inconsistent, unscientific approach to the coverage of tripartite idioms. Generally, idioms tend to be defined by their problematic status rather than their linguistic uniqueness, thus veiling their rich lexical, semantic and syntactic features. This requires a more analytical look at how these characteristics can be influential in systematising the overall accessibility of idioms for dictionary users as well as resolving a perpetual practical difficulty. Figure 1 presents an alternative entry model for tripartite Italian idioms, whether as a verb phrase (VP) or a nominalised form, on the basis of their syntactic composition.

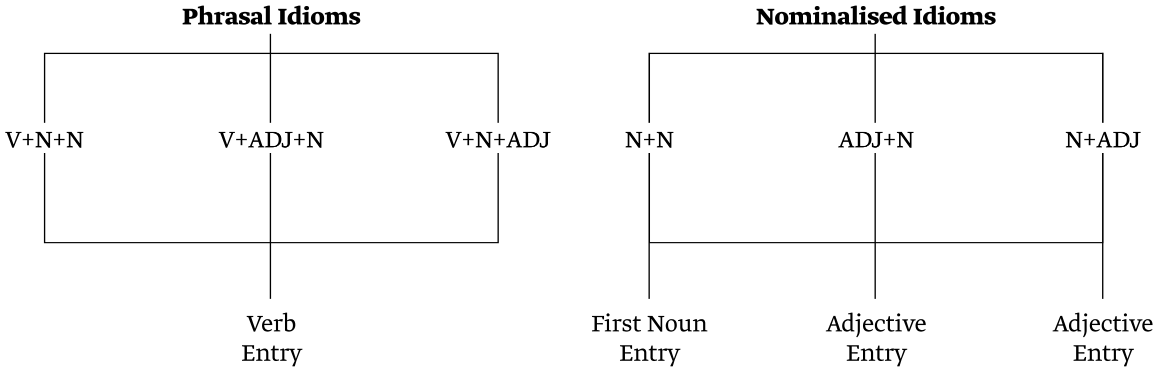


Figure 1: A Theoretical Framework for the Lexicographical Treatment of Tripartite Italian Idioms.

A universal verb entry strategy for tripartite idioms with HFV elements contrasts with current lexicographical practice as HFV entries are overpopulated and thus are considered too long for including such expressions. Tripartite idioms are generally VP structures, thus giving the verb element an elevated syntactic importance and identity for dictionary users. Its syntactic position at the head of the expression also increases it probability as a likely point of consultation for users. Nominalised forms of tripartite idioms present a greater organisational challenge, which unlike their VP equivalents, requires a more tailored entry model with due consideration afforded to syntactic order and word class. Therefore, recording N+N structures in their N1 entry and both ADJ+N and N+ADJ types in their adjective entries. These choices are predicated on the following pertinent criteria; the N1 element assumes the role of the syntactic head in binomial idioms, whereas the idiomaticity of those containing adjective and noun element is preserved by the retention of the adjective, compare, for example, the disparate meanings of *giocare a carte* and *giocare a carte scoperte* due to the presence of the idiomatically-inducing adjectival element *scoperte*. In conclusion, the multifaceted nature of idioms is complex, but also provides a substantive platform for choosing their most appropriate point of entry in a dictionary. The current systemic failure of dictionaries to address this issue reinforces the notion of idioms being subservient to words in both the lexicon and lexicography. Therefore, understanding and reprioritising the notion of an idiom and its associated features is an important objective for lexicographical practice in the twenty-first century.

5 References

Monolingual Italian Dictionaries

[ISC] *Il Sansoni Coletti* Dizionario della Lingua Italiana (2007). Milano: RCS Libri S.p.A.

[ZLZ] *Lo Zingarelli* (2009). Bologna. Zanichelli.

[HDIT] *Grande Dizionario Hoepli Italiano* (2008). Milano: Ulrico Hoepli Editore S.p.A.

Bilingual Italian-English Dictionaries

[ISI] Il Sansoni Inglese (2006). Milano: Edigeo.

[ZIR] Il Ragazzini (2009). Bologna: Zanichelli.

[HDIN] Grande Dizionario Hoepli Inglese (2007). Milano: Ulrico Hoepli Editore S.p.A.

[SHSIED] Sansoni-Harrap Standard Italian and English Dictionary by Vladimiro Macchi, Four Volume, Firenze-Roma: Sansoni Editore, 1970-1975.

Other Publications

Al-Kasimi, A. (1977). *Linguistics and Bilingual Dictionaries*. Leiden: E.J. Brill.

Botha, W. (1992). The Lemmatization of Expressions in Descriptive Dictionaries, in H. Tommola, K. Varantola, T. Salmi-Tolonen and J. Schopp (eds.). *EURALEX '92 Proceedings I-II*. Euralex International Congress, Tampere, August 4-9, 1992. Department of Translation Studies, University of Tampere. pp. 493-502.

Burger, H. (1989). Phraseologismen im allegemeinen einsprachigen Wörterbuch in F.J. Hausmann, F. Josef et al. (eds.), *Wörterbücher: Ein internationales Handbuch zur Lexikographie*. Volume I, Berlin/New York: De Gruyter. pp. 593-599.

Gouws, R. H. (1991). Toward a Lexicon-based Lexicography. *Dictionaries: Journal of Dictionary Society of North America*, **13**. pp. 75-90.

Harras, G. and Proost, K. (2005). The Lemmatisation of Idioms in H. Gottlieb, J.E. Mogensen and A. Zettersen (2005) (eds). *Symposium on Lexicography XI, Proceedings of the Eleventh International Symposium on Lexicography*. Copenhagan, May 2-4 2002 Tübingen: Max Niemeyer. pp. 277-291.

Lorentzen, H. (1996). Lemmatization of Multi-word Lexical Units: In which Entry? in in M. Gellerstam, J. Järborg, S.G. Malmgren, K. Norén, L. Rogström and C.R. Papmehl (eds.), *EURALEX '96 Proceedings I-II*. Seventh Euralex International Conference, Göteborg, August 13-18, 1996, Department of Swedish, Göteborg University. pp. 415-421.

Moon, R. (1998). Fixed Expressions and Idioms in English: A Corpus Based Approach. Oxford: Clarendon Press.

Mulhall, C. (2010). A Semantic and Lexical-Based Approach to the Lemmatisation of Idioms in Bilingual Italian-English Dictionaries in A. Dykstra and T. Schoonheim (eds.) (2010) *Proceedings of the XIV Euralex International Congress*. pp. 1355-1371.

Tomaszczyk, J. (1986). The Bilingual Dictionary under Review in M. Snell-Hornby (1986) (ed.) *ZüriLEX 1986 Proceedings, Euralex International Congress*. Zurich, 9-14 September, 1986 University of Zurich. pp. 289-297.

Especialización y Prototipicidad en Binomios N y N

Ignacio Rodríguez Sánchez
Universidad Autónoma de Querétaro
igrodsan@uaq.mx

Abstract

El trabajo que aquí se presenta es una investigación sobre binomios de estructura *N y N* (dos sustantivos unidos por la conjunción *y*). Esta investigación tiene un carácter exploratorio y se inscribe en la corriente neofirthiana de la lingüística de corpus. A nivel teórico y metodológico partimos de un enfoque guiado por datos, basándonos en los siguientes corpus: Corpus del Español, CORDE, Googlebooks, y EsTenTen11 (Sketchengine).

Aquí se abordan asuntos relacionados con la naturaleza de este tipo de colocaciones: su frecuencia, la estadística de la información mutua, la dispersión y el grado de reversibilidad para identificarlas; las relaciones que se establecen entre ambas partes; la especialización del nodo (como primera o segunda parte de los binomios) y, finalmente, la prototipicidad de algunos elementos Se concluye que el concepto de binomio tal como se entendía a partir de su definición clásica, se diluye, y se propone una visión integrada en un sistema dinámico de interacciones complejas e impredecibles.

Keywords: sustantivo y sustantivo; binomio; colocación; lingüística de corpus

1 Antecedentes

La definición de Sinclair del principio de idiomaticidad representa para algunos el momento de un cambio de paradigma en la lingüística: "*The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments*" (Sinclair, 1991: 110). Emparentadas con este trabajo, cristalizan visiones del lenguaje como la gramática de patrones (Hunston & Francis, 2000), la teoría de activación léxica (Hoey, 2005), los trabajos de Biber sobre grupos léxicos (1999; Biber & Barbieri, 2007), el análisis coloconstruccional, y otros trabajos como Wray (2002, 2008) y Corrigan et al. (2009a, 2009b).

La definición clásica de binomio corresponde a Malkiel (1959), que en su estudio comparativo de binomios irreversibles en varias lenguas los define como "[...] *the sequence of two words pertaining to the same form-class, placed on an identical level of syntactic hierarchy, and ordinarily connected by some kind of lexical link*". Según García-Page (2008: 347), un "binomio fraseológico comprehende, básicamente, las construcciones simétricas compuestas por dos sintagmas coordinados y los esquemas prepositivos, y, marginalmente ciertas construcciones asindéticas o yuxtapuestas". Para muchos autores (Almela Pérez, 2006; García-Page, 1998, 2008; Malkiel, 1959) un binomio digno de estudio es básicamente irreversible, mientras que para otros (Moon, 1998) puede no serlo. Un binomio irreversible sería el que no permitiera revertir el orden de sus dos componentes, como por ejemplo *a tontas y a locas*, *coser y cantar*, *cal y canto*. A diferencia de otras expresiones fraseológicas, el significado del binomio a veces sí se puede deducir de la suma de sus partes.

En este trabajo esperamos mostrar que es productivo estudiar los binomios desde una perspectiva que los incluya como colocaciones y no exclusivamente como expresiones idiomáticas. En la tradición fraseológica, sin embargo, hay posturas encontradas sobre este asunto. García-Page (2008: 12) argumenta que la colocación no es una estructura fija y que por tanto no debe ser objeto de estudio de la fraseología (concepción estrecha de la fraseología). Sin embargo, él mismo reconoce la dificultad de establecer definiciones claras: "¿Es *mesa redonda* o *dinero negro* una locución (…), una colocación o un compuesto?" (p.13). Por otra parte, otros estudiosos coinciden con nuestra visión: Corpas Pastor (1996: 52) propone una división de la fraseología en la que las colocaciones tienen perfecta cabida y que coincide plenamente con la de los fraseólogos anglosajones (concepción ancha de la fraseología).

Evert (2009: 1212) señala que el concepto de colocación es uno de los más controvertidos de la lingüística. Las diferencias entre lo que los neofirthianos y los fraseólogos entienden por ese mismo término ha creado una gran confusión en todos los campos. Desde nuestro punto de vista, coincidimos con Stubbs (1996: 172) sobre el hecho de que las intuiciones de los hablantes nativos sobre las colocaciones son muy imprecisas y no pueden de ninguna manera documentar con detalle dichas colocaciones. A una conclusión similar llega Alderson (2007) en el estudio en que compara datos de corpus y apreciaciones de lingüistas sobre la frecuencia de ciertas palabras.

Desde la psicolingüística, la teoría de activación léxica (*Lexical Priming Theory*) de Hoey (2005) supone un marco teórico adecuado para estudiar el fenómeno que nos proponemos abordar. Esta teoría considera que los hablantes hacemos de modo subconsciente complicadas asociaciones léxicas (semánticas, pragmáticas, de colocaciones y de coligaciones) con un género, estilo y situación social. Finalmente, Hoey sostiene que, también de manera subconsciente, percibimos la posición que ocupa una palabra en un texto, la cohesión que esta produce o deja de producir y las relaciones textuales que contribuye a formar. La teoría de activación léxica se apoya en principios psicolingüísticos como que las palabras de mayor frecuencia se activan antes y más en la mente de los hablantes, lo cual favorece un acceso rápido y fácil al lexicón. Esta activación favorece (y a la vez limita y restringe) la combinación entre las palabras. Si, como dice Giammarresi (2010: 262), almacenamos en el lexicón secuencias formulaicas enteras para ahorrar esfuerzo en el procesamiento, entonces es más lógico esperar que, dada una elección entre dos formas de transmitir un mensaje, una formulaica y otra no formulaica, sea la primera opción la que se genere antes.

En español, además del texto de García Page (2008) hay dos excelentes trabajos sobre binomios: el ya mencionado de Almela Pérez (2006) sobre binomios irreversibles y otro del propio García-Page (1998) sobre binomios antitéticos. Ambos trabajos tienen un carácter descriptivo basado en (no guiado por) corpus. Estos trabajos contienen listados de binomios que, aparentemente, se han ido recogiendo de manera intuitiva. Para nuestra investigación y por motivos de espacio nos referiremos exclusivamente a algunos datos que nos proporciona Almela Pérez (2006).

Los criterios que usa Almela Pérez (2006, págs. 141-146) para definir qué es un binomio son:

(1) Los binomios constan de dos lexemas.

(2) Son una secuencia infratextual.

(3) Tienen una estructura paralelística.

(4) Sus formas –léxicas y funcionales- son inmutables.

(5) Forman una secuencia indescomponible.

(6) Los miembros son inseparables.

(7) Tienen un significado composicional o idiomático.

2 Preguntas de Investigación

Nuestra intención es señalar que, aparte de los criterios mencionados arriba, hay explicaciones de tipo psicolingüístico (basadas en datos cuantitativos) que contribuyen a explicar las formas que adoptan los binomios:

(1) ¿La frecuencia con que ocurre el binomio y los índices de relación (en concreto la Información mutua –IM en adelante) pueden servir para identificar los binomios? Es decir, si una colocación binomial tiene 20 casos en un corpus y una información mutua de 8 ¿habremos identificado a un binomio irreversible?

(2) ¿Qué relaciones se establecen entre N_1 y N_2? ¿La frecuencia de uso de cada una de las palabras que constituyen el binomio va asociada a su posición en el binomio? Tomando el ejemplo de *aventuras y desventuras*, el hecho de que *aventuras* sea más frecuente que *desventuras* ¿nos está diciendo algo sobre el orden en que ese binomio se lexicalizó?

(3) ¿Se especializan algunas palabras en ser N_1 o de un binomio (siendo N_1 la primera parte del binomio, y N_2 la segunda)? Es decir, si sabemos que la palabra *señor* siempre aparece como N_2 cuando se combina con ciertos sustantivos de un mismo campo semántico (*amigo y señor, amo y señor, dueño y señor, esposo y señor, marido y señor, padre y señor, primo y señor, tío y señor* que aparecen como N_1), ¿no es acaso lógico que el binomio *rey y señor* también haya lexicalizado en el mismo orden que sus cohipónimos?

(4) Aparte de las relaciones semánticas, ¿qué otro tipo relaciones se establecen entre los diferentes colocativos de un binomio?

3 Extracción de datos a partir de un corpus

Los datos para este estudio proceden en primera instancia de Corpus del Español ([Davies, 2002] en adelante, CDE). En este corpus se realizó la búsqueda [NN*] Y [NN*] de los siglos XIX y XX, con un límite de 5000 casos. Del resultado obtenido se eliminaron los casos con frecuencia menor a cuatro y quedaron 2482 colocaciones binomiales. Por razones prácticas, se eliminaron posteriormente los casos de binomios con sustantivo repetido (por ejemplo, *años y años*). Nuestra base de datos original tampoco incluyó los binomios en los que se intercalan artículos, posesivos o preposiciones entre los dos sustantivos.

En una investigación previa (Rodríguez, 2013: 291) se mencionó la conveniencia de contrastar esos datos con los de corpus más grandes, que es lo que se empieza a hacer en esta investigación, complementando los datos del CDE con los del CREA, Googlebooks (interface de Mark Davies), y EsTenTen11 (Sketchengine).

4 Resultados

4.1 Información Mutua y Frecuencia

En respuesta a la primera pregunta de investigación, se calculó un índice de relación (IM) entre los dos términos del binomio para cada caso, siguiendo la fórmula basada en Oakes (1998, págs. 63-65). Según Evert (2009: 1229), la IM se debe combinar siempre con una frecuencia mínima (en nuestro caso 4) para equilibrar el sesgo hacia palabras de alta frecuencia.

Tal y como se esperaba, esta estadística mostró que la inmensa mayoría binomios tenían una IM significativa. Solo 140 binomios de los 2482 tenían una IM inferior a 3, que es, según Hunston (2002: 71), la cifra a partir de la cual se suele considerar que una colocación es significativa. Ejemplos de colocaciones y binomios cuya IM mutua es menor a 3 pero con cierto grado de fijación serían: *tiempo y forma, fondo y forma, padre y señor, tierra y libertad, forma y manera, cuerpo y sangre, vida y muerte* (5% de los casos). Esto apunta la imposibilidad de distinguir exclusivamente por métodos estadísticos una colocación de una no colocación (tal y como mencionan Evert [2009: 1242] y Cantos & Sánchez [2002]).

La frecuencia y la IM, pues, pueden ayudar a descubrir formulaicidad de un binomio pero puede que no sean los únicos criterios que los identifiquen. Junto a estos dos criterios, el de la dispersión es otro elemento útil, que contribuye a eliminar elementos estilísticos propios de los textos incluidos en el corpus. Por ejemplo, “Silvicultura y pesca” es un binomio que se repite 67 veces en el corpus pero solo aparece en un texto, que es una enciclopedia.

Sobre la segunda pregunta de investigación, es decir el orden distributivo de los dos términos de los binomios (las relaciones de N_1 con N_2) encontramos dos tipos de explicaciones. Por un lado, García-Page (2008: 347) señala que las posibles explicaciones del orden de los binomios son tanto de tipo semántico ("los principios del "egocentrismo" (...), de jerarquía social (...), de ordenación cronológica u espacial (...), de disposición de contrarios") como "fonéticos, morfológicos y léxicos". Estamos de acuerdo con él cuando comenta que este es un tipo de trabajo pendiente en el español, pero creemos que la semántica cognitiva (la tesis de la cognición corporeizada que desarrollaron casi simultáneamente Talmy [1988] y Johnson [1990] parece ser el referente de "los principios del egocentrismo") tendría algo que añadir en este sentido, en especial la integración (*blending*) conceptual que sugieren Fauconnier & Turner (2003).

Por otro lado, García-Page (2008: 348) hace referencia a que "el orden distributivo también se ha querido ver a veces en el esquema que dispone el constituyente silábicamente más corto en el primer lugar del binomio y el más largo en el segundo (...) una prueba del valor icónico de los binomios". De ser cierta esta apreciación -no se aportan datos empíricos en español- nosotros optaríamos por ofrecer para este fenómeno una explicación de carácter estadístico y psicolingüístico: simplemente que la primera parte del binomio es más frecuente que la segunda, como mencionamos antes al hablar de la teoría de activación léxica de Hoey. Este orden basado en la frecuencia permitiría al hablante un acceso más rápido al lexicón y facilitaría, por tanto, la fluidez en la comunicación. Lo corto de las palabras más frecuentes (y en primer término) vendría explicado por el corolario a la ley de Zipf (1949) que se refiere a que existe una relación directa entre la longitud de una palabra y su frecuencia (Davies, 2006: 164).

Así, en el binomio *tiempo y forma*, N_1 (*tiempo*) tiende a ser más frecuente que N_2 (*forma*) y en el binomio *oferta y demanda*, N_1 (*oferta*) tiende a ser más frecuente que N_2 (*demanda*). Esta tendencia viene confirmada por los datos de la Tabla 1.

Casos en que N_1 es más frecuente que N_2	Casos en que N_2 es más frecuente que N_1	Casos en que N_1 es tan frecuente como N_2	**Total**
784	462	17	**1263**

Tabla 1: Comparación de la frecuencia de uso de N_1 y N_2 de los binomios.

4.2 Dispersión

El valor de la estadística de información mutua combinado con una medida simple de dispersión contribuye a una mayor precisión en la identificación de binomios. Del total de 2483 binomios identificados inicialmente, 320 aparecen en el CDE de Mark Davies en solo un texto (independientemente de la frecuencia), lo cual equivale al 12.9% del total de binomios. Asimismo, se procedió a buscar en el CORDE el número de documentos en los que aparecía cada uno de los 2483 binomios. En este caso se obtuvo que 446 (18%) aparecían como máximo en un solo documento (de estos, 277 no aparecen ni una sola vez).

El número de binomios que reunían ambas condiciones (solo un documento del CDE de Mark Davies y uno o ninguno en el CORDE) es de 198 (es decir, el 8%) .

4.3 Reversibilidad

En cuanto a la irreversibilidad del binomio, ya se ha mencionado que esta es la característica esencial del binomio (Almela Pérez, 2006: 155; Malkiel, 1959: 113; García-Page, 2008: 329). García-Page es el único que reconoce que hay excepciones a la irreversibilidad. Según nuestra investigación, la reversibilidad de los binomios es un fenómeno que ocurre con mayor asiduidad de lo que suele considerarse.
En primer lugar, la en cuanto a la irreversibilidad, hay 346 binomios revertidos (del total de 2483 binomios), lo cual supone un 28% del total. Quince ejemplos y sus frecuencias aparecen en la Tabla 2
En la lista de Almela Pérez (2006) hay aproximadamente unos 90 binomios que coinciden con la estructura de los binomios que analizamos nosotros en nuestra base de datos. De estos 90 analizamos una muestra de 30 binomios que encajan con la estructura que se examina en este trabajo. Así, en la lista de binomios irreversibles que da Almela aparecen binomios que en realidad son reversibles, como *día y noche* y *noche y día* (Almela Pérez, 2006: 148-9); y otros casos como *pan y agua*, *cuerpo y alma*, *besos y abrazos*, *calidad y cantidad*, *cielo y tierra*, *uñas y dientes*, y *pies y manos*). Según nuestras estimaciones, un 30% de los binomios que Almela Pérez lista como irreversibles, no lo son en realidad.

Binomio	**FREC.**	**Binomio revertido**	**FREC.**
hombres y mujeres	426	hombres y mujeres	53
oro y plata	213	plata y oro	54
día y noche	171	noche y día	107
cuerpo y alma	136	alma y cuerpo	22
puertas y ventanas	136	ventanas y puertas	15
blanco y negro	136	negro y blanco	4
mujeres y niños	120	niños y mujeres	17
flora y fauna	116	fauna y flora	20
calles y plazas	100	plazas y calles	24
pies y manos	85	manos y pies	22
usos y costumbres	82	costumbres y usos	9
radio y televisión	80	televisión y radio	6
sangre y fuego	73	fuego y sangre	14
petróleo y gas	65	gas y petróleo	7
flor y nata	46	nata y flor	5

Tabla 2: Ejemplos de pares de binomios revertidos en el CDE con sus frecuencias.

La Tabla 2 muestra que en la mayoría de los casos, hay una gran diferencia entre las frecuencias de los dos binomios. Hay varias posibles explicaciones para esta situación. Por un lado podemos estar ante casos en que los procesos de lexicalización de la colocación no han terminado todavía pues el binomio supuestamente irreversible todavía no ha desplazado completamente al binomio revertido. Otra explicación, tal vez complementaria, es que un binomio, puede usarse solamente en determinados contextos.

Para nosotros resulta, pues, importante señalar que más que hablar de binomios irreversibles tal vez sea más fértil hablar de un cierto tipo de colocaciones, en N_1 *y* N_2, con un alto grado de lexicalización.

4.4 Especialización como N_1 o N_2

Con la muestra de 30 binomios N1 y N2 de Almela Pérez realizamos un análisis más pormenorizado.

Binomio	N_1 y N	Nº de tipos	N y N_1	Nº de tipos	N_2 y N	Nº de tipos
acoso y derribo	acoso y N	5	N y acoso	2	derribo y N	0
agua y ajo	agua y N	107	N y agua	95	ajo y N	6
ajos y cebollas	ajos y N	7	N y ajos	2	cebollas y N	0
alfa y omega	alfa y N	3	N y alfa	0	omega y N	0
armas y bagajes	armas y N	100	N y armas	70	bagajes y N	4
besos y abrazos	besos y N	19	N y besos	13	abrazos y N	12
bombo y platillo	bombo y N	3	N y bombo	1	platillo y N	0
bromas y veras	bromas y N	16	N y bromas	12	veras y N	3
cal y canto	cal y N	12	N y cal	12	canto y N	0
calidad y cantidad	calidad y N	80	N y calidad	75	cantidad y N	22
capa y espada	capa y N	11	N y capa	12	espada y N	19
cara y cruz	cara y N	33	N y cara	45	cruz y N	18
carne y hueso	carne y N	46	N y carne	41	hueso y N	14
carretera y manta	carretera y N	5	N y carretera	3	manta y N	2
causas y efectos	causas y N	24	N y causas	16	efectos y N	21
cielo y tierra	cielo y N	25	N y cielo	11	tierra y N	90
ciencia y conciencia	ciencia y N	61	N y ciencia	32	conciencia y N	38
cruz y raya	cruz y N	18	N y cruz	16	raya y N	1
cuenta y riesgo	cuenta y N	23	N y cuenta	37	riesgo y N	16
cuerpo y alma	cuerpo y N	59	N y cuerpo	33	alma y N	43
día y noche	día y N	34	N y día	11	noche y N	24
garbo y salero	garbo y N	15	N y garbo	7	salero y N	2
golpe y porrazo	golpe y N	10	N y golpe	4	porrazo y N	0
ida y vuelta	ida y N	0	N y ida	0	vuelta y N	7

Binomio	N_1 y N	Nº de tipos	N y N_1	Nº de tipos	N_2 y N	Nº de tipos
moco y baba	moco y N	3	N y moco	0	baba y N	1
pan y agua	pan y N	54	N y pan	30	agua y N	107
pecho y espada	pecho y N	36	N y pecho	15	espalda y N	13
sangre y fuego	sangre y N	103	N y sangre	84	fuego y N	34
uñas y dientes	uñas y N	8	N y uñas	8	dientes y N	25
viento y marea	viento y N	19	N y viento	15	marea y N	2

Tabla 3: Análisis de 30 binomios según el número de combinaciones con otros sustantivos (CDE de Mark Davies).

La Tabla 3 muestra en cuántos tipos de colocaciones aparecen como nodos N_1 y N_2. Por ejemplo en cuanto al primer binomio de la tabla, *acoso y derribo*, encontramos 5 combinaciones del N_1, *acoso*, con otro sustantivo (*acoso y abuso, acoso y vigilancia, acoso y protección, acoso y persecución* y *acoso y derribo*); dos casos en que *acoso* aparece como segunda parte del binomio (*violación y acoso* y *persecución y acoso*). No se registran casos en que el N_2, *derribo*, sea primera parte de otro binomio. Es decir que, según nuestros datos, parece que *acoso* se especializa como N_1 pues aparece predominantemente en esa posición. Si hacemos una comparación entre las columnas tres y cinco, tenemos que, de los treinta ejemplos tomados, hay solo 3 en que N y N_1 tiene más tipos de binomios que N_1 y N (se trata de los casos *capa y espada, cara y cruz* y *cuenta y riesgo*). En estos casos, *espada, cruz y riesgo* son nodos que se combinan más veces como N_2. Hay también tres casos en los que la colocación N_1 y N, tiene los mismos tipos que *N y* N_1 (*cal y canto, ida y vuelta* y *uñas y dientes*). Excluidos estos 6 casos, el 80% de los casos el nodo más productivo es N_1.

Si comparamos del mismo modo el número de tipo de colocaciones N_1 *y N* con N_2 *y N* (columnas 3 y 7), tenemos resultados muy similares: 83% de predominio de N_1 y solo 17% de predominio de N_2 (*capa y espada, cielo y tierra, ida y vuelta, pan y agua, y uñas y dientes*).

¿Se puede deducir pues que las palabras que están en la primera parte de un binomio tienen una potencia combinatoria mayor y esta ayuda a mantener en esa posición? La respuesta es sí, pero con salvaguardas: los casos en que no hay especialización clara (por ejemplo *valor*) y los de binomios que se especializan como N_2 como *confianza, director* y *libertad*.

4.5 Prototipicidad

En cuanto a las relaciones entre colocativos, un análisis somero realizada con los corpus más grandes arrojan resultados comparables a las relaciones de N_1 y N_2.

Aparte de las relaciones semánticas esperables (hiponimia, sinonimia, hiperonimia, etc.) y las relaciones morfológicas lógicas (misma categoria gramatical), resulta llamativo hallar relaciones sobre la estructura léxica relacionadas en el sonido (*clang association*): con el número de sílabas, la posición de la sílaba tónica, similitudes fonéticas y el uso de afijos. Aquí nos referimos a relaciones entre colocativos, no entre elementos de un mismo binomio como en *troche y moche* y *tomo y lomo*.

Por ejemplo en la Tabla 4 vemos que de las 10 primeras colocaciones de *sumisión* ordenadas por su IM, hay 5 que son palabras llanas de cuatro sílabas 4 de las cuales tres terminan en *-ismo,* y hay dos parejas que son parecidas en su inicio y su final: *obsecuencia* y *obediencia,* y *abyección* y *abnegación.*

	Binomio	**Frecuencia**	**IM**
1	servilismo	42	8.24
2	obediencia	190	7.74
3	obsecuencia	17	7.15
4	vasallaje	12	6.9
5	entreguismo	10	6.64
6	docilidad	11	6.61
7	conformismo	15	6.33
8	pasividad	31	6.13
9	abyección	6	6.13
10	abnegación	10	5.97

Tabla 4: Colocativos en binomios de *sumisión y* (Fuente: www.SketchEngine.co.uk, EsTen-Ten11, American, TreeTagger).

No es infrecuente hallar este tipo de situación. La Tabla 5 muestra que entre las diez colocaciones más frecuentes de *privaciones,* tiene gran parecido varias parejas: *austeridades y penalidades, escaseces y estrecheces, pobrezas y durezas.*

	Palabra	Frecuencia	IM
1	escaseces	161	9.98
2	fatigas	1338	9.8
3	abstinencias	49	9.72
4	austeridades	45	9.55
5	penalidades	641	9.51
6	sufrimientos	1379	9.44
7	estrecheces	172	9.37
8	pobrezas	56	9.15
9	miserias	979	9.07
10	penurias	282	8.94
11	durezas	43	8.91

Tabla 5: Colocativos en binomios de *privaciones* y (Fuente: Googlebooks Spanish).

5 Conclusiones

En primer lugar conviene recordar que los trabajos sobre binomios, y el nuestro no es una excepción, simplifican el valor del binomio pues en realidad esta estructura nunca deja de ser parte de una secuencia mayor con una función específica.

Hecha esta acotación, creemos que vale la pena explorar los binomios partiendo de principios básicos de la psicolingüística, tomando en cuenta la frecuencia de las palabras que aparecen como N_1 y como N_2. Aunque haya motivos de toda índole, parece ser que la frecuencia de uso del N_1 tiende a ser más alta que la del N_2. También hemos mostrado que hay nodos que se especializan en N_1 o en N_2, y que normalmente los nodos que aparecen como N_1 forman más binomios que los $N_{2.}$

El tamaño y la configuración del corpus son factores capitales en estudios cuantitativos porque no solo determinan las frecuencias, las estadísticas de asociación; las medidas de dispersión sino que nos van a señalar la variabilidad de relaciones conceptuales (secuencia de tiempo, de espacio, causa-efecto, jerarquía, etc). Asimismo, podemos observar fenómenos insospechados como la coincidencia en algunas colocaciones en el inicio y/o el final de la palabra (efecto tina/bañera) y la similitud del número de sílabas, del patrón consonántico, la posición de la sílaba acentuada, coincidencias fonéticas.

En definitiva, creemos que estos datos exploratorios respaldan la tesis de Hoey de que los hablantes registramos el uso de palabras, su contexto, su posición y más aspectos de los que no parecemos estar conscientes. Queda pues diluido el concepto de binomio tal como lo conciben Malkiel (1959) y Almela Pérez (2006), pasando a formar parte de un sistema dinámico de interacciones complejas e impredecibles en que se influyen mutuamente el uso, el procesamiento, el aprendizaje y la estructura de la lengua (Ellis & Frey 2009).

Este trabajo señala que los métodos semiautomáticos generados en una investigación guiada por datos sobre los binomios arrojan resultados ricos y complejos que obligan a replantearnos formas básicas generadas por la introspección.

6 Referencias

Alderson, J. C. (2007). Judging the Frequency of English Words. *Applied Linguistics*, *28*(3), 383-409 doi:10.1093/applin/amm024

Almela Pérez, R. (2006). Binomios (irreversibles) en español. *LEA: Lingüística Española Actual*, *28*(2), 135-160.

Biber, D. (1999). Longman grammar of spoken and written English. Londres: Longman.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263-286. doi:10.1016/j.esp.2006.08.003

Cantos, P., & Sánchez, A. (2002). Lexical Constellations: What Collocates Fail to Tell. *International Journal of Corpus Linguistics*, *6*(2), 199-228. doi:10.1075/ijcl.6.2.02can

Corpas Pastor, G. (1996). *Manual de fraseología española*. Madrid: Gredos.

Corrigan, R., Moravcsik, E. A., Ouali, H., & Wheatley (Eds.). (2009a). *Formulaic language* (Vol. 1). Amsterdam/New York: John Benjamins Publishing Company.

Corrigan, R., Moravcsik, E. A., Ouali, H., & Wheatley (Eds.). (2009b). *Formulaic language. (Vol. 2 Acquisition, loss, psychological reality, and functional explanations).* Amsterdam/New York: John Benjamins Publishing Company.

Cowie, A. P. (2006). Phraseology. *Encyclopedia of language & linguistics*. Amsterdam: Elsevier.

Davies, M. (2002). *Corpus del Español: 100 million words, 1200s-1900s*. http://www.corpusdelespanol.org [12/032013]

Davies, M. (2006). A frequency dictionary of Spanish: core vocabulary for learners. Abingdon: Routledge.

Ellis, N. C. (1998). Emergentism, Connectionism and Language Learning. *Language Learning*, *48*(4), 631-664. doi:10.1111/0023-8333.00063

Ellis, N. C. (2008). The Dynamics of Second Language Emergence: Cycles of Language Use, Language Change, and Language Acquisition. *The Modern Language Journal*, *92*(2), 232-249. doi:10.1111/j.1540-4781.2008.00716.x

Ellis, N. C., & Frey, E. (2009). The psycholinguistic reality of collocation and semantic prosody (2). En R. Corrigan, E. A. Moravcsik, H. Ouali, & Wheatley (eds.), *Formulaic language (Vol. 2 Acquisition, loss, psychological reality, and functional explanations).* Amsterdam/New York: John Benjamins Publishing Company.

Evert, S. (2009). Corpora and collocations. En A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: an international handbook* (Vols. 1-2, Vol. 2, págs. 1212-1248). W. de Gruyter.

Fauconnier, G., & Turner, M. (2003). The Way We Think: Conceptual Blending And The Mind's Hidden Complexities. Basic Books.

García-Page, M. (1998). Binomios fraseológicos antitéticos. En G. Wotjak (Ed.), *Estudios de fraseología y fraseografía del español actual* (págs. 195-202). Madrid/Frankfurt am Main: Iberoamericana/Vervuert.

García-Page, M. (2008). Introducción a la fraseología española. Estudio de las locuciones. Barcelona: Anthropos.

van Geert, P. (2008). The Dynamic Systems Approach in the Study of L1 and L2 Acquisition: An Introduction. *The Modern Language Journal*, *92*(2), 179-199. doi:10.1111/j.1540-4781.2008.00713.x

Gries, S. T., & Stefanowitsch, A. (2007). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin/New York: Walter de Gruyter.

Hoey, M. (2005). Lexical priming: a new theory of words and language. Abingdon: Routledge.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English.* Amsterdam/New York: John Benjamins Publishing Company.

Johnson, M. (1987). The body in the mind: the bodily basis of meaning, imagination, and reason. Chicago: University of Chicago Press.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, David Tugwell. *The Sketch Engine. Proc EURALEX 2004*, Lorient, France; Pp 105-116, http://www.sketchengine.co.uk [12/032013]

Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 21, 142-155.

Moon, R. (1998). Fixed expressions and idioms in English: a corpus-based approach. Oxford / New York: Clarendon Press.

Oakes, M. P. (1998). *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

Real Academia Española: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español actual.* <http://www.rae.es> [12/03/2013]

Rodríguez Sánchez, I. (2013). Frequency and Specialization in Spanish Binomials N y N. Procedia - Social and Behavioral Sciences, 95, 284-292. doi:10.1016/j.sbspro.2013.10.649

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford / New York: Oxford University Press.

Stubbs, M. (1996). *Text and corpus analysis*. Blackwell Publishers.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science: A Multidisciplinary Journal*, *12*(1), 49–100.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam/New York: John Benjamins Publishing Company.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. (2008). *Formulaic language: pushing the boundaries*. Oxford / New York: Oxford University Press.

Zipf, G. K. (1949). Human behavior and the principle of least effort: an introduction to human ecology. Cambridge Mass.: Addison-Wesley Press

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

Luigi Squillante
Sapienza - Università di Roma
Email: luigi.squillante@uniroma1.it

Abstract

In this work we present an empirical analysis performed on Italian nominal multiword expressions (MWEs) of the form [noun + adjective] that aims at studying quantitatively their syntactic and semantic features in order to improve their automatic identification and collection. Three indices are proposed, which are able to measure syntactic and semantic frozeness of the expressions on empirical basis in a corpus of about 1.8 million words, composed of Italian texts concerning the domain of physics. The combination of the three indices can be used to create a global measure, that we call Prototypicality Index (PI), which appears to be useful in the automatic extraction of terminological MWEs. The performance of PI at extracting true positives out of a candidate list is compared to those of the well-known statistical association measures Log-likelihood and Pointwise Mutual Information. Our results show how the performance of PI can be comparable to those of association measures, although it does not involve statistical calculations. Thus, PI can be seen as a new option for lexicographers and terminologists to integrate the already available statistical methods when identifying MWEs from texts.

Keywords: multiword expressions; terminology; prototype; extraction; empirical tests

1 Introduction

Nowadays multiword expressions (MWEs) represent one of the most studied phenomena in phraseological and lexicographic studies. They include a great variety of entities lying on a *continuum* between lexicon and syntax, whose typical features include morpho-syntactic fixedness, semantic restrictions, semantic unpredictability, constructions which differ from standard syntax, conventionality, institutionalization, etc. Their interpretation generally crosses the boundaries between words (Sag et al. 2002), and one of the best definitions to refer to such entities is proposed by Calzolari et al. (2002:1934), according to whom a MWE is "a sequence of words that acts as a single unit at some level of linguistic analysis".

Despite their apparent anomalous behavior, MWEs are a very important and frequent phenomenon in every language: in his famous *idiom principle* Sinclair (1991) states that idiomatic and morpho-syntactically restricted combinations are as normal and natural in discourse as free combinations, while Jackendoff (1997) attests that the number of MWEs stored in the lexicon of any speaker is equal to that of simple words.

Throughout the twentieth century, linguists have developed a great amount of studies which examined the aspects of MWEs on a theoretical perspective, often leading to competing analyses, controversy on interpretations or overlapping terminology. In recent years, however, computational and corpus-based studies have become one of the dominant lines of research in this field, since quantitative features, such as the fact that MWE components tend to cooccur in text with higher frequencies, have proved to be very effective in the automatic treatment of MWEs, leading to the development and improvement of several association measures (AMs) in order to identify, study and automatically extract MWEs from texts (just to mention some works: Evert 2004; Evert 2008; Kilgarriff 2006; Ramisch et al. 2010; Seretan 2011).

When analyzing a corpus, by means of computational tools, linguists are usually able to create a list of candidate expressions of MWEs where each candidate has an association score assigned by AMs. In general, the primary goal is to identify the largest possible number of true positives (candidates that represent real MWEs) within a certain threshold of significance based on the score assigned to each candidate, e.g. to provide raw material for lexicography. In this process, AMs generally consider statistical quantities, such as the number of cooccurrences of the components, the number of occurrences of the single components, the size of the corpus, etc., often with no reference to any explicit linguistic behavior. Nevertheless, considering syntactic or semantic features of MWEs from a computational and corpus-linguistic point of view is useful to improve the performances of automatic extraction tools (as shown, for other languages, in Bannard 2007; Weller & Fritzinger 2010; Cap et al., 2013), as well as to develop a better understanding of the typical features of MWEs on empirical bases (cf. Squillante 2014), which are both aspects of preeminent interest for lexicographers dealing with multiword phenomena.

2 Motivations

Our work presents an empirical study conducted on the Italian language which, unlike other major languages like English or German, still lacks well-founded computational studies in lexicography dealing with complex expressions like MWEs. Although “GRADIT - Grande Dizionario Italiano dell’Uso” (De Mauro, 1999-2007), known as the most comprehensive lexicographic resource for Italian, has a highly corpus-oriented perspective and explicitly focuses on the quantitative presence of MWEs in Italian, no explicit computational methods were involved in identifying the expressions. Similarly, the most recently published Italian collocation dictionaries (Urzì 2009; Lo Cascio 2012; Tiberii 2012) still rely mostly on intuition and only partly replicate data collection strategies, without considering a defined and explicit methodology based on corpora. Thus, there is a need to investigate computational techniques for lexicographic analyses of Italian MWEs, especially because Italian morpho-syntax differs from those of the above-mentioned Germanic languages.

The nature of our study is twofold: on the one hand we focus on empirical evidences in order to study the prototypical concept of MWE; on the other hand we compare the traditional statistical measures with syntactic or semantic tests for the identification and the extraction of MWEs from texts.
Finally, our work is focused on terminology. In fact, especially in technical domains, MWEs appear in high number even in small corpora, since specialized languages are a powerful source of multiword terminology and we see it as a matter of importance that they are identified and collected so that they can be included in the respective dictionaries and multiword terminology collections.

3 Methodology

3.1 Corpus and Prototype of MWE

As a first approach, in our study we opted to focus on the field of physics. The choice of physics is interesting since its lexicon, unlike other scientific domains such as that of medicine, is still primarily composed of highly polysemous every-day words which are put together in MWEs to form technical expressions, pursuing the established tradition started with Galileo Galilei in the seventeenth century, as recalled by Migliorini (1994:398).
In order to have an empirical base to perform our analysis, we built a corpus of about 1.8 million words collecting Italian texts concerning physics, including educational books (6,2% of the total), Wikipedia pages (34,5%), academic textbooks (20,7%), theses and dissertations (38,6%).
Our corpus was POS-tagged with TreeTagger (Schmid, 1994) and enhanced by means of a semi-automatic and manual post-tagging process in order to improve the tagging quality, e.g. to correct macroscopic systematic errors and include unrecognized technical lemmas in the dictionary. The final accuracy of the tagged corpus is evaluated at around 96% by manually checking 300 random sentences of the corpus.
We chose to analyze only nominal MWEs of the form [noun + adjective] in a first approach, representing the unmarked Italian nominal phrase. In physics, in fact, the use of nominal phrases is dominant and nominalization is often attested to be a standard feature of special languages. This is also supported by the fact that the majority of MWEs labeled by GRADIT as part of the special language of physics are nominal (2668), while only 9 belong to any other grammatical category.
Although MWEs can exhibit a great variability of behaviors, as it has been mentioned in the introduction, we chose to focus on features which could be investigated and tested on corpora, and we started with the initial hypothesis that the prototype of a MWE is an expression:

- that does not allow for interruptions or insertions of other words between its components;
- whose word order is not modifiable;
- whose components cannot be substituted by their synonyms.

The expression *relatività generale* 'general relativity' is a clear example of a terminological MWE which satisfies these three conditions, since it cannot be interrupted (cf. **relatività più generale* 'more general relativity'), it does not allow a modification in the order of its components, although this is possible for Italian nominal phrases (cf. **generale relatività*) and it cannot be modified by substituting one of its components with a synonym (cf. **relatività universale* 'universal relativity' or **relatività totale* 'total relativity').

However, although these features involving fixedness are typically associated to nominal MWEs in Italian, they do not always appear together in all expressions. For example, interruptibility is allowed for *punto debole* 'weak point', which admits *punto più debole* 'weaker point'; *infrarosso lontano* 'far infrared' is attested together with *lontano infrarosso*; while *gas ideale* 'ideal gas' can be substituted by *gas perfetto* 'perfect gas'. Because of this, the concept of prototype is thought of just as a model which could help to order the expressions on a continuous scale from a maximum grade of fixedness on several levels (adhesion to the prototype) to more flexible expressions.

The reason for considering the hypothesis of such a prototype comes from studies like those of Masini (2009) and Squillante (2014), which show how the nucleus of the prototype seems to include those expressions that are generally referred to as *polirematiche* in the Italian lexicographic tradition and exhibit syntagmatic *and* paradigmatic frozeness, needing the cooccurrence of their components in order to acquire their specific meaning (e.g. *luna di miele* 'honeymoon'; *essere al verde* 'to have no money', lit. 'to be at green'). Terminological expressions are generally part of this group.

When fixedness becomes less strict and modification is allowed, the *continuum* of MWEs moves towards those expressions that we can call *lexical collocations,* which show only preference for the cooccurrence of their components (e.g. *capelli castani* 'chestnut brown hair' or *compilare un modulo* 'to fill a form'), being «not fixed but recognizable phraseological units» (Tiberii, 2012).

3.2 Three Indices for the Measure of Empirical Frozeness

Following Squillante (2014), we implemented a computational tool that performs empirical tests concerning the above-mentioned features of modifiability for each candidate expression. Each of the features is quantified by an index whose value is computed on the basis of the comparison between the occurrences of the modified expression and those of the regular basic unmarked form in the corpus, i.e. the lemmatized form, regardless of inflection (which our analysis proved to be not a relevant feature in discriminating MWEs from standard expressions). All the queries are made on surface forms or POS categories, depending on the test, and do not involve syntactic structures as they would arise from parsing.

Given an expression, the index of interruptibility (I_i) counts the number of the occurrences of the sequence in its basic form [noun + adjective], say n_i, and the occurrences of the same sequence with one word occurring between the two components (n_{bf}), calculating the following ratio:

$$I_i = \frac{n_i}{n_{bf} + n_i}¶$$

In this way, a high number of interrupted expressions with respect to those which are not interrupted let the index acquire a high value. The sum in the denominator let the index be limited between 0 and 1.

In an analogous way, the index concerning the reverse order (I_o) compares the number of occurrences of the inverted sequence [adjective + noun] (n_o) with those of the basic form n_{bf}, according to the formula:

$$I_o = \frac{n_o}{n_{bf} + n_o} \P$$

Finally, the index concerning the feature of substitutability compares the number of occurrences of the basic form with the occurrences of all the sequences in which one of the two components is replaced by one of its synonyms (if present). If the number of occurrences of the *i*-th synonym of the first and the second component are called respectively $n_{s1,i}$ and $n_{s2,i}$, the total number of substituted sequences for the expression is:

$$n_s = \sum_i n_{s1,i} + \sum_i n_{s2,i} \P$$

and the index I_s is given by the formula:

$$I_s = \frac{n_s}{n_{bf} + n_s} \P$$

The calculation of I_s is subjected to the availability of an external synonym list. In our study, as a first approach, we chose the GNU-OpenOffice Thesaurus for the Italian language[1] for practical reasons, since it was immediately available, easily manageable and proved to be good enough for our purpose. However, one can integrate the tool with other more specific resources in the future, in order to improve the quality of the results.

The values of the three indices can be merged into a single function that we call Prototypicality Index (PI), representing the adherence of the expression to the hypothesized prototype. We consider the following formula:

$$PI = \frac{n_{bf}}{n_{bf}^{max}} \cdot \frac{1}{1 + I_i + I_o + I_s} \P$$

whose value increases when the values of the three indices decrease (thus, a high PI value means high fixedness), and in which the three features are weighted in the same way by the operation of

[1] http://linguistico.sourceforge.net/pages/thesaurus_italiano.html.

sum. In this way an expression with a very high value for just one of the indices can have a resulting PI value similar to that of an expression with average values distributed on all the three indices. Therefore, this structure is useful to take into account the flexibility of the nature of MWEs. Finally, the PI considers a correction factor, given by the normalized ratio between the frequency of the expression and that of the most frequent candidate expression n_{bf}^{max}. This correction factor, which is bounded between 0 and 1, is needed to take into account the fact that low occurrences for the basic form in the corpus reduce the reliability of the empirical tests, since the presence or the absence of modifications cannot be tested on a large set of expressions.

4 Analysis and Results

As a first analysis, we considered the whole set of nominal MWEs labeled as part of the lexicon of physics in GRADIT. The considered set consists of a total amount of 1.551 MWEs, 595 of which are attested to occur in our corpus.

The resulting values of the three indices (considered separately) indicate that 73% of the attested expressions are never interrupted, 93% never appear in reverse order and 64% do not attest any substitution of their components. The empirical evidence, hence, suggests that the syntactic fixedness, more than paradigmatic frozeness, seems to be relevant in outlining the prototype of nominal MWEs in physic Italian terminology. It must be underlined that the absence of modifications in the corpus does not mean that the expression does not allow them in general, nevertheless the empirical evidence can be considered a good approximation in our computational perspective.[2]

Since the list of physics-related MWEs extracted from GRADIT is supposed to include only terminological expressions with a completely definite phraseological status, we can consider them as a gold standard for further analyses.

In fact, the PI can be used as a new measure for the automatic extraction of MWEs from texts.

On the basis of the PI values, it is possible to assign each expression of a list of candidates a score and order the expressions according to it.

In order to have empirical evidence of the performance of the PI, we considered an input list from our corpus, composed of all the bigrams of the form [noun + adjective] which were extracted automatically, forming a set of about 22.700 expressions.

If we order the list according to PI we obtain results which appear analogous to those generally produced by statistical AMs, since PI is able to filter out most non-MWE candidates, which get very low scores and are pushed to the end of the list. At the same time, expressions appearing with very high

[2] It must be said that some noise in this kind of approach is unavoidable, since it can happen that few expression can exhibit modifications, but the modified expressions are not MWEs anymore, as in the case of *forza debole* 'weak force' meaning one of the four fundamental interactions, which is attested together with *debole forza*, meaning just that the intensity of a generic force is weak.

scores at the top of the list have high probability of representing true MWEs. Table 1 and Table 2 show, respectively, the top and the end of the list sorted according to PI.

Rank	MWE candidate	English translation	PI value
1	Campo magnetico	Magnetic field	0.9565
2	Campo elettrico	Electric field	0.6133
3	Momento angolare	Angular momentum	0.5717
4	Meccanica quantistica	Quantum mechanics	0.5205
5	Calorimetro elettromagnetico	Electromagnetic calorimeter	0.4748
6	Modello standard	Standard model	0.4259
7	Valore medio	Mean value	0.4206
8	Massa invariante	Rest mass	0.3683
9	Energia cinetica	Kinetic energy	0.3630
10	Campo gravitazionale	Gravitational field	0.3423
11	Campo elettromagnetico	Electromagnetic field	0.3314
12	Relatività generale	General relativity	0.3155
13	Buco nero	Black hole	0.2997
14	Meccanica classica	Classic mechanics	0.2591
15	Carica elettrica	Electric charge	0.2395

Table 1: Top-15 of the candidate list made of all the [noun + adjective] bigrams attested in our corpus sorted according to the Prototyicality Index values.

Rank	MWE candidate	English translation	PI value
22686	Entità indipendente	Independent entity	$7.0651 \cdot 10^{-6}$
22687	Caso tale	Case such	$6.8689 \cdot 10^{-6}$
22688	Fotone due	Photon two	$4.8725 \cdot 10^{-6}$
22689	Condizione fondamentale	Fundamental condition	$4.4757 \cdot 10^{-6}$
22690	Sistema vivente	Living system	$4.3961 \cdot 10^{-6}$
22691	Parte maggiore	Bigger part	$4.3766 \cdot 10^{-6}$
22692	Ambito magnetico	Magnetic range	$3.9057 \cdot 10^{-6}$
22693	Condizione finale	Final condition	$2.6518 \cdot 10^{-6}$
22694	Dimensione media	Average dimension	$2.5493 \cdot 10^{-6}$
22695	Forma standard	Standard shape	$2.2079 \cdot 10^{-6}$

Table 2: End of the candidate list made of all the [noun + adjective] bigrams attested in our corpus sorted according to the Prototyicality Index values.

In order to evaluate the performance of the PI, we chose to compare its results on our candidate list with two well-known statistical association measures, Log-likelihood (Dunning 1993), hereafter LL, and Pointwise Mutual Information (Church & Hanks 1990), hereafter PMI, which are widely used in corpus-linguistics to identify MWEs. Both AMs can be seen as representatives of two general groups of measures which quantify two different aspects of word combinations: LL measures how unlikely it is that the two words are independent while PMI investigates "how much the observed cooccurrence frequency exceeds expected frequency" as stated in Evert (2008: 1128). In this way, their use can provide two different perspectives on the statistical extraction of MWEs.

By means of the computational tool "mwetoolkit" (Ramisch et al. 2010), each bigram of our candidate list is assigned a LL and a PMI value, so that all the expressions can be ordered according to their statistical scores. The performance of PI and the two measures is evaluated on the basis of the rate of the retrieval of true positives in the lists: we compare how many true MWEs are detected while going through the lists, according to the ordering established by the scores of statistical measures and PI.

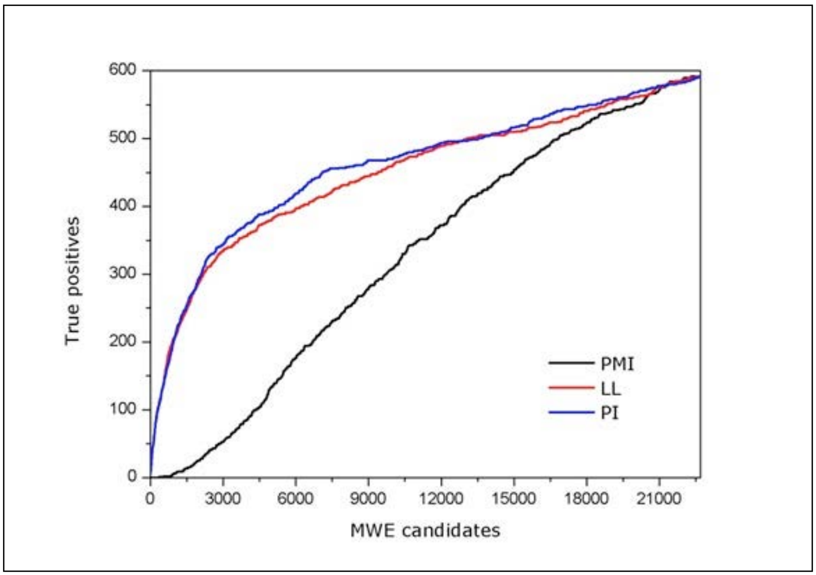


Figure 1: Comparison between the extraction rates of true positives of Pointwise Mutual Information (black), Log-likelihood (red) and Prototipicality Index (blue).

Figure 1 shows the curves representing the extraction rates for the three measures. As one can see, LL and PI had quite similar performances at identifying true positives, thus indicating that syntactic and semantic tests on empirical data can provide good results when used in extraction tasks. The poorer result of PMI can be justified by the fact that no frequency threshold was applied at the beginning and this AM is known for overestimating low-frequency expressions which are often false positives (Evert 2008).

We noted that for the first 1.800 candidates (corresponding to a 40% of true MWEs retrieved) LL obtained slightly better results with respect to PI, but for the remaining 20.900 candidates, the PI was almost always the better choice. This seems to indicate that on large scales the PI can be more useful to lexicographers, who are generally interested in retrieving the largest possible number of MWEs and not only those in the first positions of the lists generated by statistics.

As an additional analysis, we considered also a frequency threshold on the input candidate lists, in order to minimize the problems related to low-frequency expressions, which especially affect PMI. Thus, we filtered our list, keeping only expressions with a frequency $f \geq 30$ (for a total of 301 expressions) and performed the same procedure as above.

Since a frequency of at least 30 occurrences can provide a good empirical basis for the tests, we decided to consider in this case also a "pure" variant of the PI, which is not corrected by the frequency information and is given by the following formula:

$$PI_p = \frac{1}{1 + I_i + I_o + I_s}¶$$

Figure 2 shows the extraction rates for the four measures. Once again LL and PI are the best choices and their performances are almost equal. This time PMI appears to be more useful, as one could expect, although its extraction rate is less effective than LL or PI. Finally PI$_p$ shows an extraction rate which is clearly better than that of PMI for the first 80 candidates, while for the remaining candidates its performance can be comparable to PMI. At the end of the process the number of true positives retrieved was 101 for LL and PMI, 99 for PI and 98 for PI$_p$.

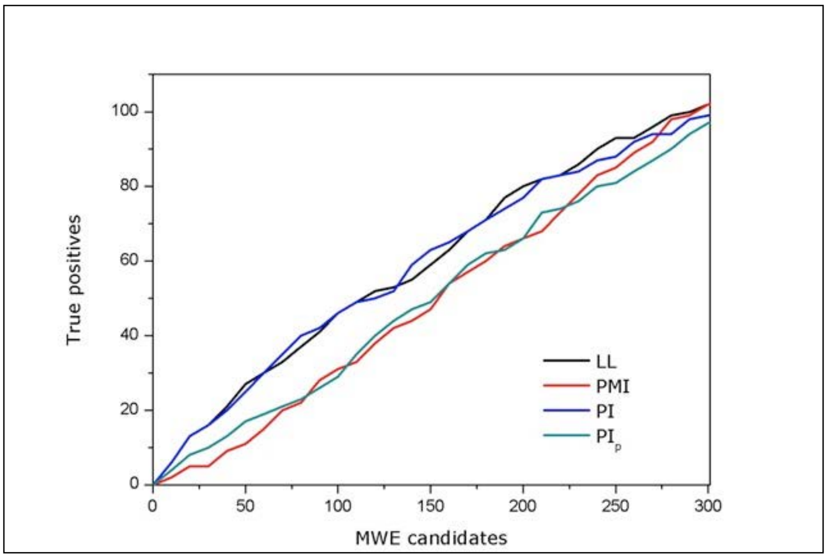


Figure 2: Comparison between the extraction rates of true positives for Log-likelihood (black), Pointwise Mutual Information (red), Prototypicality Index (Blue) and the pure version of PI (green) on a candidate list with a frequency threshold of 30 occurrences.

5 Conclusions and future work

In this work we have shown how syntactic and semantic features can play an important role in studying MWEs from a computational perspective. In the case of Italian nominal MWEs of the form [noun + adjective] belonging to the special language of physics, empirical tests performed on a corpus of 1.8 million words suggested that syntactic and semantic frozeness are effective features when outlining the prototype of this kind of expressions, although semantic substitutions are more tolerated than syntactic modifications.

The three indices that quantify empirical frozeness considered in this work proved effectiveness in extraction tasks of MWEs when merged in a function that we called Prototypicality Index, which produced results that can be considered comparable to those of statistical association measures.

Such results show how our methodology can be seen as a new option for lexicographers and terminologists, to integrate the already available statistical methods when identifying MWEs from texts, thus providing one more perspective in the extraction task which can be useful to have a more complete and general overview of the phenomena as well as to create complete terminological dictionaries or resources.

Moreover, as mentioned above, the PI works better on larger scales and appears to be useful to lexicographers who are interested in retrieving more efficiently MWEs when considering a high coverage, thus dealing with expressions spanning throughout the candidate list and not focusing only on its top. This feature of PI can be explained by the fact that syntax and semantics, unlike statistical features, show more strength and reliability when dealing with less frequent expressions.

Nevertheless, the fact that a simplified version of the PI, which does not involve frequency information, produced worse results (but still similar to AMs) on a limited candidate list composed by expressions with more than 30 occurrences, shows that frequency inevitably plays a role in helping the retrieval of true positives.

However, the empirical results presented in this work must be tested on larger and more general corpora, as well as on corpora of other specialized domains, in order to evaluate the usefulness of the PI for general and specialized lexicography.

Future works must include the development of tools which can deal with other pattern of nominal MWEs as well as other grammatical categories, such as verbal or adverbial MWEs, where the above-mentioned features of modifiability are to be used in different ways when defining the prototype.

Lastly, the tools developed are to be made available, e.g. as a part of corpus research workbenches, for lexicographers and terminologists.

6 References

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003).* Sapporo, Japan.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C. & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Canary Islands.

Cap, F., Weller, M. & Heid, U. (2013). Using a Rich Feature Set for the Identification of German MWEs. In *Proceedings of Machine Translation Summit XIV.* Nice, France.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, 16(1), pp. 22-29.

De Mauro, T. (1999-2007). GRADIT, Grande Dizionario Italiano dell'Uso. Torino: UTET.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. In *Computational Linguistics*, 19(1), pp. 61-74.

Evert, S. (2004). The Statistics of Word Cooccurrencies: Word Pairs and Collocations. PhD Thesis. University of Stuttgart.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, M. Kytö (eds.) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, pp. 1212-1248.

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge: MIT Press.

Kilgarriff, A. (2006). Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress.* Torino: Dell'Orso, pp. 997-1004.

Lo Cascio, V. (2012). *Dizionario combinatorio compatto italiano.* Amsterdam: John Benjamins Publishing Company.

Masini, F. (2009). Combinazioni di parole e parole sintagmatiche. In M. Catricalà, P. Pietrandrea, E. Lombardi Vallauri, P. Di Giovine, D. Cerbasi, L. Mereu, L. Gaeta, G. Fiorentino, P. D'Achille, M. Grossmann, E. Jezek, F. Masini, A. Pompei, E. Bonvino, F. Orletti, M. Frascarelli (eds.) Spazi linguistici. Studi in onore di Raffaele Simone. Roma: Bulzoni, pp. 191-209.

Migliorini, B. (1994). *Storia della lingua italiana.* Milano: Bompiani [I. ed 1960].

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). Mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the 7th International Conference on Language Resources and Evluation (LREC 2010).* Valletta, Malta.

Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd CICLing (CICLing-2002), vol. 2276/2010 of LNCS.* Mexico City, Mexico.

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

Seretan, V. (2011). *Syntax-based Collocation Extraction.* Berlin: Springer.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Squillante, L. (2014). Towards an Empirical Subcategorization of Multiword Expressions. To appear in *Proceedings of the EACL 10th Workshop on Multiword Expressions.* Gothenburg, Sweden.

Tiberii, P. (2012). Dizionario delle Collocazioni. Le combinazioni delle parole in italiano. Bologna: Zanichelli.

Urzì, F. (2009). *Dizionario delle Combinazioni Lessicali.* Luxembourg: Convivium.

Weller, M. & Fritzinger, F. (2010). A hybrid approach for the identification of multiword expressions. In *Proceedings of the SLCT 2010 Workshop on Compounds and Multiword Expressions.* Linköping, Sweden.

Historical Lexicography and Etymology

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

Diego Dotto
Opera del Vocabolario Italiano - CNR
dotto@ovi.cnr.it

Abstract

Il progetto *Divo* (*Dizionario dei volgarizzamenti*) si propone uno studio analitico del lessico di traduzione dei volgarizzamenti medievali e allo stesso tempo la costruzione di strumenti *ad hoc* per lo studio di questo lessico. Il progetto si compone di tre fasi: la prima è la compilazione di una bibliografia filologica secondo il modello *TLIon*; la seconda è la costruzione di un corpus lemmatizzato, in cui ciascun testo volgare è associato paragrafo per paragrafo all’originale latino; il terzo e ultimo punto è lo studio del lessico di traduzione, cioè il lessico identificato come una traduzione diretta dal latino.

Keywords: Lessicografia; Italiano antico

1 Lo studio lessicale dei volgarizzamenti dei classici dal cantiere del *DiVo*

Il progetto *DiVo*, ideato, promosso e diretto da Elisa Guadagnini e Giulio Vaccaro presso l’Opera del Vocabolario Italiano e la Scuola Normale Superiore di Pisa, ha l’obiettivo di studiare analiticamente il lessico dei volgarizzamenti italoromanzi dei testi classici e tardo-antichi, con limiti fissati al VI secolo (in particolare all’opera di Boezio), per i testi di partenza, e al XIV secolo per i testi di traduzione (con l’inclusione, in casi particolari, di volgarizzamenti realizzati a cavallo tra il XIV e il XV secolo). Per fare questo, è in corso – ma si tratta di un lavoro già in larga parte disponibile alla consultazione da parte della comunità scientifica – la costruzione di tre strumenti digitali, gratuitamente e liberamente accessibili in rete:[1]

- la *Bibliografia filologica – DiVo DB*: un repertorio analitico di schede sulla tradizione dei testi latini e volgari oggetto dello studio, consultabile all’indirizzo http://tlion.sns.it/divo/;
- il *corpus DiVo*: un corpus interrogabile per forme, lemmi e iperlemmi che raccoglie esaustivamente i volgarizzamenti disponibili in edizione affidabile, con l’associazione paragrafo per paragrafo del

[1] *DiVo DB* rientra nella rete *TLIon – Tradizione della letteratura online* (*TLIon DB*), diretta da Claudio Ciociola, per cui cfr. *infra*. I *corpora* sono gestiti dal software lessicografico GATTO 3.3, ideato e sviluppato da Domenico Iorio-Fili presso l’Opera del Vocabolario Italiano, nella versione *Gattoweb*, realizzata dallo stesso Domenico Iorio-Fili con la collaborazione di Andrea Boccellari – si tratta dello stesso software che gestisce il corpus di riferimento dell’italiano antico, il *corpus OVI dell’Italiano antico* (in abbreviazione *corpus OVI*).

testo latino di partenza e con note filologiche sulla tradizione latina e volgare, consultabile all'indirizzo http://divoweb.ovi.cnr.it/;

- il *corpus CLaVo*: un corpus che permette ricerche a partire da latino per forme e lemmi, con l'associazione paragrafo per paragrafo del testo volgare d'arrivo, consultabile all'indirizzo http://clavoweb.ovi.cnr.it/.

Le analisi lessicali condurranno alla redazione di voci per il *TLIO* (*Tesoro della lingua italiana delle origini*) e a studi specifici di carattere onomasiologico e semasiologico sul lessico di traduzione.

I tre strumenti, nell'integrazione reciproca tra dati e meta-dati da un lato, tra linguistica, filologia e informatica dall'altro, puntano a porsi come un punto di riferimento nella ricerca lessicale sui volgarizzamenti, e più in generale sui testi dell'italiano antico, considerato il ruolo, da un punto di vista quantitativo e qualitativo, dei testi di traduzione nella documentazione italiano antica (cfr. *infra*). Nella lessicografia storica dell'italiano, questo ruolo era riconosciuto sin dal Vocabolario della Crusca (1612) e prima ancora tale consapevolezza era ben presente alle riflessioni che ne ispirarono l'elaborazione, nella fattispecie di Salviati e di Borghini: dalla valutazione dell'errore di traduzione in rapporto all'uso linguistico, da considerare positivamente come un'attestazione linguistica fede degna quando restituisce un uso possibile nell'architettura della lingua, a prescindere dal fatto che essa possa provenire da un'incomprensione totale o parziale del dettato del testo di partenza, all'effetto di trascinamento del latino su numerosi lemmi, perlopiù crudi latinismi, che trovano nei volgarizzamenti un'attestazione prevalente o addirittura esclusiva, con, sullo sfondo, il problema della "naturalità", tema centrale per le teorie linguistiche del XVI secolo.[2]

Passeremo brevemente in rassegna le caratteristiche principali di questi tre strumenti, per poi portare alcuni esempi delle analisi lessicali, mostrando come con la loro interazione si possa ottenere un raffinamento delle nostre conoscenze sul lessico dell'italiano antico.

Speciale attenzione sarà dedicata al rapporto tra linguistica e filologia, soprattutto in riferimento al problema dell'estrazione dei dati lessicografici da *corpora* informatici (e alla diversa e relativa affidabi-

[2] Cfr. Guadagnini (2013: 62-65), che osserva però come la sensibilità per il problema della testimonianza dei volgarizzamenti come fonte lessicografica sia andata gradualmente perdendosi nel XVIII e XIX secolo. Riprende questo filo interrotto il *TLIO* con la *Bibliografia dei volgarizzamenti*, un repertorio sintetico di schede dedicato a tutti volgarizzamenti presenti nel *corpus OVI dell'italiano antico*, finalizzato ad agevolare il redattore nel reperimento e nel confronto con il testo originale di partenza, in modo da fornire un'interpretazione corretta di un contesto o evidenziare un'accezione particolare di un lemma (cfr. Artale 2003: 299 e Beltrami 2010: 246-247). Naturalmente per "testo originale" va intesa un'approssimazione ricostruibile a partire dalle edizioni moderne di riferimento, dallo spoglio degli apparati delle stesse o dalle ipotesi formulabili integrando i dati disponibili – sono invece rari i casi in cui abbiamo un testo specificamente noto, e anche in questi casi, in realtà, occorrerebbe distinguere tra la lettura "reale" di una determinata lezione nel testo di partenza e la lettura mentale da parte del volgarizzatore.

lità delle edizioni su cui si fonda la costituzione di un corpus).[3] Si aggiunga che, tra tutte le discipline linguistiche, la lessicografia sconta una certa inerzia, legata in particolare all'effetto di trascinamento che caratterizza la tradizione lessicografica – a qualsiasi livello, ogni dizionario è inevitabilmente in dialogo con i dizionari che lo hanno preceduto – e alla scarsa o modesta affidabilità filologica dei dizionari, un problema ancora più grave e delicato quando si tratta di dizionari storici, che a loro volta sono oggetti storici.[4]

La prospettiva che ispira il progetto *DiVo* è così espressa da Elisa Guadagnini:

> la circolarità ineliminabile fra qualità delle edizioni e qualità (vale a dire affidabilità) dei *corpora* testuali e degli studi che ne derivano è una tara che inficia la scientificità dei risultati soltanto per chi abbia la feticistica presunzione di estrapolare – dai *corpora* e dalle edizioni – dei dati di verità: la consapevolezza che qualunque testo restituito da qualunque tipo di edizione è di per sé un testo "ricostruito" consente invece, a nostro avviso, di preservare il valore, ma ancora prima il senso e la legittimità, di strumenti o di analisi che coprano vasti insiemi di materiali in una prospettiva ampiamente comparatistica, al netto del margine di oscillazione, di variabilità, di potenziale cambiamento nella lezione o nell'interpretazione, che è sempre postulabile per ogni dato testuale. (Burgassi e Guadagnini 2014, i.c.s.)

Corollario di questa impostazione è che il *corpus DiVo* si fondi su edizioni di testi con differenti gradi di affidabilità, posto che sono state escluse le edizioni inaffidabili. L'inclusione più problematica ha riguardato soprattutto le edizioni sette-ottocentesche che tengono ancora il campo nonostante il rinnovamento degli studi e dei metodi. La loro esclusione si sarebbe rivelata una soluzione facilmente percorribile, ma di fatto scarsamente funzionale alla necessità di uno studio ad ampio raggio del lessico dei volgarizzamenti: infatti, da un lato, in negativo, un corpus dei volgarizzamenti dei classici fondato in maniera esclusiva su edizioni critiche moderne avrebbe scontato una serie così ampia di lacune che il suo valore rappresentativo sarebbe stato quasi annullato; dall'altro lato, in positivo, è opportuna una parziale e ragionata rivalutazione di una parte di queste edizioni, le quali, fondate su un manoscritto unico, restituiscono di norma una lezione fede degna almeno sul piano della sostanza, ciò che più importa per chi sia interessato al lessico. Va da sé che, per esempio, un uso dell'intero

[3] Con focalizzazione sul rapporto tra filologia e storia della lingua (cui è legittimo senz'altro sostituire linguistica, in particolare con riferimento alle varietà linguistiche antiche), questo rapporto "è spesso così vincolante da configurare un circolo vizioso: non abbiamo una buona edizione perché ci mancano sufficienti conoscenze storicolinguistiche perché non disponiamo di edizioni affidabili dei testi donde dovremmo attingere" (Stussi 1993: 214). Da un'altra angolatura, centrata invece sulla linguistica, cfr. i principi che fondano la *Grammatica dell'italiano antico* (Salvi e Renzi 2010: 7-16): "il circolo filologia-linguistica, per cui ognuno dei due punti di vista presuppone in realà l'altro [...] si può alla volte spezzare emettendo ipotesi e provando ad applicarle".

[4] Cfr. Beltrami (2011) e Picchiorri (2013; 2014). Nella prospettiva del *TLIO*, cfr. la ricostruzione storica del dibattito sull'avvio dei lavori per il vocabolario in Vaccaro (2013). La condizione dell'esistenza stessa del *TLIO* è la scommessa di "fare" un dizionario storico fondato su spogli di prima mano, a loro volta fondati sulle edizioni esistenti disponibili, spezzando il circolo vizioso che altrimenti avrebbe costretto a una dilazione continua dell'elaborazione del vocabolario a favore della preparazione dei testi (Beltrami 2011: 342).

corpus DiVo per ricerche che investano la forma dei testi (i livelli della fonologia, della morfologia o alcuni aspetti della morfosintassi) è a carico della (ir)responsabilità di chi consulta il corpus, mentre è una responsabilità di chi lo costruisce mettere a disposizione tutti i dati per una corretta valutazione dell'edizione presente nel corpus, chiarendo le metodologie seguite dall'editore e verificandone l'affidabilità. Viceversa un utilizzo parziale anche per la forma è possibile, ma solo per sottocorpora, selezionando i testi editi secondo criteri rigorosi anche sul piano formale o particolarmente interessanti su questo fronte, com'è il caso, per esempio, dei volgarizzamenti testimoniati da un autografo.[5]
Insomma alle spalle e davanti al lessicografo, come a chiunque interroghi un corpus informatizzato di testi, sta (o dovrebbe stare) sempre "una continua valutazione critica dei dati" (Beltrami 2011: 348).

2 La Bibliografia filologica del DiVo - DiVo DB

Premessa e allo stesso tempo risultato della valutazione delle edizioni da inserire nel corpus dei volgarizzamenti dei classici è la *Bibliografia filologica* del *DiVo* (*DiVo DB*), repertorio di schede con la presentazione dei dati essenziali sull'opera: cenni biografici sull'autore del volgarizzamento, datazione, identificazione della coloritura linguistica, indicazione della tipologia testuale e del genere, catalogazione della tradizione diretta mediante il censimento dei testimoni manoscritti e delle stampe antiche, sintesi sulla storia della tradizione, identificazione dell'edizione di riferimento e panorama bibliografico articolato per punti[6]. Le schede sui testi di partenza (perlopiù opere latine, per cui cfr. Zago 2012) sono invece sintetiche, fornendo dati sull'autore, sulla datazione, sulla tipologia testuale, sul genere e sull'edizione di riferimento sulla base di una valutazione dello stato degli studi.
Come ha scritto Giulio Vaccaro (Guadagnini e Vaccaro 2014: 127), il lavoro alla base di *DiVo DB* "è stato foriero di robuste novità nel campo delle acquisizioni testuali", e di conseguenza anche per i dati lessicografici in uscita.
Si pensi per esempio al campo delle retrodatazioni: tradizionalmente il volgarizzamento delle *Collazioni dei santi Padri* di Giovanni Cassiano era datato in modo generico al XIV secolo (nel *GDLI* e come fuori corpus nel *TLIO*) sulla base dell'edizione ottocentesca fondata sul manoscritto 1637 della Biblioteca statale di Lucca, datato al 1442 (Bini 1854). Ma l'indagine sulla tradizione manoscritta di questo testo, peraltro non unitario, ma suddivisibile in due distinti volgarizzamenti per le collazioni I-X e XI-XXIV, ha portato al rinvenimento di un testimone senese del volgarizzamento A ,databile su base paleografica alla fine del XIII secolo (Siena, Biblioteca degli Intronati, I V 8), che produrrà sicuramente numero-

[5] Sui criteri che hanno guidato l'inclusione delle edizioni nel corpus, anche in rapporto al *corpus OVI dell'Italiano antico*, che ha criteri leggermente diversi, cfr. Dotto (2013: 75-78) e Guadagnini e Vaccaro (2014: 119-127). Nel *corpus DiVo*, per esempio, si leggono (o si leggeranno) le edizioni dei volgarizzamenti delle *Heroides* di Filippo Ceffi secondo il riconosciuto autografo Vaticano Palatino Latino 1644 (Zaggia 2009) o del *De brevitate vitae* di Seneca secondo la copia di mano di Andrea Lancia (Frullani 5 della Biblioteca Moreniana di Firenze), da riconoscere verosimilmente come il volgarizzatore (cfr. per ora De Robertis e Vaccaro 2013).

[6] Il censimento dei volgarizzamenti dei classici è passato attraverso una versione a stampa: cfr. Artale, Guadagnini, Vaccaro 2010.

se retrodatazioni: per esempio la prima attestazione della voce *inquietudine* del *TLIO*, la cui più antica attestazione compare nel volgarizzamento fiorentino delle *Pistole* di Seneca, databili *ante* 1325, potrà essere retrodata proprio grazie a questo ritrovamento, sfruttando le potenzialiatà di uno strumento come il *TLIO*, pubblicato direttamente in rete e pertanto facilmente aggiornabile.[7]

3 Il corpus DiVo (Corpus del Dizionario dei volgarizzamenti)

Il *corpus DiVo* raccoglie le edizioni affidabili dei volgarizzamenti dei testi classici e tardo-antichi in una qualsiasi varietà italoromanza con limite posto al XIV secolo. Esiste un vincolo all'esaustività per i volgarizzamenti delle opere classiche. Sono inclusi nel corpus anche i testi che non sono volgarizzamenti diretti da latino, ma hanno un intermediario romanzo (di norma il francese), per esempio le *Pistole* di Seneca, che dipendono da un volgarizzamento francese, o in un'altra varietà italoromanza (di norma il toscano), per esempio l'*Istoria d'Eneas* siciliana di Angilu di Capua, che dipende dalla compilazione fiorentina tradizionalmente attribuita ad Andrea Lancia. Inoltre, in casi motivabili con il valore storico-culturale o linguistico del testo, oltre ai volgarizzamenti veri e propri, con la resa generalmente puntuale del dettato latino, sono consultabili nel corpus anche alcune compilazioni che corrispondono al montaggio di volgarizzamenti distinti: è il caso dei *Fatti de' Romani*, vasta compilazione di storia romana di origine antico francese, che riusa Sallustio, Cesare, Lucano e Svetonio e che è circolata in Italia secondo differenti redazioni.

Attualmente, col rinnovo di marzo 2014, il corpus comprende 150 testi volgari, per complessive 5.941.061 occorrenze e 169.845 forme grafiche distinte. Per dare un'idea della sua entità e della ragione per cui esso potrà integrare utilmente il corpus di riferimento dell'italiano antico (*corpus OVI*), ricordo che quest'ultimo comprende 2316 testi per complessive 23.157.266 occorrenze e 467.098 forme grafiche distinte.

I punti di forza del corpus sono:

- l'associazione paragrafo per paragrafo del testo latino di partenza, in modo da rendere immediatamente conto del rapporto tra testo tradotto e corrispondente traduzione;
- un sistema di annotazione, che contempla la lemmatizzazione e l'iperlemmatizzazione sugli àmbiti semantici sensibili per lo studio del lessico dei volgarizzamenti.[8]

[7] Per questa ragione nel *corpus DiVo* sono e saranno consultabili un'edizione a uso interno del volgarizzamento A sulla base del manoscritto senese, a cura del progetto *DiVo*, e l'edizione Bini che si fonda sul manoscritto di Lucca, che accorpa entrambi i volgarizzamenti. Si rinvia alla relativa scheda in *DiVo DB*.

[8] Sui criteri dell'associazione del latino al testo volgare, cfr. Burgassi 2013: l'associazione dei volgarizzamenti dei classici è completa, per cui il corpus ad oggi presenta 79 testi con latino associato, che coprono 3.172.419 occorrenze. La lemmatizzazione e l'iperlemmatizzazione sono invece ancora in costruzione (per un primo abbozzo dei principi e delle soluzioni tecniche che si seguiranno, cfr. Dotto 2012).

4 Il corpus CLaVo (Corpus dei classici latini volgarizzati)

Il *corpus CLaVo* è il corpus "gemello" del *corpus DiVo* in quanto raccoglie tutte le opere latine classiche contenute nel *corpus DiVo*, con l'associazione paragrafo per paragrafo di ciascun volgarizzamento. Al momento contiene 26 opere latine, associate a 45 volgarizzamenti, per complessive 913.656 occorrenze di 78.587 forme grafiche distinte, ma non è ancora completo; è consultabile per forme e in parte anche per "lemmi muti", una speciale lemmatizzazione, curata da Anna Zago, che fornisce una prima griglia di coppie forma-lemma per agevolarne l'interrogazione. A regime esso conterrà più di 80 testi, per circa 1.300.000 occorrenze e oltre 120.000 forme grafiche distinte; grazie al lemmatizzatore semi-automatico di Gatto 4.0 potrà contare su una lemmatizzazione vera e propria largamente esaustiva.[9]

La funzione primaria del *corpus CLaVo* è la possibilità di recuperare agevolmente attraverso lo scaricamento dei contesti tutte le rese traduttive dello stesso lemma latino a partire dal testo latino di partenza, attraverso i meccanismi del prestito, del calco o della riformulazione volgare (un equivalente o una perifrasi).[10] Così per es., come vedremo in parte più avanti, è possibile recuperare i traducenti del lat. *lictor*: *littore*, *berroviere*, *giustiziere*, *masnadiere*, *messo*, *sergente*, ecc. Uno strumento simile ha fortissime potenzialità, per sondare sia il lessico tecnico e "storico" (cfr. *infra*), sia il lessico generico. In questo modo infatti è possibile tracciare relazioni onomasiologiche all'interno del volgare, per inventariare le relazioni sinonimiche e soprattutto per apprezzare le frequenze d'uso, i valori connotativi e denotativi dei diversi lemmi in sincronia e i cambiamenti che i medesimi lemmi hanno subito in diacronia, nella "breve" diacronia del XIII e XIV secolo con interesse centrato sull'italiano antico, come nella "lunga diacronia" per una migliore intelligenza di determinati lemmi, come nel caso dei "latinismi latenti" (cfr. Burgassi e Guadagnini 2014).

Da notare che l'ordinamento cronologico dei testi non si riferisce a quello dei testi latini, ma a quello dei testi volgari: questa soluzione intende agevolare lo studio di come siano cambiate le modalità traduttive nel XIII e XIV secolo, per verificare analiticamente l'ipotesi già di Cesare Segre (1953) dell'esistenza di una prima fase, in cui i volgarizzatori tendono a ricorrere a equivalenti volgari, e di una seconda, per certi aspetti preumanistica, benché l'attività del volgarizzare sia operazione schiettamente anti-umanistica, in cui i volgarizzatori privilegiano il prestito dal latino.

5 Esempi di analisi lessicali

Uno dei problemi più cospicui dei volgarizzatori impegnati nella traduzione dei classici era quello di rendere il lessico materiale o "storico": si tratta di un lessico non-marcato in latino perché non indivi-

[9] Su Gatto 4.0, destinato a sostituire Gatto 3.3, e in particolare sul lemmatizzatore semi-automatico, cfr. Iorio-Fili 2012.

[10] Un'altra potenzialità del *corpus CLaVo*, che non si può discutere qui, è il suo riuso indipendente dall'analisi lessicale dei volgarizzamenti in quanto banca dati di testi classici in sé e per sé, interrogabile grazie agli strumenti del software lessicografico Gatto 3.3.

dua settori disciplinari specifici, ma risulta "speciale" solo in una prospettiva storica perché rinvia a referenti che sono tipici della società e della cultura antica e sono scomparsi in quella medievale. Rientrano in questa categoria i lemmi che si riferiscono a oggetti di uso quotidiano o alla misurazione del tempo, sostituita in séguito dalla cronologia cristiana, o ancora all'esistenza di popoli scomparsi e territori ridisegnati rispetto alla prospettiva e alle conoscenze geografiche dei volgarizzatori medievali. È il caso del lessico della cariche e degli uffici, che conteneva ampie di zone di discontinuità tra il mondo antico e quello medievale. In questi casi il volgarizzatore poteva oscillare tendenzialmente tra una traduzione orientata sulla lingua di partenza, il latino, ricorrendo al prestito diretto o una traduzione orientata sulla lingua d'arrivo, il volgare, cercando nel repertorio lessicale della propria lingua un equivalente che desse conto del significato del lemma latino. Possibilità collaterale, in realtà caratteristica di molti volgarizzamenti dei classici, specialmente nella prima metà del XIV secolo a Firenze, era quella di accompagnare il testo con apparati paratestuali in forma di chiose, marginali o interlineari, o di glossari, all'inizio o alla fine del testo.[11] È quanto avviene per il lat. *lictor* "ufficiale che accompagnava vari magistrati romani recando con sé fasci di verghe con una scure in mezzo":[12]

(1) gli maggiori Romani con gradissima diligenza ritennero questa usanza: che alcuno non s'interponesse tra 'l consolo e ' pressimani *lictori*, tutto ch'egl'andassero insieme per cagione d'oficio. (*Valerio Massimo* (Vb), a. 1326 [fior.]) / cfr. Val. Max. II, 2, 24: "Maxima autem diligentia maiores hunc morem retinuerunt, ne quis se inter consulem et proximum *lictorem*, quamuis officii causa una progrederetur, interponeret"

– nel volgarizzamento parziale di Valerio Massimo, Vb, troviato un prestito diretto, ma nello stesso testo nelle chiose troviamo l'equivalenza tra "ufficiale" e "littore" (2) e in un glossario, peraltro presente in uno dei due testimoni che tramanda il volgarizzamento Vb, tra "sergente" e "littore" (3):

(2) xii imperialissimi onori erano li xii *officiali* de' consoli, li quali si chiamavano '*lictori*' e portavano le 'nsegne de' consoli. (*Chiose a Valerio Massimo* (Vb), 1326 [fior.])

(3) Ciascuno consolo aveva xii *sergenti* li quali erano chiamati *littori*. (*Gloss. degli uffici romani* (red. Marc.), XIV pm. [tosc.])

L'equivalenza tra il termine generico "sergente" e "littore" è documentata anche nel volgarizzamento della *Deca quarta*:[13]

(4) sopra tutti con portamento eccelso rendere superbe leggi intorniato di *sergenti* chiamati *littori*; i quali sempre e sudditi stanno con le verghe del ferro al dosso, e con le securi sopra le teste; e questo avvi-

[11] Le chiose, come i glossari, servivano a colmare la "distanza epistemica" tra la cultura antica e quella medievale: "Il commento è il termometro delle difficoltà della comunicazione. Il caso più ovvio è quello della distanza cronologica e geografica tra emittente e ricevente: sono i testi antichi o quelli in altre lingue ad essere fregiati più spesso di commento. Si potrebbe parlare, meglio, di distanza epistemica: si terrebbe così conto, oltre che della distanza cronogeografica, anche di quella culturale" (Segre 1992: 4).

[12] I testi citati e le relative abbreviazioni bibliografiche corrispondono a quelle del *corpus DiVo*: per accedere alla bibliografia: *DiVoWeb* (o *CLaVoWeb*) > *Altre funzioni* > *Accesso ai dati bibliografici* (con i link a *DiVo DB* per ulteriori approfondimenti).

[13] Se ne hanno riscontri anche in àmbito francese antico: cfr. la *Base de civilisation romaine (XII^e-XV^e s.),* s.v. *lictor*.

ene quanti anni ora l'uno ed ora l'altro signore rinnovando sortiscono. (*Deca quarta*, a. 1346 [fior.]) / cfr. Liv., XXXI, 39, 9: "Praetor Romanus conventus agit: eo imperio evocati conveniunt, excelso in suggestu superba iura reddentem, stipatum lictoribus vident, virgae tergo, secures cervicibus imminent"

Un caso interessante è discusso da Massimo Zaggia (1991: 611-613) in una sua recensione al volgarizzamento anonimo della prima Epistola di Cicerone al fratello Quinto (Piva 1989):

(5) Per le quali cose Gneo Ottavio poco tempo fa fu reputato da tutti molto soave e benigno, nel cui reggimento il primo *littore o berroviere* tacette senza vietare la venuta <e non bisognava dire 'l tale vuole parlare>, ciascuno parlò quante volte gli piacque e quanto lungamente egli volle. (*Ep. a Quinto*, XIV sm. [tosc.]) / cfr. Cic. *Q. fr.*, 21: "apud quem primus *lictor* quievit, tacuit accensus"

Il manoscritto base seguito dall'editrice legge "littore o sergiente", mentre il resto della tradizione ha "littore o berroviere". In questo caso Maria Antonia Piva opta per l'espunzione di "sergente" / "berroviere" ritenendo che si tratti di una glossa che ha così costituito una dittologia sinonimica, secondo un processo normale nella tradizione dei volgarizzamenti.[14] All'operazione della Piva, Zaggia (1991: 612) obietta che "se l'eliminazione delle esplicazioni sinonimiche sembra legittima quando queste risultano trasmesse da un solo ramo della tradizione [...], non pare autorizzata quando tutti i testimoni concordano nel trasmettere una dittologia". A ulteriore sostegno della propria obiezione, Zaggia riporta il seguente esempio, in cui tutta la tradizione concorda nella dittologia "berroviere e littore":

(6) Sia ogni tuo *berroviere e littore* dimostratore non della sua benignità e dolcezza, anzi della tua, e quelli frusti e quelle scure o mannaie che portano più dimostrino segno della dignità dell'ufficio tuo che della signoria e forza. (*Ep. a Quinto*, XIV sm. [tosc.]) / cfr. Cic. *Q. fr.*, 13: "sit *lictor* non suae sed tuae lenitatis apparitor"

Nel *corpus DiVo*, si è scelto di seguire l'ipotesi formulata da Zaggia, che preferisce la dittologia "littore o berroviere" a causa del riscontro di (6), inserendo una nota che dà conto della divergenza rispetto all'edizione di riferimento e si giustifica l'intervento apportato.

Un altro caso che chiama in causa il delicato rapporto tra edizioni e dati lessicografici ricavabili da esse. A fronte del lat. *praesultor*, che evidentemente a causa della propria rarità non doveva risultare perspicuo, le tre redazioni del volgarizzamento toscano di Valerio Massimo (in bibliografia con le sigle Va V1 V2) presentano i seguenti esiti:[15]

(7) Iove comandò a un popolare latino in sogno che dicesse al consolo che no· gli piacea ne' prossimi giochi di Circe quello *prosentuoso vendicatore*... (Valerio Massimo (red. Va), a. 1336 [tosc.]) / cfr. Val. Max., I, 7, 4: "<T>. Latinio homini ex plebe Iuppiter in quiete praecepit ut consulibus diceret sibi *praesultorem* ludis circensibus proximis non placuisse"

[14] Se "sergente" è termine generico, non così per "berroviere", che in italiano antico era ben attestato nel significato di "funzionario con mansioni esecutive al servizio di un ufficiale pubblico (capitano, podestà, priore ecc.) o di un signore" (cfr. *TLIO* s.v.).

[15] Nel *corpus DiVo* le redazioni del volgarizzamento di Valerio Massimo si leggono per la prima volta integralmente grazie alle edizioni di lavoro approntate da Vanna Lippi Bigazzi, che ringraziamo di cuore qui. Cfr. la scheda in *DiVo DB*.

Da dove nasce la lezione "prosentuoso vendicatore"? Da un'erronea segmentazione, per cui il volgarizzatore deve aver letto "praes ultorem" in luogo di "praesultorem". La lezione non è però confermata nella redazione V1, almeno secondo l'edizione DeVisiani (1867-1868) (8), mentre l'edizione di Vanna Lippi Bigazzi conferma per V1 la lezione di Va (9):

(8) Iove comandò a uno latino del popolo in sogno, che dicesse al consolo, che non li piacea nelli prossimi giuochi circensi quello *presultore.* (Valerio Massimo (red. V1, ed. De Visiani), a. 1336 [fior.])

(9) Iove comandoe a uno latino del popolo in sogno che dicesse al consolo che no· lli piacea <vedere> ne li prossimi giuochi di Circe quello *presumptuoso vendicatore*... (Valerio Massimo (red. V1, ed. Lippi Bigazzi), a. 1336 [fior.])

Spulciando l'apparato dell'edizione De Visiani (1867-1868: 80), si apprende che il suo manoscritto base ("St. e Cod.") legge in realtà "presentuoso vendicato" e solo un manoscritto, il Palatino 27 della Biblioteca Palatina di Parma, legge il crudo latinismo, non altrimenti attestato, "presultore".

Se passiamo all'ultimo anello delle redazioni del Valerio Massimo, V2, troviamo un'altra lezione, frutto probabilmente di un emendamento come nel caso di "presultore":

(10) Iove in sogno comandò ad uno uomo latino de la minuta gente che dicesse al consolo che a lui non era piaciuto quello *antisaltatore* ne li prossimi giuochi Circesi... (Valerio Massimo (red. V2), c. 1346 [tosc.])

Da un punto di vista lessicografico, importa constatare che "presultore" andrà scalato cronologicamente, visto che la lezione genuina di V1 sarà certo l'errore di traduzione "presuntuoso vendicatore" per trascinamento da Va. I due emendamenti restituiti dalla tradizione corrispondono inoltre a due distinte tipologie di resa traduttiva: la prima per prestito, la seconda per calco.

Va notato infine che nel contesto dello stesso esempio il volgarizzatore della *Deca prima* di Tito Livio, Filippo da Santa Croce ricorre a una complessa perifrasi a dimostrazione della delicatezza della resa di *praesultor* o dell'affine *praesultator*:

(11) e fugli avviso che Giove gli dicesse, che *quegli che la prima danza aveva alla festa menata*, gli dispiacque... (Filippo da Santa Croce, *Deca prima*, 1323 [tosc.]) / cfr. Liv., II, 36, 2: "visus Iuppiter dicere sibi ludis praesultatorem displicuisse"

Gli esempi portati vorrebbero dimostrare l'assunto iniziale: una "continua valutazione critica dei dati" è condizione necessaria per l'integrazione di linguistica e filologia.

6 Riferimenti bibliografici

Artale, E. (2003). I volgarizzamenti del *corpus TLIO*. In *Bollettino dell'Opera del Vocabolario Italiano*, 8, pp. 299-377.

Base de civilisation romaine (XII^e-XV^e s.), ed. F. Duval, CNRTL CNRS-ATILF. Accessed at: http://www.cnrtl.fr/lexiques/civirom/ [02/02/2014].

Artale, E., Guadagnini, E., Vaccaro, G. (2010). Per una bibliografia dei volgarizzamenti dei classici (il *Corpus DiVo*). In *Bollettino dell'Opera del Vocabolario Italiano*, 15, pp. 309-366.

Beltrami, P.G. (2010). Lessicografia e filologia in un dizionario storico dell'italiano. In C. Ciociola (ed.), Storia della Lingua Italiana e Filologia. Atti del VII Convegno ASLI (Pisa-Firenze, 18-20 dicembre 2008). Firenze: Franco Cesati, pp. 235-248.

Beltrami, P.G. (2011). Il mito dell'edizione per lessicografi e il *Tesoro della Lingua Italiana delle Origini*. In A. Overbeck, W. Schweickard, H. Völker (eds.), Lexicon, Varietät, Philologie. Romanistische Studien Günter Holtus zum 65. Geburtstag. Berlin: De Gruyter, 2011, pp. 341-349.

*Bibliografia dei volgarizzamenti [*del corpus TLIO*]*, ed. E. Artale. Accessed at: http://tlio.ovi.cnr.it/BibVolg/ [02/02/2014].

Bini, T. (ed.). 1854. Volgarizzamento delle Collazioni dei SS. Padri del venerabile Giovanni Cassiano. Lucca: Giusti.

Burgassi, C. (2013). Notizie dal *DiVo*. Teoria e pratica dell'associazione latino-volgare, In P. Larson, P. Squillacioti, G. Vaccaro (eds.), «Diverse voci fanno dolci note». L'Opera del Vocabolario Italiano per Pietro G. Beltrami. Alessandria: Edizioni dell'Orso, pp. 85-96.

Burgassi, C. e Guadagnini, E. (2014). Prima dell'«indole». Latinismi latenti dell'italiano. In *Studi di lessicografia italiana*, XXXI, i.c.s.

Corpus CLaVo. Corpus dei classici latini volgarizzati, eds. C. Burgassi, D. Dotto, E. Guadagnini, G. Vaccaro. Accessed at: http://clavoweb.ovi.cnr.it/ [02/02/2014].

Corpus DiVo. Corpus del Dizionario dei Volgarizzamenti, eds. C. Burgassi, D. Dotto, E. Guadagnini, G. Vaccaro. Accessed at: http://divoweb.ovi.cnr.it/ [02/02/2014].

Corpus OVI dell'Italiano antico, ed. E. Artale e P. Larson. Accessed at: http://gattoweb.ovi.cnr.it/ [02/02/2014].

De Robertis, T. e Vaccaro, G. (2013). Il *Libro di Seneca della brevitade della vita humana* in un autografo di Andrea Lancia. In *Studi di filologia italiana*, 71, i.c.s.

De Visiani, R. (ed.). 1867-1868. Valerio Massimo, De' fatti e detti degni di memoria della città di Roma e delle stranie genti. Bologna: Romagnoli.

DiVo DB. DiVo – Bibliografia filologica, eds. E. Guadagnini e G. Vaccaro. Accessed at: http://tlion.sns.it/divo/ [02/02/2014].

Dotto, D. (2012). Note per la lemmatizzazione del corpus DiVo. In *Bollettino dell'Opera del Vocabolario Italiano*, 17, pp. 339-366.

Dotto, D. (2013). Notizie dal *DiVo*. Un primo bilancio sulla costituzione del corpus. In P. Larson, P. Squillacioti, G. Vaccaro (eds.), «Diverse voci fanno dolci note». L'Opera del Vocabolario Italiano per Pietro G. Beltrami. Alessandria: Edizioni dell'Orso, pp. 71-83.

GDLI. (1961-2002). *Grande dizionario della lingua italiana*, ed. S. Battaglia, [poi G. Bàrberi Squarotti]. Torino: UTET.

Guadagnini, E. (2013). Notizie dal *DiVo*. Parole tradotte e lessicografia dell'italiano. In P. Larson, P. Squillacioti, G. Vaccaro (eds.), «Diverse voci fanno dolci note». L'Opera del Vocabolario Italiano per Pietro G. Beltrami. Alessandria: Ed. dell'Orso, pp. 59-70.

Guadagnini, E. e Vaccaro, G. (2014). Un contributo allo studio del "Volgarizzare e tradurre": il progetto *DiVo*. In Lingue, testi, culture. L'eredità di Folena, vent'anni dopo. Atti del XL Convegno Interuniversitario di Bressanone (12-15 luglio 2012). Padova: Esedra, pp. 113-127.

Iorio-Fili, D. (2012). Il lemmatizzatore semi-automatico di Gatto4. In Dizionari e ricerca filologica. Atti della Giornata di Studi in memoria di Valentina Pollidori (Firenze, Villa Reale di Castello, 26 ottobre 2010). Alessandria: Ed. Dell'Orso, pp. 41-56.

Picchiorri, E. (2013). Sulla genesi di un errore nel Battaglia. In *Studi linguistici italiani*, 39(1), pp. 134-136.

Picchiorri, E. (2014). Problemi filologici nei dizionari storici italiani dal *GDLI* al *TLIO*. In J.M. Brincat, R. Coluccia, F. Möhren (eds.), Actes du XXVII^{e} Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 5: Lexicologie, phraséologie, lexicographie. Nancy: ATILF, i.c.s.

Piva, M.A. (ed.). 1989. Anonimo trecentesco. *Volgarizzamento della prima Epistola di Cicerone al fratello Quinto*. Bologna: Commissione per i Testi di lingua.

Salvi, G. e Renzi, L. (eds). (2010). *Grammatica dell'italiano antico*. Bologna: il Mulino.

Segre, C. (ed.). 1953. *Volgarizzamenti del Due e Trecento*. Torino: UTET.

Segre, C. 1992. Per una definizione del commento ai testi. In O. Besomi e C. Caruso (eds.), Il commento ai testi. Atti del Seminario di Ascona (2-9 ottobre 1989). Basel-Boston-Berlin: Birkhäuser, pp. 3-14.

Stussi, A. (1993) *Lingua, dialetto e letteratura*. Torino: Einaudi.

TLIO. Tesoro della lingua italiana delle origini, ed. P. Squillacioti. Accessed at: http://tlio.ovi.cnr.it/TLIO/ [02/02/2014].

TLIon DB. Tradizione della letteratura italiana online, ed. C. Ciociola. Accessed at: http://www.tlion.it/ [02/02/2014].

Vaccaro, G. (2013). Veniamo da molto lontano e andiamo molto lontano. Documenti per la storia dell'Opera del Vocabolario Italiano dalle origini al 1992. In *Bollettino dell'Opera del Vocabolario Italiano*, 18, ic.s.

Zaggia, M. 1991. Rec. a Piva 1989. In Rivista di letteratura italiana, IX.3, pp. 611-616.

Zaggia, M. (ed.). 2009. Ovidio, «Heroides». Volgarizzamento fiorentino trecentesco di Filippo Ceffi. I. Introduzione, testo secondo l'autografo e glossario, Firenze, SISMEL-Ed. del Galluzzo.

Zago, A. (2012). La bibliografia dei testi latini (e greci) inclusi nel *corpus DiVo*. In *Bollettino dell'Opera del Vocabolario Italiano*, 17, pp. 367-391.

Acknowledgements

Questo contributo rientra nel progetto *DiVo* (*Dizionario dei Volgarizzamenti*), ospitato dall'Opera del Vocabolario Italiano - CNR e dalla Scuola Normale Superiore di Pisa, finanziato dal MIUR all'interno del programma FIRB – Futuro in Ricerca 2010.

Informatiser le Französisches etymologisches Wörterbuch: la nécessaire prise en compte de l'utilisateur

Pascale Renders, Esther Baiwir
F.R.S-FNRS/Université de Liège
pascale.renders@ulg.ac.be, ebaiwir@ulg.ac.be

Résumé

« Know your user » (Atkins & Rundell 2008 : 5). Ce conseil ne s'applique pas uniquement à la conception et la rédaction d'un nouveau dictionnaire. La transformation d'un dictionnaire imprimé en dictionnaire électronique peut également bénéficier d'une étude réévaluant les parcours de consultation suivis par les utilisateurs et les difficultés qu'ils rencontrent. Dans une étude préalable à l'informatisation du *Französisches Etymologisches Wörterbuch* de Walther von Wartburg, la prise en compte du point de vue des utilisateurs a mis en évidence deux facettes de l'ouvrage, vu tantôt comme un recueil de monographies, tantôt comme un thesaurus. Après un résumé de l'avancement du projet (qui est maintenant dans sa phase de production), cette communication expose la façon dont les deux visions ont influencé la modélisation informatique du discours lexicographique, permettant ainsi de résoudre une grande partie des problèmes rencontrés par les utilisateurs et ouvrant la voie à une mise à jour de l'ouvrage.

Keywords: FEW; digitalization; user perspective

1 Introduction

Il est aujourd'hui admis que le concepteur d'un dictionnaire doit d'abord définir précisément le public auquel il s'adresse:

> [...] the most important single piece of advice we can give to anyone embarking on a dictionary project is : know your user. [...] This doesn't imply a superficial concern with 'user-friendliness', but arises from our conviction that the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for. (Atkins & Rundell 2008 : 5)

Cette remarque s'applique-t-elle aussi pour un dictionnaire existant, déjà publié, qu'on voudrait transformer en dictionnaire électronique ? Si l'objectif de l'informatisation est d'augmenter les potentialités de consultation et de résoudre des problèmes d'utilisation, ne faut-il pas avoir une idée précise de l'identité des utilisateurs du dictionnaire, de ce qu'ils y cherchent, de la façon dont ils le consultent, des difficultés qu'ils rencontrent et des fonctionnalités qu'ils voudraient y trouver ? Plus généralement, l'utilisation effective de l'ouvrage diverge-t-elle de ce qui était prévu lors de la conception du projet initial et peut-elle être améliorée par le changement de medium ?

Ces questions ont été le point de départ d'une étude qui s'interrogeait sur les possibilités d'informatisation du *Französisches etymologisches Wörterbuch* de Walther von Wartburg (ou FEW), dictionnaire de référence en linguistique historique française et romane. Cette étude a été achevée en 2011 ; l'informatisation du FEW est en cours depuis octobre 2012. Après un résumé du projet et de son avancement, nous proposons d'exposer comment le point de vue de l'utilisateur a modifié l'analyse des structures du dictionnaire et comment ce changement de perspective a été pris en compte dans l'informatisation proprement dite.

2 Le projet d'informatisation du FEW

L'idée d'informatiser le FEW vient de ses utilisateurs. Le premier fut Wooldridge, déjà en 1990 (1990 : 239) puis en 1998 : « [l]a seule façon de mettre au jour ce qui concerne le français du XVIe siècle dans le FEW serait d'informatiser les 25 volumes… puis de les interroger à partir de repères comme « fr. », « mfr. », « 16^{e} s. », etc. » (Wooldridge 1998 : 211). Il suffit d'avoir consulté une fois l'ouvrage pour comprendre l'intérêt à la fois de celui-ci et de son informatisation. Le FEW, rédigé de 1922 à aujourd'hui (l'équipe de rédaction s'attelle depuis quelques années à la refonte du premier volume, cf. www.atilf.fr/few), est un dictionnaire de référence en linguistique romane, contenant le lexique de tous les parlers galloromans (français, francoprovençal, occitan, gascon et dialectes). Il est néanmoins sous-exploité, en raison de la complexité de ses structures (cf. Büchi & Chambon 1995). Son informatisation, ardemment souhaitée par la communauté scientifique, est censée à la fois résoudre les problèmes d'accessibilité de l'ouvrage et permettre sa mise à jour ; toutefois, un contenu dense et l'existence de nombreux caractères spéciaux non Unicode (plus d'une centaine) rendaient, jusqu'il y a peu, le projet utopique. Après quelques tentatives partielles, une étude de faisabilité fut entamée, financée par le laboratoire ATILF à Nancy (où est hébergée la rédaction du FEW, cf. www.atilf.fr), par la Fondation FEW (Suisse) et, en majeure partie, par l'Université de Liège en Belgique (bourse de doctorat).

Cette étude fut concluante : l'informatisation des 25 volumes du FEW était non seulement souhaitable, mais possible, sous la forme d'un balisage XML inséré a posteriori dans le discours lexicographique (Renders 2011). La réussite de l'entreprise nécessitait toutefois que soient pris en compte trois contraintes fortes, parmi lesquelles l'obligation de respecter les structures du dictionnaire, y compris dans leurs incohérences et leurs défauts, avec l'interdiction formelle de réécrire les articles pour normaliser le tout. La deuxième contrainte était la nécessité de pouvoir automatiser complètement l'insertion du balisage. La troisième contrainte, enfin, consistait à s'assurer que le résultat de l'informatisation répondrait effectivement aux attentes des utilisateurs, notamment en résolvant les divers problèmes d'accès auxquels ils se heurtaient. La prise en compte de ces trois contraintes mène en pratique à l'élaboration d'un compromis qui, seul, permet une conclusion positive.

Sur la base de la méthodologie mise au point dans cette étude, l'informatisation du FEW se déroule en trois phases :

(1) l'acquisition du texte des 25 volumes, accompagné d'un balisage typographique minimal ;
(2) le balisage XML complet des informations lexicographiques, de façon totalement automatisée, à l'aide d'un logiciel construit à cet effet ;
(3) la mise en ligne des articles balisés et leur exploitation via une interface de consultation.

Ces trois phases (développées plus longuement dans Renders à paraître) comportent quelques particularités par rapport à d'autres projets d'informatisation. Pour la première phase, une saisie manuelle du texte a été préférée à une numérisation, cette dernière s'étant avérée peu fructueuse en raison des nombreux caractères spéciaux non reconnus par les logiciels OCR. Une solution de double saisie, déjà utilisée pour d'autres dictionnaires (par exemple le *Deutsche Wörterbuch* des frères Grimm, cf. dwb.uni-trier.de/de/die-digitale-version), a été proposée et apportée par le Center for Digital Humanities de Trèves, qui possède une expertise reconnue dans ce domaine (cf. Kompetenzzentrum.uni-trier.de).

En ce qui concerne le balisage XML, il a pour particularité d'être pensé selon la perspective de l'utilisateur, contrairement par exemple au balisage du *Trésor de la Langue Française informatisé* qui fut automatisé selon les structures du dictionnaire uniquement (cf. Dendien & Pierrel 2003). Cette particularité le distingue également des dictionnaires conçus dès le départ dans une perspective électronique, ces derniers présentant généralement un balisage pensé pour les besoins de leur rédaction. L'automatisation du balisage et sa vérification s'effectuent à l'Université de Liège. Il est prévu que le balisage inséré soit plus tard converti au standard TEI (cf. www.tei-c.org). Le résultat du processus est la création d'articles au format XML qui pourront être exploité sous la forme d'une base de données.

La mise en ligne finale des articles informatisés nécessite quant à elle l'affichage de caractères spéciaux non standards (non Unicode) et, de ce fait, la création d'une police de caractères spécifique comprenant la totalité de ceux-ci. Seule la phase de mise en ligne est directement concernée par cette police : les phases précédentes nécessitent certes la reconnaissance de tous ces caractères spéciaux (sous la forme de codes ou d'entités XML), mais pas leur affichage. L'Atelier National de Recherche Typographique (www.anrt-nancy.fr) a proposé de créer cette police, tandis que l'ATILF se charge de développer l'interface d'interrogation.

Les trois phases sont successives, mais ne requièrent pas obligatoirement le traitement de la totalité du FEW à chaque étape. L'informatisation peut s'effectuer article par article. Actuellement (juillet 2014), trois des 25 volumes sont en cours de traitement. Ces volumes seront très certainement interrogeables en ligne avant que d'autres volumes ne soient saisis : la première phase est, en effet, la plus coûteuse et dépend donc fortement des financements apportés. En raison de cet obstacle financier à une informatisation rapide, il a parallèlement été décidé de mettre le FEW à la disposition de tous en mode image. Les 25 volumes sont accessibles depuis février 2014 à l'adresse https://apps.atilf.fr/lecteurFEW. Une possibilité d'interrogation minimale (par étymons et par lexèmes) de ces images est prévue, en attendant l'interface d'interrogation complète qui accompagnera la mise en ligne du FEW en mode texte.

3 L'utilisation du FEW

Les difficultés d'utilisation étant la raison principale du projet, il nous semblait évident que l'avis de l'utilisateur était à prendre en compte dès le départ, c'est-à-dire non seulement lors du développement de l'interface de consultation (phase 3), mais aussi dans la définition du balisage à insérer dans le dictionnaire (phases 1 et 2).

Afin de rencontrer au mieux les besoins des utilisateurs, il était d'abord nécessaire de les connaître, suivant le conseil donné par Atkins & Rundell (2008 : 5). Plusieurs questions se posaient, concernant d'abord l'utilité du dictionnaire, ensuite les parcours de consultation actuellement suivis dans la version imprimée du FEW et les problèmes rencontrés. Enfin, il s'agissait de s'interroger sur les parcours à mettre en place dans la version électronique pour répondre à ces problèmes. Les attentes des utilisateurs devaient, rappelons-le, dialoguer avec deux autres contraintes : l'obligation de respecter les structures du dictionnaire – c'est-à-dire le produit lexicographique tel qu'il a été pensé lors de sa rédaction et présenté dans la version imprimée – et la nécessité de pouvoir automatiser le balisage XML. Des demandes inconciliables avec ces contraintes seraient d'emblée soit rejetées, soit revues de façon à élaborer un compromis réaliste.

3.1 Le FEW, son utilité, ses utilisateurs

Si l'on en croit son titre, le FEW est un dictionnaire étymologique du français, ce qui pourrait faire croire qu'il est essentiellement utilisé pour connaître l'étymon des lexèmes de la langue française. En réalité, le titre de l'ouvrage est réducteur (cf. Büchi & Chambon 1995 : 947-948). D'une part, l'étymologie-histoire pratiquée par le FEW le mène à donner davantage d'informations que les autres dictionnaires étymologiques. Le FEW présente en effet une étymologie intégrante (cf. Malkiel 1976), c'est-à-dire que l'information étymologique représente le critère organisateur des données. La conséquence est une structure complexe à plusieurs niveaux (super-, macro-, micro- et infrastructure, cf. Büchi 1996 : 5-6). D'autre part, le domaine couvert par le FEW dépasse la langue française pour embrasser de façon presque exhaustive la totalité des lexèmes du domaine galloroman. Il s'agit donc d'un ouvrage de référence pour tous les parlers et dialectes concernés.

L'étendue du domaine linguistique pris en compte explique que chaque lexème soit associé dans le FEW à une étiquette géolinguistique, qui précise l'état de langue (ancien français, moyen français, français moderne ; ancien gascon etc.) ou le dialecte (lorrain, champennois, picard etc.) auquel il appartient. Des références bibliographiques complètent et précisent la chronologie suggérée par l'information géolinguistique. Le FEW sert donc prioritairement à étymologiser, localiser et dater un lexème dans un sous-domaine linguistique, même si d'autres utilisations sont possibles, par exemple pour connaître le sens d'un mot, sa graphie, sa forme phonique ou sa catégorie grammaticale.

La nature des données contenues dans le FEW, à savoir le lexique des parlers du domaine galloroman, explique qu'il soit utilisé en linguistique historique, dans les études concernant le lexique du français et des autres langues ou dialectes du domaine galloroman. Il est systématiquement utilisé par les étymologistes des autres langues romanes, ainsi que des langues non romanes. De manière générale, le FEW constitue une référence pour l'étude historique de toute langue qui a été en contact étroit avec le français. Mais pour incontournable qu'elle soit, cette référence est toujours un simple outil au service du chercheur, dont l'objet d'étude n'est évidemment jamais le FEW en lui-même. Ainsi, dialectologues, philologues, éditeurs, lexicographes consulteront avidemment le FEW, mais pour mieux construire leur objet propre — nous y reviendrons. Diverses catégories d'utilisateurs consultent donc l'ouvrage, avec des besoins variés et avec des ressources différentes face à la complexité du discours lexicographique. Ce sont majoritairement des spécialistes en leur domaine, mais une partie est constituée par les étudiants et par un public d'amateurs. Tous, y compris les spécialistes, rencontrent des difficultés de consultation et souhaitent une informatisation rapide de l'ouvrage.

3.2 Parcours de consultation et de lecture

L'avis des utilisateurs a pu être recueilli de diverses manières, d'abord via les publications scientifiques des disciplines concernées (voir par exemple Rey 1971 : 103-104 ; Roques 1991 : 94 ou encore Wooldridge 1998 : 211 ; cf. Renders 2011 : 8-15), ensuite via la diffusion d'un questionnaire au sein de la communauté internationale des chercheurs en linguistique française et romane lors du *XXVe Congrès International de Linguistique et Philologie Romanes* (cf. Renders 2010 ; pour les résultats, Renders 2011) et, enfin, via de nombreuses rencontres individuelles. Il a ainsi été possible d'obtenir un aperçu des pratiques actuelles d'utilisation du FEW et, dans un second temps, de connaître les souhaits des utilisateurs dans l'optique d'un FEW informatisé, souhaits en relation étroite avec les difficultés qu'ils rencontrent dans la consultation de la version imprimée. Les attentes exprimées sont révélatrices de la façon dont les utilisateurs voudraient exploiter le FEW et, plus généralement, de la façon dont ils perçoivent l'ouvrage et ses structures.

L'analyse des comportements d'utilisation du FEW a distingué deux activités distinctes et successives lors de l'utilisation du FEW : d'une part, la consultation, opération consistant à repérer dans le dictionnaire l'endroit où se trouve l'information que l'on recherche ; d'autre part, la lecture, opération consistant à s'approprier de façon complète l'information recherchée ainsi que l'analyse qu'offre le FEW en rapport avec cette information. Pour chacune de ces deux opérations, l'étude a montré des divergences entre les parcours actuels, permis par la version imprimée, et les parcours souhaités dans une version électronique.

En ce qui concerne la consultation, les souhaits des utilisateurs d'un futur FEW numérique induisent des itinéraires totalement nouveaux par rapport aux itinéraires traditionnels. En effet, dans la version imprimée, les points d'entrée dans le dictionnaire sont réduits aux étymons-vedettes (à condition de connaître la langue de l'étymon, qui détermine la partie superstructurelle et donc le volume à consulter) et aux lexèmes (à condition soit d'avoir une idée de leur étymon, soit de trouver ces lexème – ou des lexèmes apparentés – dans les index du FEW qui ne sont nullement exhaustifs : cf. ATILF 2003, qui remplace les divers index situés en fin de volume). En pratique, l'utilisation du FEW s'apparente souvent à un jeu de piste.[1] Dans la perspective d'une version électronique, les points d'entrée ne seraient toutefois plus réduits aux seuls lexèmes et lemmes (qui deviendraient en outre plus facilement repérables), mais s'étendraient fructueusement à tout type d'information présent dans le discours lexicographique (étiquettes géolinguistiques, sources bibliographiques, dates etc.). Ce mode de consultation « transversale », qui mène à plusieurs endroits dans le dictionnaire, est impossible dans la version papier du FEW, mais très attendu dans l'optique de son informatisation.

En ce qui concerne la lecture, l'étude a mis en évidence la complexité des itinéraires traditionnels, due à la nécessité, pour s'approprier l'analyse approfondie des données fournies par l'ouvrage, de mettre ces données en relation et en contexte à plusieurs niveaux. Par ailleurs, un aller-retour est constamment requis entre le dictionnaire et son *Complément*, qui explicite les nombreuses abréviations géolinguistiques et bibliographiques propres au FEW. Il est intéressant de constater que ces parcours présentent des variantes selon l'expérience qu'a l'utilisateur des structures du dictionnaire, selon ses compétences en linguistique française et dialectale et selon son besoin d'explicitation des abréviations. Les difficultés d'accès de l'ouvrage expliquent que de nombreux souhaits soient émis dans l'optique d'une informatisation, tels que la résolution des nombreuses abréviations, l'explicitation des sigles bibliographiques et des sources, la traduction des termes allemands ou, encore, la mise en évidence du plan des articles longs. La plupart de ces besoins se résolvent par des mises en relation (avec le commentaire de l'article, avec le *Complément*, avec des outils externes) qui, certes, sont possibles dans la version imprimée du FEW, mais seraient grandement facilitées par une informatisation de son contenu. Il ne s'agit donc pas de modifier les itinéraires de lecture classiques de la version imprimée, mais de faciliter la mise en relation de données qui, dans le discours lexicographique, ne sont pas situées côte à côte. Ce faisant, on ouvre la voie à des parcours de lecture hypertextuels qui n'étaient pas identifiés comme tels dans l'analyse des comportements d'utilisation du FEW papier, mais qui étaient sous-jacents.

3.3 Deux visions du FEW

L'analyse des parcours effectifs et des parcours souhaités par les utilisateurs a fait apparaître deux modes a priori contradictoires de consultation et de lecture du dictionnaire. Les difficultés de lecture s'expliquent en effet par la vision de l'article du FEW comme un discours construit et structuré, dans lequel chaque information est à mettre en relation avec celles qui l'entourent. Rappelons que les ar-

[1] Il faut ajouter à ces difficultés d'accès le problème des classements multiples, cf. Baldinger 1980.

ticles du FEW classent et hiérarchisent les données différemment, de façon à retracer l'histoire particulière de chaque famille lexicale, ce qui fait de chaque article une monographie à part entière :

> L'ouvrage se présente, en fait, comme un ensemble structuré de monographies, dont la forme lexicographique n'est qu'un auxiliaire au service de la « visée globalisante » (Swiggers 1990 : 347) de Wartburg, qui l'anime et la domine. (Büchi & Chambon 1995 : 952)

Cette particularité explique que les articles du FEW se lisent davantage qu'ils ne se consultent. Il s'agit de ce que nous appelons la *dimension monographique* (ou dimension M) du FEW, qui n'est accessible que par l'opération de lecture. Les demandes, partagées par les utilisateurs, d'un plan de l'article et d'une traduction du commentaire sont tout à fait liées à cette dimension monographique du FEW : elles visent à atteindre plus aisément le classement des données et l'analyse qui en découle.

L'enthousiasme des utilisateurs pour l'informatisation du FEW est toutefois davantage à expliquer par une autre vision de l'ouvrage, que nous appelons la *dimension thesaurus* (ou dimension T) du FEW. Dans cette vision du FEW comme un thesaurus, les utilisateurs sont intéressés par la masse de données qui s'y trouve et par les informations qui sont associées à chaque lexème. C'est le lexème, et non plus l'article, qui constitue leur centre d'intérêt. Le FEW est en effet le seul dictionnaire où se trouvent rassemblés tous les lexèmes des langues et dialectes du domaine galloroman, ce qui en fait un ouvrage des plus précieux dans de nombreuses sous-disciplines linguistiques. Cette dimension thesaurus est à l'œuvre lorsque l'utilisateur imagine des modes de consultation transversale, qui permettraient d'accéder directement à un groupe de lexèmes partageant un point commun malgré leur dispersion lexicographique due à leur appartenance à des familles lexicales différentes. Les besoins de mise en relation avec le *Complément* et avec d'autres dictionnaires, qui permettraient d'accéder directement à l'intégralité des références bibliographiques associées à un lexème, font partie de cette vision du FEW comme thesaurus. Enfin, c'est cette dimension qui explique que le FEW se consulte généralement à partir des lexèmes, alors que les entrées de la nomenclature sont des étymons.

Les deux dimensions dégagées ci-dessus ne sont pas inconciliables, mais étroitement imbriquées et complémentaires dans la construction du discours lexicographique. Si la consultation gagne à être envisagée dans une dimension thesaurus où chaque lexème est individualisé et accessible séparément, la lecture, quant à elle, ne peut s'effectuer que dans une dimension monographique, c'est-à-dire dans une mise en relation et une contextualisation des données.

Ces besoins, exprimés par les utilisateurs, ne sont pas résolus par les itinéraires de consultation et de lecture permis par le discours lexicographique sous sa forme actuelle. Les difficultés d'accès aux données du FEW ont toujours été attribuées à la présentation condensée et hautement structurée du discours lexicographique ; l'étude montre que ces difficultés proviennent également, si pas davantage, du fait que les utilisateurs veulent consulter le FEW dans une optique plus large que celle pour laquelle il a été conçu.

4 La prise en compte de l'utilisateur dans la version électronique

Partant de ce constat, il s'est avéré essentiel de résoudre l'inadéquation entre la conception initiale du dictionnaire et l'utilisation qui en est souhaitée, à la fois en tant que recueil de monographies et en tant que thesaurus. En dimension M, il s'agit de faciliter la lecture de l'article en tant qu'ensemble structuré et autonome et d'optimiser les itinéraires de mise en relation entre les informations (à la fois au sein d'un article et hors article). En dimension T, il s'avère nécessaire d'ouvrir l'accès aux nouveaux itinéraires de consultation attendus par la communauté (consultation transversale). Enfin, les deux dimensions sont concernées par la nécessité de permettre la mise à jour du FEW (intégration des ajouts et corrections apportés ailleurs) et sa mise en réseau avec d'autres ressources lexicographiques. Ces nouveaux parcours doivent être pris en compte non seulement au moment de créer l'interface de consultation, mais aussi dans la modélisation préalable et la formalisation XML du discours lexicographique.

4.1 Dimension monographique : itinéraires de lecture

Rétablir l'autonomie de l'article en dimension M nécessite tout d'abord de permettre son extraction hors du volume imprimé, tout en conservant les informations qui étaient données par le contexte physique et la situation de l'article au sein de l'ensemble. Rappelons que le FEW possède une superstructure divisant l'ouvrage en fonction de la langue d'origine des lexèmes et que les étymons-vedettes ne sont dès lors pas nécessairement partout classés par ordre alphabétique. La solution informatique consiste à « redescendre » au niveau de l'article les informations données dans les niveaux supérieurs (essentiellement le numéro de volume ainsi que la page où débute l'article). En pratique, ces informations sont automatiquement explicitées au début de l'article, dans les attributs des balises XML identifiant l'article et la colonne. Par exemple, la version XML de l'article MASCŬLĪNUS (FEW 6/1, 424b) commence ainsi :[2]

```
<art book="1" ici="1" id="0" lang="german" type="doc-com" volume="6">
<col pg="424" s="b"/>
```

Afin de faciliter la lecture de l'article en tant qu'ensemble structuré, il a été décidé de proposer au lecteur un plan résumant cette structure. Ce résumé ne consiste pas en une réécriture (impossible à automatiser et donc inconcevable, conformément au respect des trois contraintes précitées), mais, plus simplement, en l'affichage automatique de la première unité lexicale de chaque paragraphe, précédée du marquage alphanumérique situant le paragraphe dans la numérotation microstructurelle de l'article (réexplicitée si nécessaire) et, lorsqu'il existe, du marqueur textuel (titre de section) explicitant le

[2] L'attribut *ici* (*in-column index*) indique l'ordre de l'article dans la colonne, *lang* la métalangue utilisée dans le commentaire; l'attribut *type* indique si l'article est divisé en une partie documentaire (*doc*) et une partie de commentaire (*com*) (cf. Büchi 1996 : 78).

critère de regroupement des lexèmes au sein du paragraphe. Le plan de l'article CHOCOLATL (FEW 20, 63b-64a) se présente par exemple ainsi dans sa version XML :

<!--article map
1 Mfr. chocholate m. „breuvage fait avec des amandes de cacao" (1598)
1 Ablt. — Nfr. chocolatière f. „vase où l'on prépare, où l'on sert le chocolat en boisson" (seit 1680)
2 Nfr. chicolate f. „chocolat" (1658)
-->

Ce plan, affiché en tête d'article, donne immédiatement au lecteur une vision synthétique de la structure de l'article (classement des lexèmes au sein des différentes subdivisions) : ici, trois paragraphes structurés en deux parties numérotées, la première partie reprenant des dérivés (*Ablt.* pour *ableitungen*). Ce faisant, il aide à saisir d'un coup d'oeil l'organisation générale de la famille lexicale traitée.

Enfin, les itinéraires de lecture sont également facilités par la création de liens hypertextuels internes et externes. Les liens internes concernent d'une part les notes et appels de note (identifiés et associés de façon à faciliter le passage de l'un à l'autre), d'autre part les références renvoyant, dans le commentaire, au marquage alphanumérique structurant les matériaux. Par exemple, toujours dans l'article CHOCOLATL, le commentaire explique l'origine étymologique de chacune des deux parties numérotées. Tant les marqueurs alphanumériques que les références à cette numérotation (*<pref>*) ont été automatiquement reconnus et balisés, ce qui permettra la création de liens hypertextuels :

<pref id="1">1</pref> ist aus <lang>sp.</lang> <form><i>chocolate</i></form> entlehnt [...]. Unklar ist auch das verhältnis von <pref id="2">2</pref> zu <pref id="1">1</pref>. Der erste beleg von <pref id="2">2</pref> kommt von den kleinen Antillen, wodurch wahrscheinlich gemacht wird, dass diese form sich selbständig verbreitet hat.

Les liens externes concernent d'une part les renvois à d'autres articles du FEW (mis en oeuvre par le balisage automatique des étymons, volumes et pages), d'autre part les renvois externes au dictionnaire. Le balisage des sigles bibliographiques permet en effet de créer un lien hypertextuel vers leur explicitation (fournie dans la base de données contenant le *Complément* au FEW) et vers la ressource électronique si cette dernière existe.

Ces trois nouveautés apportées par la version électronique (autonomie de l'article, résumé de sa structure, liens hypertextuels) permettent de faciliter et d'optimiser les parcours de lecture du FEW dans sa dimension monographique.

4.2 Dimension thesaurus : itinéraires de consultation

En dimension T, il s'agit en priorité de permettre une consultation du FEW via les lexèmes (unités lexicales) et les informations qui y sont associées. L'autonomie de chaque unité lexicale est rétablie en balisant au sein d'un même élément XML les informations qui la composent (étiquette géolinguistique, signifiant, catégorie grammaticale, définition, références) et en rétablissant celles de ces informations qui seraient implicites, cas de figure très fréquent dans le FEW puisque certaines informations (étiquette géolinguistique, catégorie grammaticale, signifiant, définition) ne sont jamais répétées si elles ont déjà été citées dans l'unité précédente (cf. Büchi 1996 : 117 et Renders 2011 : 76-81). Le principe de « redescente » des informations s'applique donc également ici. Dans l'article ACCUSATIVUS par exemple (FEW 24, 94b), qui comporte uniquement deux lexèmes *accusatif*, l'un substantif masculin (« cas auquel on met le complément direct »), l'autre adjectif (« qui concerne l'accusatif »), l'étiquette géolinguistique et le signifiant associés au deuxième lexème n'ont pas été répétés. La version XML rétablit ces informations grâce à l'insertion automatique d'une balise <imp>, identifiant le type d'information manquant et son contenu implicite :

<unit><imp contents="Mfr." type="geoling"/><imp contents="accusatif" type="form"/><gram>adj.</gram> <def>„qui concerne l'accusatif"</def>

Ce balisage permet d'extraire de chaque article toutes les unités lexicales qu'il contient et de leur rendre leur autonomie, indépendamment de leur insertion dans le discours monographique. Une liste de toutes les unités est systématiquement créée pour chaque article ; cette liste attribue en outre à chaque lexème son étymon (étymon-vedette de l'article) et sa référence FEW. Toujours pour l'article ACCUSATIVUS, la liste créée est la suivante :

<fiche etymon="accusativus" lang="Mfr." lang="nfr." forme="accusatif" gram="m." def="„cas auquel on met le complément direct"" ref="(seit ca. 1170, EdConf, FrMod 21, 217)" N="FEW 24/1, 94b, ici 1, §1, u1"></fiche>

<fiche etymon="accusativus" lang="(imp.) Mfr." forme="(imp.) accusatif" gram="adj." def="„qui concerne l'accusatif"" ref="(1380, Aalma 98; Pom 1671–1700; Lar 1866–1948)" N="FEW 24/1, 94b, ici 1, §1, u2"></fiche>

Le balisage rend ainsi possible une consultation transversale du FEW via les informations associées aux lexèmes : consultation par parler (étiquette géolinguistique), par catégorie grammaticale, par élément de définition, par signifiant, par référence bibliographique, par étymon. La consultation par critères chronologiques nécessite en outre que les dates correspondant à chaque sigle bibliographique soient explicitées, ce qui s'effectue via le *Complément* au FEW.

Enfin, des consultations à cheval entre la dimension monographique et la dimension thesaurus, à savoir selon la langue d'origine des lexèmes (souvent implicite dans le FEW, car associée à la partie superstructurelle où se trouve l'article) ou selon le type de descendance (emprunts ou lexèmes héréditaires) sont également rendues possibles par le balisage.

4.3 Sortir du FEW

Comme évoqué au point 3.1., les utilisateurs du FEW sont divers, de même que leurs objets d'étude. Le parcours dialectique entre le FEW et ces diverses entreprises peut être illustré par un exemple qui nous est familier, celui de l'*Atlas linguistique de la Wallonie* (ALW). Tout au long de l'analyse des matériaux de l'ALW, les rédacteurs tissent un dialogue entre leur ouvrage et le formidable outil qu'est le FEW

: d'une part, en extrayant de celui-ci les données servant à éclairer les matériaux wallons, d'autre part, en amenant divers compléments, amendements ou ajustements. Au fil de ce cheminement se développe « une véritable réévaluation de l'état de l'art représenté par le FEW » (Chauveau & Buchi 2011 : 12).

Si les spécialistes s'accordent à reconnaître un certain intérêt aux compléments apportés par l'ALW au dictionnaire de Wartburg, ceux-ci sont de toute évidence difficilement accessibles. Cette conclusion peut être étendue à d'autres ouvrages, qu'ils soient atlantographiques, lexicographiques ou philologiques. Dès lors, il nous semble que prendre en compte l'utilisateur du FEW passe également par une intégration, à quelque niveau que ce soit, de ses apports à la construction d'un savoir commun.

Le balisage de l'œuvre selon les modalités exposées ci-dessus permettra une navigation optimisée et des consultations transversales internes. Ce balisage pourrait ensuite être exploité pour développer des liens externes vers divers projets apparentés, tels que l'ALW pour des apports ponctuels ou des entreprises telles que DEAF, Godefroy, TLFi etc. pour des apports plus systématiques. Ces divers ouvrages intégrant déjà dans leur programme lexicographique des références au FEW, leur mise en réseau nécessite uniquement que le FEW informatisé dispose d'url pérennes qui puissent servir à la fois de référence et de lien hypertextuel. Toutefois, l'intégration des apports externes pourrait aller plus loin qu'une simple mise en réseau et conduire à un FEW évolutif. Cette idée encore utopique est en phase avec la dimension T du FEW, puisqu'elle permettrait des consultations basées sur des critères (géolinguistiques, chronologiques, étymologiques ou autres) qui correspondraient à l'état actuel de la science et non à une version périmée. Le balisage du FEW en dimension T, et plus particulièrement le rétablissement de l'autonomie des unités lexicales, devrait permettre ces consultations sans que la dimension M du FEW n'en soit affectée.

5 Conclusion

Loin de constituer une modélisation théorique et gratuite du dictionnaire, les deux dimensions décelées dans la conception et l'utilisation effective du FEW d'une part, mais également dans l'exploitation que rêvent d'en faire ses utilisateurs, ont très concrètement guidé la réflexion dans le cadre du projet. Nous espérons avoir montré comment ces deux dimensions s'articulent et devront continuer à le faire, de la façon la plus explicite possible, dans la future version électronique.

En outre, nous avons montré en quoi les fonctionnalités permettant l'exploitation du dictionnaire dans ses deux dimensions ne concernaient pas seulement la création de l'interface utilisateur, mais devaient être prévues dès la phase de modélisation du discours lexicographique. Aborder les structures du dictionnaire selon le point de vue de l'utilisateur est la condition nécessaire pour pouvoir rendre compte de la façon dont le FEW est, non plus conçu et rédigé, mais perçu et utilisé.

Bien entendu, c'est au niveau de l'interface, toujours en développement à l'heure où nous écrivons ces lignes, que l'utilisateur prendra pleinement conscience des potentialités des nouveaux parcours qui lui sont offerts. Gageons cependant que cet avant-goût aiguisera encore davantage l'appétit des chercheurs!

6 Bibliographie

ALW = Remacle, L. et al. (1953–). Atlas linguistique de la Wallonie. Tableau géographique des parlers de la Belgique romane d'après l'enquête de Jean Haust et des enquêtes complémentaires. Liège: Vaillant-Carmanne.

ATILF (2003). Französisches Etymologisches Wörterbuch. Index A–Z. Paris: Champion.

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

Baldinger, K. (1980). Etymologies doubles dans le FEW. In H. J. Izzo (éd) Italic and Romance: Linguistic Studies in Honor of Ernst Pulgram. Amsterdam: Benjamin, pp. 189-194.

Büchi, E., Chambon, J.-P. (1995). Un des plus beaux monuments des sciences du langage : le FEW de Walther von Wartburg (1910-1940). In G. Antoine, R. Martin (éd.) Histoire de la langue française, 1914-1945. Paris: CNRS Editions, pp. 935-963.

Büchi, E. (1996). *Les structures du* Französisches Etymologisches Wörterbuch*. Recherches métalexicographiques et métalexicologiques*. Tübingen: Niemeyer.

Chauveau, J.-P., Buchi, E. (2011). État et perspectives de la lexicographie historique du français. In *Lexicographica. International Annual for Lexicography*, 27, pp. 101-122.

Complément = Chauveau, J.-P., Greub, Y. & Seidl, C. (2010). Französisches Etymologisches Wôrterbuch. Eine darstellung des galloromanischen sprachschatzes. Supplement zur 2. Auflage des Bibliographischen Beiheftes. Bâle: Zbinden.

DEAF = Baldinger, K. *et al.* (1974–). *Dictionnaire Étymologique de l'Ancien Français*. Québec/Tübingen/Paris: Presses de l'Université Laval/Niemeyer/Klincksieck. Site internet : http://deaf-server.adw.uni-heidelberg.de.

Dendien, J., Pierrel, J.-M. (2003). Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. In *Traitement automatique des langues (TAL)*, 43 (2), pp. 11-37.

FEW = von Wartburg, W. *et al.* (1922-2002). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes* (25 vol.). Bonn/Heidelberg/Leipzig-Berlin/Bâle: Klopp/Winter/Teubner/Zbinden.

Godefroy = Godefroy, F. (1881-1895). Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle (8 vol.). Paris: Vieweg.

Malkiel, Y. (1976). *Etymological dictionaries. A tentative typology*. Chicago-London: The University of Chicago Press.

Renders, P. (2010). L'informatisation du *Französisches Etymologisches Wörterbuch* : quels objectifs, quelles possibilités ?. In M. Iliescu, H. Siller-Runggaldier & P. Danler (eds.) Actes du XXVe Congrès International de Linguistique et de Philologie Romanes (Innsbruck, 3-8 septembre 2007). Berlin/New York: De Gruyter, vol. 6, pp. 311-320.

Renders, P. (2011). Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du *Französisches Etymologisches Wörterbuch*. PhD Thesis. Université de Liège, Liège, BE.

Renders, P. (à paraître). Mise en ligne, mise à jour et mise en réseau du *Französisches Etymologisches Wörterbuch*. In D. Trotter, A. Bozzi & C. Fairon (éds.) Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Nancy: ATILF.

Rey, A. (1971). Le dictionnaire étymologique de W. von Wartburg : structures d'une description diachronique du lexique. In *Langue française*, 10, pp. 83-106.

Roques, G. (1991). L'articulation entre étymologie et histoire de la langue. In *Travaux de linguistique*, 23, pp. 91-95.

Swiggers, P. (1990). Lumières épistolaires sur l'histoire du F.E.W. : Lettres de Walther von Wartburg à Hugo Schuchardt. In *Revue de linguistique romane*, 54, pp. 347-358.

TLFi = CNRS/Université Nancy2/ATILF (2004). *Trésor de la langue française informatisé* (cédérom). Paris: CNRS Éditions (site internet : http://www.cnrtl.fr/definition).

Wooldridge, T. R. (1990). Le FEW et les deux millions de mots d'Estienne-Nicot : deux visages du lexique français. In *Travaux de linguistique et de philologie*, 28, pp. 239-316.

Wooldridge, T. R. (1998). Le lexique français du XVIe siècle dans le GDFL et le FEW. In *Zeitschrift für romanische Philologie*, 114, pp. 210-257.

A Morphological Historical Root Dictionary for Portuguese

João Paulo Silvestre, Alina Villalva
Centro de Linguística da Universidade de Lisboa
jpsilvestre@fl.ul.pt, alinavillalva@campus.ul.pt

Abstract

The project of a Morphological Historical Root Dictionary (MHRD) for Portuguese aims to build a specialized dictionary containing a critical selection of lexical units: the first stage is devoted to adjectives, namely those that can be found in Figueiredo (1913). Its main goals are the clarification of morphological and semantic issues in the evolution of the lexicon, and the assessment of the communicative adequacy of the words that are registered in current Portuguese dictionaries, particularly by signaling unused or seldom used words.

The methodology we established is three-sided: it first relies on the lexical analysis of the selected words; then, it seeks for lexicographic information in old Portuguese dictionaries (16th to 19th centuries), in Portuguese textual corpora and in etymological dictionaries; finally, it contrasts the Portuguese data with data from other romance languages, searching for lexical and semantic loans.

To conclude, we propose a prototype dictionary entry, applying the above-described methodology to the adjective *bravo.*

Keywords: word roots; morphology; Portuguese; historical lexicography

1 Preliminary remarks

The entry list of most contemporary Portuguese dictionaries still echoes lexicographic approaches that most of the other European languages have discarded along the past century. The main problem resides in the fact that they give room to the vast majority of the contents of previous dictionaries, thus accumulating, with an identical status, words that make part of the contemporary lexicon and a huge amount of unused words. This abundance of entries renders other problems usually found in dictionaries, such as graphic alternates, inadequate meanings and wrong etymological information, to name a few.

Furthermore, in the last two decades, mainstream dictionary publishing houses have rendered their workforce to the uncompelling issue of the orthographic 'entente' between Portugal, Brazil and (eventually) all other Portuguese-speaking countries, which motivated 'new', 'updated' paper editions. The remaining workforce of dictionary companies is fully devoted to the introduction of 'neologisms' that will haunt future editions.

There is an urge, then, to review the wordlists, to clean each entry, to correct the information given, to expunge old unused words. A dictionary may, of course, be cumulative about the selection of entries, but it must mark those that are unused, although they can be found in old literary texts.

2 Diachronic incoherence in Portuguese contemporary lexicography

Take, for instance, the case of the verb *abundar* 'to abound'. *Infopedia* is an online dictionary that claims to be the most complete dictionary of European Portuguese, covering general, technical and scientific vocabularies. Surprisingly, it registers words like *bondar*, *abondar* and *avondar* that are certainly not part of those vocabularies:

(1) Avondar. Verbo transitivo e intransitivo. Ver abundar. Do latim abundāre, «abundar» (*Infopédia*)

(2) Abondar. Verbo intransitivo. Regionalismo. Ser suficiente; bastar; bondar (Do latim abundāre, «idem» (*Infopédia*)

(3) Bondar.Verbo intransitivo. Popular. Bastar; ser suficiente (Do latim abundāre, «trasbordar; abundar» (*Infopédia*)

Probably, the form *avondar* is the oldest in Portuguese – it can be found in 14th to 17th century textual sources; *abondar* can also be found between the 14th century and the 19th; *abundar* starts in the 16th century and is the only form in contemporary usage[1]. Notice that the third form, i.e. *bondar*, marked in *Infopedia* as a 'popular' form has very few registers in the *Corpus do Português* database[2]. In fact, it has only two, one of which comes from an oral corpus, and it is quite difficult to understand.
A search in non-contemporary lexicographic sources hekps to consolidate the hypothesis above: *avondar* occurs in 16th and 17th century dictionaries, already marked as peripheral; *abondar* occurs in the 16th century, and in an 18th century it is considered as an error. The contemporary form, i.e. *abundar*, which is graphically closer to the spelling of the Latin verb (i.e. *abundare*), appears in the 17th century. Curiously, the recovery of all these variants began in 19th century dictionaries, such as Morais and Figueiredo. *Infopedia* replicates them, particularly Figueiredo.

(4) Auondar. Vide Abondar (Cardoso 1569)

(5) Avondar. Vide Abundar (Pereira 1697)

(6) Abundar. Ter abundancia. Erro, Abondar (Feijó 1734)

(7) Avondar, n. Abastar, ser bastante em numero. Antiq. [antiquate] (Silva 1831)

(8) Bondar v. i. Prov. Sêr bastante, sufficiente: mas isso não bonda. Alter. de abundar. (Figueiredo 1899)[3]

[1] Examples were taken from the database *Corpus do Português.*

[2] Only 7 matches found in texts, all from the 19th century (*Corpus do Português*).

[3] Examples were taken from the database *Corpus Lexicográfico do Português.*

Filtering a general dictionary such as *Infopedia* is a desirable, but difficult task that requires highly trained manpower. This task is out of range for individual good will and low budget.

As we mentioned before, general contemporary Portuguese dictionaries either accumulate information from previous dictionaries, regardless of the errors they perpetuate and the real usage of the words, or they are produced on the basis of modern *corpora*, that integrate limited amounts of data, thus ignoring all the words that are not represented there. Lexical *corpora*, apart from coverage limitations, also frequently lack morphological tagging.

3 Planning a specialized historical dictionary

The MHRD project was designed in order to give a constructive response to such a negative perspective in the field of contemporary dictionary making. This project aims to build a specialized dictionary, which will contain a critical selection of the lexicon of Portuguese. Its main goals are:

- the clarification of the process of morphological and semantic evolution of the lexicon;
- the assessment of the communicative adequacy of the words that are registered in current Portuguese dictionaries, mostly by signalling unused or seldom used words.

MHRD will, thus, include simple and complex lexicalized roots, documented in a set of selected early lexicographic sources for Portuguese.

3.1 Lexicographic sources

Revisiting old dictionaries is thus obligatory, but instead of aiming to consider all of them, in the preliminary stage of this project, we decided to make a selection based on an analysis of lexicography in Portugal. The set of dictionaries that form this *canon* was selected for qualitative reasons, since they all played an important role either for the quality of the information they provided or for the normative role that they assumed. The selection, ranging from the 16th century to the end of the 19th century, includes:

- Jerónimo Cardoso (1569) *Dictionarium Latinolusitanicum / Lusitanicolatinum* — The first printed dictionary with extended word list in Portuguese (about 12 thousand entries). It is a testimony of ancient lexical choices and word forms, prior to the systematic imitation of Latin in neologisms and spelling.
- Bento Pereira (1697) *Prosodia in Vocabularium Bilingue, Latinum, et Lusitanum* — Extensive Latin-Portuguese learners dictionary, which represents the increase of lexical variety in the 17th century, by adapting many Latin words into Portuguese. In the corpus, there are about 50 thousand Portuguese word forms.

- Raphael Bluteau (1712-28) *Vocabulario Portuguez e Latino* — The first dictionary with examples from literary Portuguese texts, with more than 40 thousand entries. It makes a systematic collection of terminology and neologisms resulting from loans.
- António Morais Silva (1789) *Diccionario da Lingua Portugueza* — It is the first monolingual dictionary of the Portuguese, with a modern lexicographical technique. We also considered the fourth revised and extended edition of this dictionary, published in 1831. It was an authoritative dictionary throughout the nineteenth century. As a general rule, it notes old or unused words.
- Cândido de Figueiredo (1899) *Novo Diccionário da Língua Portuguesa* — A cumulative dictionary, which collects ancient and modern words without consistently noting their effective use in contemporary language. Served as lexical corpus to dictionaries in the 20th century, which reproduced the word list with little critical review.

Finally, in order to ascertain the usage of words in contemporary Portuguese, we consulted the *Corpus de Referência do Português Contemporâneo*. Completion and crosschecking of lexical analysis relies on the consultation of etymological dictionaries (Corominas and Pascual 1991) and historical dictionaries (*Le Trésor de la Langue Française Informatisé* and *El Nuevo Tesoro Lexicográfico de la Lengua Española*).

3.2 Root identification

Since feasibility is one of our main concerns, we decided to limit our research to simple (unanalysable) roots. We believe that, once we have achieved to isolate the core set of roots, we will be better equipped to identify and describe derived and compound words. Simple words, those that are projected from single roots, have a supplementary advantage: they are usually old words and old words tend to accumulate or to change meanings. Those moments, which are not easy to detect, are seldom documented.

The identification of the core set of roots is certainly crucial for a better understanding of the Portuguese lexicon, but no existing general or specialised Portuguese dictionary or lexical corpus provides this information. Two specialized dictionaries deserve to be mentioned, however. The first one is the *Dicionário de Raízes e Cognatos* [cf. Goes (1921)]. Apart being based on 19th century sources, it mainly deals with the small subset of neoclassical roots. The second one is the most important Portuguese morphological dictionary. It was made in Brazil, by gathering lexical information from unspecialized sources, such as general language dictionaries [cf. Heckler, Back, Massing (1984-1988)]. Both of them offer interesting data, but their consultation also needs extensive critical reading.

Most Portuguese words, irrespectively of their longer or shorter existence in the Portuguese lexicon, come from a relatively stable set of roots, which can be documented in morphologically simple words (as free roots) or in complex words (as free roots in compositional words and as bound roots in lexicalized words). This typology of roots (based on Villalva and Silvestre [in print]) also foresees cases of bound roots in compositional complex words:

Root	Head		Complement		Modifier	
	Free root	Bound root	Of a derivational suffix	Of another root	Of a simple word	Of a complex word
Type 1 e.g. rat-	rat-o ‘mouse’		rat-ice ‘cunning’	rat-icida ‘rat poison’		
Type 2 e.g. gastr-			gástr-ico ‘gastric’	gastr-onomia ‘gastronomy’		
Type 3 e.g. graf-	graf-ar ‘to write’	pluvió-graf-o ‘rain gauge recorder’	gráf-ico ‘graphic’	graf-ologia ‘graphology’		
Type 4 e.g. super-					super-amigo ‘super-friend’	super-confortável ‘super-comfortable’

Table 1: Typology of roots.

Therefore, a specific methodology had to be established. Bearing feasibility in mind, we decided to devote our initial research to adjectives. The approach we decided to take is three-sided: it relies on the lexical analysis of the selected words; it seeks lexicographic information in old Portuguese dictionaries (16th to 19th centuries) and Portuguese textual *corpora* and in etymological dictionaries; it contrasts the Portuguese data with data from other romance languages, searching for lexical and semantic loans.

The first stage of the project is devoted to adjectives, the second to nouns and the third one to verbs.

3.3 Adjectives

The lexical analysis of adjectives considers their morphological, syntactic and semantic properties. Notice that, from a morphological point of view, Portuguese adjectives are not significantly different from nouns, which raises a practical problem for the selection of roots.

Adjectives and nouns, they both require number inflection (cf. Table 2, i) and they both comprise a subset that allows for gender variation (cf. ii) and a subset of invariable forms (cf. iii).

i)	ii)	iii)
casa $_{\text{Nsingular}}$ ‘house’ casas $_{\text{Nplural}}$ ‘houses’ leve $_{\text{ADJsingular}}$ ‘light’ leves $_{\text{ADJplural}}$ ‘light’	gato $_{\text{Nmasculine}}$ ‘male cat’ gata $_{\text{Nfeminine}}$ ‘female cat’ novo $_{\text{ADJmasculine}}$ ‘new’ nova $_{\text{ADJfeminine}}$ ‘new’	casa $_{\text{Nfeminine}}$ ‘house’ carro $_{\text{Nmasculine}}$ ‘car’ leve $_{\text{ADJ}}$ ‘light’

Table 2: Adjectives - morphology.

There are, nevertheless, differences that can be spotted. As far as number is concerned, although its specification is compulsory for nouns and adjectives alike, in nouns it has semantic relevance (singular refers one entity, plural refers more than one), in adjectives it is semantically irrelevant: adjectives have no quantifiable meaning - number inflection is merely relevant for syntactic agreement (*casa* $_{\text{singular}}$ *nova* $_{\text{singular}}$ ‘new house’; *casas* $_{\text{plural}}$ *novas* $_{\text{plural}}$ ‘new houses’).

Gender is even more diverse. All nouns have to have a gender value, which is lexically determined, irrespective from their possibility to participate in gender contrasts (cf. *gato* / *gata*; *carro* and *casa*, in table 3, i). In general, animate nouns can participate in gender contrasts either by thematic alternation (cf. i), by a morphological resource (cf. ii) or lexically (cf. iii), but some animate nouns do not, which eventually creates a mismatch between grammatical gender and the gender of the referent (cf. iv):

i)	ii)	iii)	iv)
gato‘male cat’ gata ‘female cat’ aluno ‘male student’ aluna ‘female student’	galo ‘rooster’ galinha ‘hen’ marquês ‘marquis’ marquesa ‘marchioness’	cavalo ‘horse’ égua ‘mare’ homem ‘man’ mulher ‘woman’	testemunha $_{\text{feminine}}$ ‘witness (male or female) cônjuge $_{\text{masculine}}$ ‘spouse (husband or wife) águia $_{\text{feminine}}$ ‘eagle (male or female)’ águia-macho $_{\text{feminine}}$ ‘male eagle’ águia-fêmea $_{\text{feminine}}$ ‘female eagle’ rinoceronte $_{\text{masculine}}$ ‘rhino’ rinoceronte-macho $_{\text{masculine}}$ ‘male rhino’ rinocerente-fêmea $_{\text{masculine}}$ ‘female rhino’

Table 3: Adjectives – gender contrast.

Inanimate nouns are never allowed to participate in gender contrasts. It is possible to find pairs of words that apparently share the same root, although they belong to different thematic classes. They are not in a gender contrast - they are different words (*casa* $_{\text{feminine}}$ ‘house’; *caso* $_{\text{masculine}}$ ‘case’).

For adjectives, gender is as irrelevant as number. It may be syntactically important, for agreement, but a great deal of adjectives is invariable, so the syntactic relevance is also questionable – it is probably just a vestige from Latin declension.

Apart from these morphosyntactic properties, adjectives and nouns also share the possibility to undergo evaluative affixation. It is particularly relevant to notice that the most productive suffix (i.e. *–inh{o, a}(s)* is equally available, but its semantic effect on nouns differs from its semantic outcome in adjectives. In the first case, it is typically a diminutive or valuative (cf. table 4, i); in the second case its reading is ambiguous – typically, it can either be an attenuative or a superlative (cf. Table 4, ii). On the other hand, the superlative forming suffix (i.e. *-íssimo*) only adjoins to adjective bases, but a large set of unquestionable adjectives are not scalable and thus they do not allow the adjunction of this suffix (cf. Table 4, iii).

i)	ii)	iii)
casinha 'small house' carrinho 'small car' gatinho 'small (dear) cat'	novinho 'pretty/very young'	*casíssima 'very+house' novíssimo 'very+new' *teatralíssimo 'very+theatrical'

Table 4: Adjectives - afixation.

In sum, morphology can provide some clues to help setting adjectives apart from nouns, but it fails to draw a neat borderline. Syntactic distribution has, thus, to be considered as well, in order to characterise adjectives as a word class; it is also relevant to establish adjective subclasses. The next set of examples comprises a subset of words that can only occur in an adjective position and another subset of nouns that never occur as adjectives (*cabelo*$_{N}$ *fino*$_{ADJ}$ 'thin hair'); a third subset includes words that occur in adjectival and nominal contexts (*professor*$_{N}$ *assistente*$_{ADJ}$ 'assistant professor'; *assistente*$_{N}$ *do produtor*$_{N}$ 'producer's assistant'):

In order to get a better understanding of adjectives on the basis of syntactic criteria, probably the most relevant are those that concern word order and the possibility to occur in predicative positions as well as in non-predicative positions. Colour adjectives, for instance, can never occur in a prenominal position (*vestido vermelho* 'dress red'; **vermelho vestido* 'dress red'), but other adjectives can (*velho hábito* 'old habit'), although this is a marked word order, except for ordinal adjectives (*primeiro dia* 'first day'; **dia quinto* 'fifth day'):

The predicative *vs.* non-predicative distinction also fails to clearly set adjective subclasses: most adjectives can have both distributions (*o vestido vermelho* 'the red dress'; *o vestido é vermelho* 'the dress is red') and small sets of adjectives have exclusive distribution (*o primeiro dia* 'the first day; *o vestido é vermelho* 'the dress is red').

Finally, we need to consider the semantics of adjectives, which is probably the most difficult aspect to deal with. Several ontologies have been suggested in the literature, but none of them is able to avoid very specific world knowledge constraints.

We will also use a set of criteria (somehow in parallel with morphological and syntactic criteria above considered) that will apply to each form. The first condition concerns gradability, measured on the basis of *–íssimo* affixation and also on the basis of syntactic comparative constructions. This condition allows us to identify three subsets of adjectives: those that respond positively to both tests (cf. table 5, i), those that respond positively just to *–íssimo* (cf. ii)[4] and those that respond negatively to both of them (cf. iii). Notice that this condition has to be tested in a specific syntactic context, since the result is not always the same (cf. iv):

[4] This contrast is probably due to the fact that the evaluative suffix is closer to a rhetoric resource than to the setting of a degree.

i)	preço alto ‘high price’ preço altíssimo ‘very+high price’ preço muito alto ‘very high price’
ii)	casa enorme ‘enormous house’ casa enormíssima ‘very+enormous house *casa muito enorme ‘very enormous house’
iii)	sais minerais ‘mineral salts’ *sais mineralíssimos ‘very+mineral salts’ *sais muito minerais ‘very mineral salts’
iv)	os cavalos estão cansados ‘the horses are tired’ os cavalos cansados não ganham corridas ‘tired horses don’t win races’ os cavalos estão cansadíssimos ‘the horses are very+tired’ *os cavalos cansadíssimos não ganham corridas ‘very+tired horses don’t win races’

Table 5: Adjectives – *íssimo* afixation.

The second condition concerns the possibility to relate an adjective to another adjective, usually by opposition. Rasken and Nirenburg distinguish binary oppositions of non-gradable, complementary antonyms (cf. table 6, i), polar oppositions of gradable antonyms (cf. ii) and multiple non-gradable oppositions (cf. iii):

i)	ii)	iii)
morto vs. vivo ‘dead’ vs ‘alive’	(*muito frio*) *frio* (*pouco frio*) vs. (*pouco quente*) *quente* (*muito quente*) ‘very cold’ ‘cold’ ‘bit cold’ vs. ‘bit hot’ ‘hot’ ‘very hot’	*análise histórica* ‘historical analysis’ ……… *económica* ‘economic’ … ……… *social* ‘social’ … ……… *política* ‘political’ …

Table 6: Adjectives – sense relations.

Finally, we need to consider the semantic features of the antecedent, since this information is of crucial importance to circumscribe the meaning of adjectives. In particular, it is necessary to identify the value of animacy, humanness, countability, concreteness. Notice, for instance, that the adjective *bravo* (described in some detail in Villalva & Silvestre 2011) has accumulated different meanings, ranging from a negative pole to a positive pole. In the first case, it means ‘ferocious’ if it applies to animals (wild animals), but it will mean ‘angry’ when it gets to be applied to people.

(9) $N_{[-human]}$ *bravo* ‘ferocious’ (Cardoso 1569)

(10) $N_{[+human]}$ *bravo* ‘courageous’ (Bluteau 1712-1728)

(11) $N_{[+human]}$ *bravo* ‘angry’ (Silva 1789)

3.4 *–mente* adverbs

The classification of deadjectival adverbs may also bring some problems. Consider the case of *alto*, which is a good representative of a set of adjectives that are used to measure dimension, in its various features (height, length, weight, etc.). This is a word of Latin origin (i.e. *altus*), originally a participle of the verb *alo* 'to feed', thus meaning 'fed, grown'. As an adjective, *altus* had two basic meanings[5], related to the perception of height (A. Seen from below upwards, *high* and B. Seen from above downwards, *deep*). Both of them allowed a figurative, non-physical interpretation, respectively 'elevated, distinguished' and 'profound'.

The semantic interpretation of the Portuguese adjective *alto* (as well as the Spanish cognate, *alto*, the French *haut* or the Italian *alto*) replicates the semantics of the A. interpretation of the Latin *altus*. Derived nouns, such as *altura* 'height' and *alteza* 'highness', as well as derived verbs such as *altear* 'to heighten' and *enaltecer* 'to praise', help to consolidate that conclusion.

If no other reason existed, the adjective *alto* would not deserve much more comment, but that is not the case if we also consider the derived adverb *altamente*. All romance languages share an adverb forming resource which is based in the grammaticalization of a Latin noun (*mens, mentis*) that took place prior to the individuation of Romance languages, thus explaining its vitality still in contemporary strata. In Latin, the ADJ + *mente* sequence had an adverbial usage. Probably, -*mente* adverbs (or adverbs to be) in Romance languages were initially strictly manner adverbs. Contemporary usage is a bit more complex: apart from manner adverbs (e.g. *elegantemente* 'elegantly', *injustamente* 'unfairly'), -*mente* also forms temporal locatives such as those in table 7, generally equivalent to a locution that includes the base adjective (i.e. *anteriormente* = num momento anterior; *imediamente* = num momento imediato; *futuramente* = num momento futuro).

Past	Present	Future
anteriormente	actualmente	brevemente
antigamente	presentemente	futuramente
inicialmente		seguidamente

Table 7: Adverbs - Temporal locatives.

Although the formation of –mente adverbs is considered as a very productive word formation process, some restrictions have been identified. According to Scalise (1990), Italian -*mente* cannot be attached to several types of adjectives. Unsurprisingly, the same negative constraints apply in Portuguese, so –*mente* cannot apply to possessive (cf. **minhamente*), demonstrative (cf. **estamente*), indefinite (cf. **qualquermente*), or numeral adjectives (cf. **umamente*), nor can it be attached to qualifying adjectives denot-

[5] http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0059%3Aentry%3Daltus1

ing physical qualities (cf. **gordamente*), colour adjectives (cf. **vermelhamente*), superlatives (cf. **melhormente*) and adjectives modified by evaluative suffixes (cf. **ligeirinhamente*).

Considering these negative constraints, the adjective alto should not yield a *–mente* adverb, but *altamente* is a quite frequently used word. Scalise (1990) considers that polysemic adjectives that refer a physical property and a psychological property can derive a *–mente* adverb from the second meaning[6]. In fact, *altamente* is never related to the physical meaning of *alto* – it never means 'in a high manner'; but it not related to the psychological meaning of *alto* – it never means 'in an elevated manner'.

Evidences from ancient dictionaries confirm that there are only occurrences of polysemous meanings of *altamente* and the same applies to other adjectives indicating size or extension, such as *baixo, largo, estreito, comprido, longo, curto.* Should be noted that in the same sources the adverbs *largamente* and *longamente* are considered as synonyms (cf. 21).

(12) Profundè. aduer. *Alta* & fundamente. (Cardoso 1569)
(13) Tragice loqui Falar *altamente*. como em tragedia. (Cardoso 1569)
(14) Eminenter, adv. Excellentemente, *altamente*. (Pereira 1697)
(15) *Altamente*. Alte. Sublimiter. (Pereira 1697)
(16) *Baixamente*. Abjecte. Ignobiliter. (Pereira 1697)
(17) Spatiose, Adv. *Larga*, espaçozamente. (Pereira 1697)
(18) Largiter, Adv. *Larga*, liberal, *abundantemente*. (Pereira 1697)
(19) *Estreitamente*. Anguste. arcte. (Cardoso 1569)
(20) Conjunctissime, adv. superl. Muito junta, & *estreitamente*, muito amigavelmente. (Pereira 1697)
(21) Prolixe, Adv. *Larga, longa, liberalmente*. (Pereira 1697)
(22) Pleniter. Adv. Plena, cheia, copioza, perfeita, *compridamente*. (Pereira 1697)
(23) Perlonge, Adv. Mui longa, & *compridamente*; muito longe. (Pereira 1697)
(24) *Curtamente*. Timide. (Pereira 1697)

CRPC returns more than 8.000 matches for *altamente* and two conclusions become self-evident: it is always a modifier of an adjectiv,e and it is always a quantifier adverb, equivalent to *muito* 'very'.

(25) *situação* ***altamente*** *benéfica* 'highly beneficial situation' (*CRPC*)
(26) *factores* ***altamente*** *estimulantes* 'highly stimulating factors'(*CRPC*)

In fact, most *–mente* adverbs that occur as adjective modifiers has exactely the same meaning as *altamente*, which demonstrates that the meaning of the adverb is not related to the meaning of the base adjective:

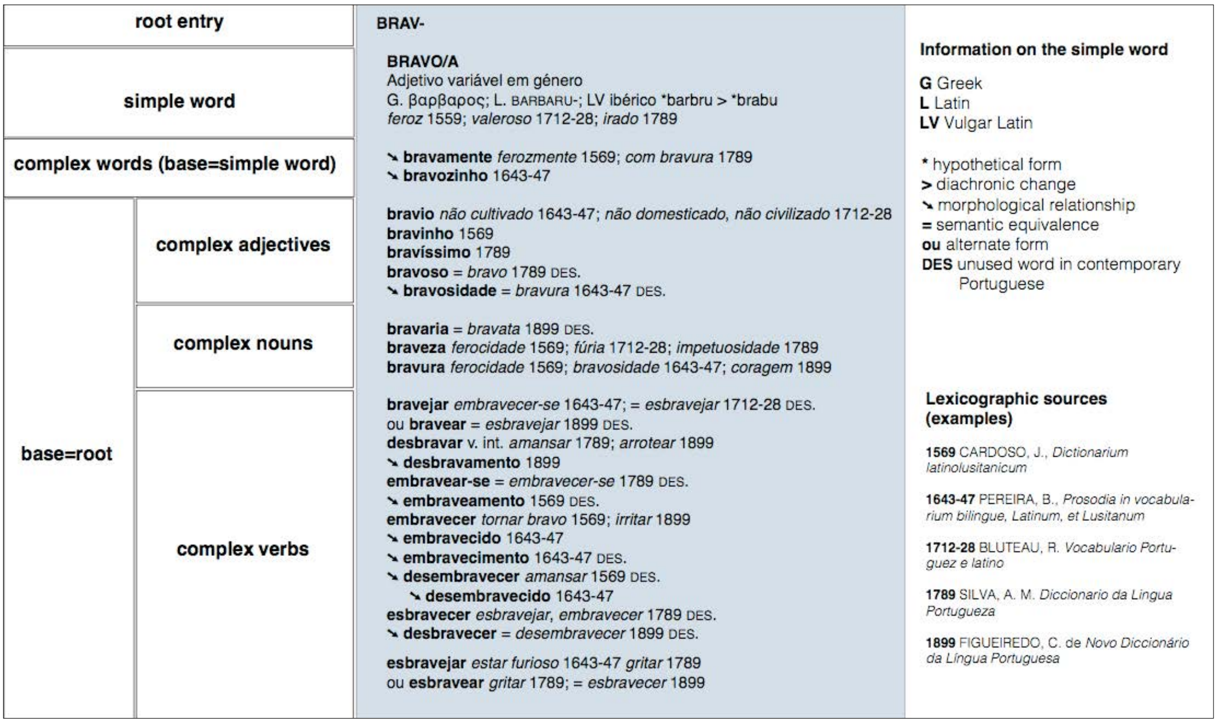
(27) *a praia continua* ***abençoadamente*** *vazia* 'the beach continues blessedly empty' (*CRPC*)
(28) *convite tão* ***abertamente*** *amoroso* 'such openly loving invitation' (*CRPC*)
(29) *coisas* ***abissalmente*** *diferentes* 'abysmally different things' (*CRPC*)

[6] Scalise (1990) presents the eaemple of aridamente that is related to the meaning 'boring' and not to the meaning 'dry' of the adjective arido.

3.5 Entry structure: the case of *bravo*

The structure of articles aims to gather information about the roots and derived words, revealing the diachronic sequence and semantic relations. The lexical research that precedes the compilation of an article can be exemplified in previous works about the semantic evolution of the adjective *bravo* (Villalva, Silvestre 2011) and about adjectives that have undergone a severe meaning shift, like *esquisito* (Silvestre, Villalva, in print). In this paper, we propose a prototype entry, applying the data collected on *bravo* (see table 8).

The entry headword is the root. The first category is the information on the simple word, indicating word class, etymology and a summary of documented semantic evolution. Complex words formed directly from the base (such as -mente adverbs) have a separate description. The other groups are the complex adjectives, verbs and names. For each of the words identified, we provide the date of first attestation in the selected dictionaries, as well as the semantic equivalence. Based on the occurrence in lexical corpora (especially CRPC), we finally evaluated the word frequency. The result is an assessment on the contemporary use of words (unused forms are explicitly marked), with which we intend to contribute to the review of dictionary wordlists and to improve word formation descriptions.



root entry		**BRAV-**	**Information on the simple word** **G** Greek **L** Latin **LV** Vulgar Latin * hypothetical form **>** diachronic change ↘ morphological relationship = semantic equivalence **ou** alternate form **DES** unused word in contemporary Portuguese **Lexicographic sources (examples)** **1569** CARDOSO, J., *Dictionarium latinolusitanicum* **1643-47** PEREIRA, B., *Prosodia in vocabularium bilingue, Latinum, et Lusitanum* **1712-28** BLUTEAU, R. *Vocabulario Portuguez e latino* **1789** SILVA, A. M. *Diccionario da Lingua Portugueza* **1899** FIGUEIREDO, C. de *Novo Diccionário da Língua Portuguesa*
simple word		**BRAVO/A** Adjetivo variável em género G. βαρβαρος; L. BARBARU-; LV ibérico *barbru > *brabu *feroz* 1559; *valeroso* 1712-28; *irado* 1789	
complex words (base=simple word)		↘ **bravamente** *ferozmente* 1569; *com bravura* 1789 ↘ **bravozinho** 1643-47	
base=root	**complex adjectives**	**bravio** *não cultivado* 1643-47; *não domesticado, não civilizado* 1712-28 **bravinho** 1569 **bravíssimo** 1789 **bravoso** = *bravo* 1789 DES. ↘ **bravosidade** = *bravura* 1643-47 DES.	
	complex nouns	**bravaria** = *bravata* 1899 DES. **braveza** *ferocidade* 1569; *fúria* 1712-28; *impetuosidade* 1789 **bravura** *ferocidade* 1569; *bravosidade* 1643-47; *coragem* 1899	
	complex verbs	**bravejar** *embravecer-se* 1643-47; = *esbravejar* 1712-28 DES. ou **bravear** = *esbravejar* 1899 DES. **desbravar** v. int. *amansar* 1789; *arrotear* 1899 ↘ **desbravamento** 1899 **embravear-se** = *embravecer-se* 1789 DES. ↘ **embraveamento** 1569 DES. **embravecer** *tornar bravo* 1569; *irritar* 1899 ↘ **embravecido** 1643-47 ↘ **embravecimento** 1643-47 DES. ↘ **desembravecer** *amansar* 1569 DES. ↘ **desembravecido** 1643-47 **esbravecer** *esbravejar, embravecer* 1789 DES. ↘ **desbravecer** = *desembravecer* 1899 DES. **esbravejar** *estar furioso* 1643-47 *gritar* 1789 ou **esbravear** *gritar* 1789; = *esbravecer* 1899	

Table 8: Root brav - Prototype entry.

4 References

Bluteau, R. (1712-1728). *Vocabulario portuguez e latino*. Coimbra-Lisboa, Collegio das Artes da Companhia de Jesu

Cardoso, J. (1569-1570). Dictionarium latinolusitanicum & vice versa lusitanico latinum. Conimbricae, Joan. Barrerius

Corominas, J., Pascual, J. A. (1981).*Diccionario crítico etimológico castellano e histórico*. Madrid, Gredos.

Corpus de Referência do Português Contemporâneo. Accessed at: http://www.clul.ul.pt/pt/recursos/183-crpc#cqp

Corpus do português: 45 million words, 1300s-1900s. Accessed at: http://www.corpusdoportugues.org [10/03/2014].

Corpus Lexicográfico do Português. Accessed at: http://clp.dlc.ua.pt [10/03/2014].

Figueiredo, C. de (1913). *Novo diccionário da língua portuguesa*. Porto, Typ. da Empr. Litter. e Typographyca

Goes, C. (1921). *Dicionário de raízes e cognatos da língua portuguesa*. Belo Horizonte, Paulo Azevedo & Cia.

Heckler, E., Back, S., Massing, E.R. (1984-1988). *Dicionário morfológico da língua portuguesa*. São Leopoldo, Unisinos.

Infopédia. Dicionário da Língua Portuguesa da Porto Editora. Accessed at: http://www.infopedia.pt/lingua-portuguesa/ [10/03/2014].

[10/03/2014].

Le Trésor de la langue française informatisé. Accessed at: http://atilf.atilf.fr/tlf.htm [10/03/2014].

Nuevo tesoro lexicográfico de la lengua española. Accessed at: http://ntlle.rae.es/ntlle/SrvltGUILoginNtlle [10/03/2014].

Pereira, B. (1697). Prosodia in vocabularium bilingue, latinum et lusitanum. Eborae, Typographia Academiae

Raskin, V., Nirenburg, S. (1995). *Lexical Semantics of Adjectives: A Microtheory Of Adjectival Meaning*. Memoranda in Computer and Cognitive Science MCCS-95-288. Las Cruces - New Mexico, New Mexico State University.

Scalise, S. (1990) Constraints on the Italian suffix –mente. In W. U. Dressler (ed.), *Contemporary Morphology*. Berlim, Mouton de Gruyter.

Silva, A. M. (1789). *Diccionario da lingua portugueza.* Lisboa, na Of. de Simão Thaddeo Ferreira

Silvestre, J. P., Villalva, A. (in print). Mutations lexicales romanes: esquisito, bizarro et comprido. In *InnTrans: Innsbrucker Beiträge zu Sprache, Kultur und Translation*. Peter Lang Verlag.

Villalva, A., Silvestre, J. P. (2011). De bravo a brabo e de volta a bravo: evolução semântica, análise morfológica e tratamento lexicográfico de uma família de palavras. In *ReVEL* v. 9, n. 17. Accessed at: http://www.revel.inf.br/files/artigos/revel_17_de_bravo_a_brabo.pdf [10/03/2014].

Villalva, A., Silvestre, J. P. (in print). *Introdução ao estudo do léxico. Descrição e análise do Português.* Petrópolis-Rio de Janeiro, Vozes.

Lexicological Issues of Lexicographical Relevance

What can Lexicography Gain from Studies of Loanword Perception and Adaptation?

Mirosław Bańko, Milena Hebal-Jezierska
Institute of Polish Language, Institute of West and South Slavic Studies, University of Warsaw
m.banko@uw.edu.pl, milena.hebal-jezierska@uw.edu.pl

Abstract

In normative dictionaries and usage guides, many loanwords are dismissed as 'unnecessary' on the grounds that they have a native synonym, which should be favoured instead. If there is no native equivalent of a loanword in the recipient language, one is often invented in order to eradicate the unwanted loan. However, more detailed studies, in particular, those referred to in this paper, suggest that there is no such thing as fully equivalent words: even if two words have the same designative meaning, they differ in other respects, which makes them mutually unexchangeable except for contexts where the difference between them is inessential.

The goal of this paper is to demonstrate that the meaning potential of lexical loans is different from that of their native synonyms, just because their form is different and differently perceived by language users. The different perception of loanwords can in turn affect their semantic development, thus causing a loanword and its native synonym to diverge. The authors of normative dictionaries and language guides should, therefore, give more consideration to lexical borrowings before they condemn them as 'unnecessary' or 'snobbish'.

Keywords: loanwords; purism; synonymy; variance

1 Introduction

Lexical loans can be divided into two groups: those which are motivated by nominative needs, i.e. the necessity to name a new object or a new phenomenon of foreign origin, and those which appear for expressive needs, because they seem to introduce certain stylistic values, emotional overtones, etc. Polish *smartfon* and Russian *смартфон*, both coming from English *smartphone*, are examples of the former group, while the English interjection *wow*, now used in many other languages, can exemplify the latter group. There are obviously more ways to fill lexical gaps, as the French alternative names for smartphone – *ordiphone* and *téléphone intelligent* – show, but we will not be dealing with them here. Instead, we will restrict ourselves to borrowings proper, i.e. words taken from a donor language with possible alterations in their pronunciation, spelling, morphological and syntactic features, sometimes also in their meaning.

By definition, loanwords borrowed for nominative needs have no commonly known synonyms in the recipient language, not at least at the moment they enter it. Loanwords adopted for expressive needs, again by definition, do have some synonymous words and that puts them in a disadvantageous position whenever purist attitudes or specific concern about the 'economy' of language come into play. Many such loanwords are dismissed as 'unnecessary', criticized as 'overused', pointed out as examples of bad taste and snobbery. Negative assessments of them are expressed in the popular press, scholarly literature, academic textbooks, language guides and dictionaries alike.

It is not our intention to question the efforts of normatively oriented linguists and lexicographers who aim at reducing the number of loanwords in a language. Critical assessment of word borrowing – and of particular borrowings – is something needed, if only because there are readers waiting for such criticism. Dictionaries have to account not only for how words are really used, but also for what the language users think about their correct use. On the other hand, lexicographers should be aware that cases of full equivalence between a loanword and its native synonym are practically non-existent. Even if two words have the same designative meaning, they differ in other respects and evoke different associations in the minds of the language users. Over the years, such associations may stabilize and become part of the designative meaning, thus making the originally synonymous words diverge.

The goal of this paper is not so much to remind us of these relatively simple truths, as to demonstrate that the adaptation of loanwords in the recipient language is guided, at least to a certain extent, by the interplay between their form and meaning. A tendency can be observed to maintain harmony between the form and meaning of loans, which can manifest itself, *inter alia*, in how the original spelling of a loan is assimilated in the recipient language and how its meaning is shaped in relation to its native synonyms. The tendency will be illustrated with w number of examples later on in the paper. Let us begin, however, with examples of purist attitudes from Polish, German and Czech lexicography and linguistics.

2 Examples of purist attitudes

In a textbook for students of Polish language and literature, Andrzej Markowski (2005), a prominent linguist, chairman of the Council of the Polish Language, gives long tables of 'overused words' and 'vogue words', usually of foreign origin, and demonstrates, by means of invented examples, how they can be replaced by other words, most of them native or borrowed so long ago that their foreign origin is no longer recognizable. The same or similar loanwords were reviled earlier in a standard dictionary of Polish usage, edited by the same author (Markowski 1999), in some of his other books, and in many popular dictionaries and usage guides, compiled by others. It is worth stressing that Markowski's position is far from extreme purism. His judgments are of a 'better/worse', not 'yes/no' type, yet his decisions are clearly not based on detailed analyses of the meaning and use of the particular words he paired. Had he looked at them more carefully, he would have found distinctions which make the words mutually unexchangeable, except for contexts where the difference between them is inessential.

The attitude to loanwords varies depending on the political and sociolinguistic situation, as well as the normative tradition in particular countries. In Germany, around 300 dictionaries of loanwords were published between 1801 and 1945, around half of them belonging to the class known as *Verdeutschungswörterbücher*, literally 'Germanizing dictionaries'. They were not intended to explain the meaning and illustrate the use of borrowed words, but rather to demonstrate how these could be replaced by native words, some of them specifically invented for this purpose (Lipczuk 2007, 2011). No doubt the regional disintegration of Germany before 1871 favoured the development of national purism, but purist attitudes developed within German society even after the unification of the country, because the rising power of the state created favourable conditions for German nationalism. Also the romantic tradition of treating the language as the embodiment of the spirit of a nation caused many Germans to believe that loanwords posed a threat not only to their language, but also to their national identity.

The example of Germany, where even international words became the object of purifying actions (cf. *Rundfunk* and *Fernsprecher*, coined to replace *Radio* and *Telephon*, respectively), is an extreme one, but similar 'nativizing' dictionaries are known from the lexicographic tradition of many countries. In Poland, which from 1795 to 1918 was partitioned among Russia, Prussia and Austria, the concern about the language was steadily expressed at the time and took on different forms. At one end of the scale was Linde's (1807-1814) six-volume dictionary of the Polish language, based on citations from about 800 sources, an attempt to save the treasures of the language and help the nation to survive the difficult time (Adamska-Sałaciak 2001). At the other end there were a number of much smaller dictionaries and usage guides whose aim and content made them similar to German *Verdeutschungswörterbücher*. Among them, Kortowicz (1891) is a good example, see Leszczyński (2000) and Czesak (2007) for information about his dictionary.

In the history of the Czech language, purist attitudes appeared in the times of Jan Hus and have been present continuously thereafter, up to the present day (Engelhardt 2001: 235). The purist trends were particularly strong at the end of the 19th century and between 1920s and 1940s. At the end of the 19th century purists tried to eliminate words of foreign origin, particularly Latin and Greek internationalisms, as well as German and French loans. A number of neologisms were formed on the basis of native words, but the new coinages often replicated the structure of the words they were supposed to replace (Engelhardt 2001:237). For instance, the Czech lexical innovation *pololetí*, patterned on the German word *Halbjahr* (literally 'half year'), was invented to substitute for the Czech internationalism *semestr* (cf. Latin *semestris* 'six-monthly'). As was often the case, the substitution failed, with both *semestr* and *pololetí* being used in present-day Czech.

At the beginning of the 20th century the purist tendencies in Czech linguistics became even stronger. Linguists aimed to eliminate not only proper loans, but also lexical and syntactic calques. Many usage guides warning language users against loans of different kinds were published. Purists were particularly eager to identify German loans everywhere, even in native constructions (Král 1917), and they formed bizarre neologisms to replace them. Many of the new coinages had a short life, but some have survived to the present. A good example is the word *rozhlas*, which was introduced with the intention to replace *broadcasting* and *radio* (from the verb *hlásit* 'report' and the suffix *roz-*, denoting the spread of something from one place).

After the communists came to power in the Czech state, especially after 1948, English loans were fought most vigorously and replaced with native words or their spelling was changed to conceal their western origin. Many native neologisms were created at that time, e.g. *silostroj* (a compound word of the meaning 'power and machine') was introduced to replace *motor*, and *samohyb* (another compound combining the meanings of 'itself' and 'move') was intended to take the place of *auto*, cf. Svobodová (2009: 33). Nowadays the tendencies to purify the Czech language are not so strong, but protective attitudes can still be observed, because some linguists are afraid that the increasing presence of foreign words poses a threat to the Czech language.

3 Why are native equivalents never fully equivalent to lexical loans?

Many linguists, philosophers and literary historians have claimed that no two words can be fully equivalent with respect to their linguistic function. One can find the same opinion among lexicographers, cf. the often-quoted passage from Urdang's introduction to *The Synonym Finder*:

> Those who work with language know that there is no such thing as a true 'synonym'. (...) Even though the meanings of words may be the same – or nearly the same – there are three characteristics of words that almost never coincide: frequency, distribution, and connotation. (Urdang 1978)

Ullmann takes a less extreme position on this point and explains why cases of absolute synonymy are very rare:

> (...) it is perfectly true that absolute synonymy runs counter to our whole way of looking at language. When we see different words we instinctively assume that there must also be some difference in meaning, and in the vast majority of cases there is in fact a distinction even though it may be difficult to formulate. Very few words are completely synonymous in the sense of being interchangeable in any context without the slightest alteration in objective meaning, feeling-tone or evocative value. (Ullmann 1964: 142)

As 'completely synonymous' he mentions technical terms, e.g. *caecitis* and *typhlitis* can both be used with reference to the inflammation of the blind gut. However, even such names differ with respect to non-designative features, e.g. they evoke different associations in the minds of the language users.

One often hears that language does not tolerate fully equivalent words and differentiates them, thus eliminating cases of full equivalence. The tendency to avoid redundant means of expression is said to be evidence that language is governed by laws of economy (Nagórko 2004: vii). However, such propositions do not explain the inner mechanism of linguistic economy and, in particular, they do not explain why there should be a difference in meaning, broadly understood, between loanwords and their native synonyms. Our position is that this has something to do with the word form itself. The unfamiliar forms of lexical loans are perceived differently from the familiar forms of their native synonyms and the difference in perception may result in different semantic development of such words.

For example, shortly after *kurort*, a 19th-century borrowing from German, appeared in Polish, a native term *uzdrowisko* (literally 'health resort') was coined with the intention to relegate the unwanted loan from the language. However, this effort failed to have the desired effect: instead of disappearing, *kurort* changed its meaning to 'popular and snobbish holiday place'. The change was very likely directed by the connotations of the word: some pre-war dictionaries (e.g. *Słownik wyrazów obcych* of 1921) informed that *kurort* was most often used with reference to health resorts in Germany and the preference for foreign places in its use is still visible in modern texts. Furthermore, the collocation image of *kurort* includes such features as exclusiveness (strangely enough, not in conflict with popularity), modernity, reputation and elegance, whereas in the collocation image of *uzdrowisko* it is tradition and aesthetic values that are best seen (see Bańko 2013a for a more detailed analysis of both words).

The influence of word form on word meaning can be studied in texts and other cultural artifacts (e.g. by collocation analysis or Google image inspection), but it can also be brought to light by means of experiments. The results published by Song and Schwarz (2010) are worth quoting here. They experimented with nonce words, some of them familiar in shape, some strange, and observed a correlation between the familiarity of a word and its perception. For instance, fictitious food additives with names difficult to pronounce were evaluated as more harmful than food additives with easy names. Similarly, roller-coaster rides in a fictitious amusement park were judged as more risky and more exciting when their names were strange and difficult. Song and Schwarz explain the effect with a mistaken projection of the difficulties the subjects experienced in processing the unfamiliar words onto the referents of the words: unaware of the source of difficulty, the subjects attributed it to the referents, judging them as more risky, more dangerous, more harmful, etc. (for a critical review of these studies see Rączaszek 2013).

It would be premature to claim that Song and Schwarz's findings can explain all the distinctions observed between a synchronic loan and its native synonyms. More experiments are needed and they should be done on real language data, not on invented words. We will next briefly describe a project designed to perform more systematic research in this regard, using both linguistic and psycholinguistic methods.

4 About the APPROVAL project

The aim of the APPROVAL project is to search for various factors bearing on the psychological perception, social reception and linguistic adaptation of loanwords.[1] Among the possible factors, the relation between word meaning and word form is of particular importance, because we assume that the form of a word is not irrelevant (contrary to the widely accepted view of the arbitrary nature of linguistic signs, a foundation stone of Saussurian linguistics). The form of a word can influence its meaning (cf. *kurort* above), but also the meaning of a word can affect its form, e.g., by hindering the process of a loanword's adaptation (the word *jazz* can be a case in point: though borrowed to Polish and Czech almost a hundred years ago, it still appears mainly in its original spelling in both languages, very likely because the foreign spelling reflects better the symbolic values associated with jazz music in Poland and the Czech Republic, see Bańko and Hebal-Jezierska 2012 for details).

We also assume that fully equivalent words do not exist, so we focus on comparative analysis of word pairs (sometimes triples, quadruples, etc.) in which one element is of foreign origin, the other native, or in which one element has the original spelling, while the other is graphically adapted to the recipient language. Fifty Polish word pairs have been subjected to in-depth analysis, based on language corpora and other data, not excluding the evidence in the language itself (e.g. we treat the frequency of a word as indicative of its importance and we pay attention to secondary uses of a word, its derivatives and idiomatic expressions, because such data reveal some of the typical associations the word calls up in the minds of its users and help to draw the stereotypical image of its referent). As this paper is being prepared, most of the 50 synonym and variant pairs have been already inspected and the results are available on the project website, see http://www.approval.uw.edu.pl/en_GB/start pl.

In order to make the observations more credible, the same research is being done on corresponding Czech word pairs which serve as a control group, e.g. the Polish pair *absurdalny – niedorzeczny* 'absurd, nonsensical' corresponds to the Czech pair *absurdní – nesmyslný*, in which the first element comes from the same root as the first element of the Polish pair. By composing the research material this way, it became possible to study the adaptation processes in two cognate languages on the basis of comparable examples, using the same methodology, the same kind of data, and even the same description format.

In total, 100 pairs in two languages will have been analyzed by the end of the project, using corpora, as well as dictionaries, web archives, digital libraries, library catalogues, Google images and other sources of language relevant data (e.g., library catalogues are being used to check frequencies of words in book titles). In addition, psycholinguistic experiments are being carried out on the Polish material to enrich and verify the observations based on language corpora and other textual and non-textual sources by means of linguistic methods (see the project website for details).

[1] The name of the project comes from 'Adaptation, Psychological Perception and Reception of Verbal Loans'. It is also meant as a reminder that in normal circumstances loanwords do not pose a threat to a language; to the contrary, they add to its wealth.

The results gained so far are encouraging and they largely support the hypothesis of there being a relationship between form and meaning in the adaptation of lexical loans. For limits of space, a few examples from Polish will have to suffice here. We will focus on selected details, with no intention to account for a full analysis of any of the words mentioned below.

4.1 strofa and zwrotka

The Polish words *strofa* and *zwrotka* both mean 'stanza', but in technical literature usually the former word is used, while in the general language the latter one is more common. The likely reason is not only that *strofa* is of Greek origin (borrowed via Latin), but also that *zwrotka* contains a familiar suffix *-k-* which in many other nouns (though not in *zwrotka*) has a diminutive function.

The difference between these two words was first observed in corpus analysis, especially in their collocation images. For example, *zwrotka*, but not *strofa*, is used in reference to popular songs and children's poems, *strofa* can be the subject of aesthetic evaluation (cf. *piękne strofy* 'beautiful stanzas') and artistic activity (cf. *pisać, układać strofy* 'write, arrange stanzas'), while *zwrotka* is less frequent in such contexts. In addition, *strofa* can be recited, but *zwrotka* is sung. Only *zwrotka* collocates with the word *refren* ('refrain'), which confirms its connection to songs.

A study of free associations, based on Osgood's semantic differential, was next carried out. Twenty-two subjects took part in it, each asked to mention up to three associations for one word, so the maximum number of associations for a word was 66. Among the associations noted more than once, *refren* 'refrain', *rymy* 'rhymes', *śpiewanie* 'singing', *muzyka* 'music' and *ognisko* 'camp-fire' were given only for *zwrotka*, while *szkoła* 'school', *poezja* 'poetry', *literatura* 'literature' and *Mickiewicz* (the best known Polish poet) were given only for *strofa*. In addition, though *piosenka* 'song' and *wiersz* 'poem' were mentioned for both words, *piosenka* had a frequency of 19 with *zwrotka* and 3 with *strofa*, while *wiersz* appeared 16 times with *strofa* and only 3 times with *zwrotka*. As can be seen, the results of corpus analysis are in line with the study of free associations.

4.2 helikopter and śmigłowiec

Though *helicopter* and *śmigłowiec* both mean 'helicopter' in Polish, their stylistic distribution is different. *Helicopter* is common in spoken language and in many other language varieties, while *śmigłowiec* tends to be used in technical literature. This is probably the reason why in the domain *lego.com/pl-pl*, belonging to the producers of Lego bricks, the Google search engine finds far more occurrences of *helikopter* than *śmigłowiec*. As far as book titles are concerned, *helikopter* can be found on the covers of children's stories, while *śmigłowiec* appears in the titles of books on aeronautical technology and military science. Among the Google images indexed with the word *helikopter*, toys and miniature models are more frequent than among images indexed with the word *śmigłowiec*.

However, a more interesting and more surprising observation about *helikopter* and *śmigłowiec* can be made when comparing their relative frequencies in certain contexts. Though on Polish-language websites *helikopter* is several times more frequent than *śmigłowiec*, the quantitative advantage of *mały helikopter* 'small helicopter' and *szybki helikopter* 'fast helicopter' over *mały śmigłowiec* and *szybki śmigłowiec*, respectively, is significantly lower. Moreover, *lekki helikopter* 'light helicopter' is less frequent than *lekki śmigłowiec*. Apparently, small, light and fast machines of this type are more often referred to with the word *śmigłowiec* than its relative frequency to the word *helikopter* would suggest. This may be due to the fact that the name *śmigłowiec* is related to the words *śmigło* 'propeller', *śmigły* 'swift' and *śmigać* 'move quickly, zip (around)'.

However, the overall picture is not quite clear yet, partly because the relative frequencies of *helikopter* and *śmigłowiec* in two reference corpora of Polish – Narodowy Korpus Języka Polskiego and Korpus Języka Polskiego PWN – are opposite to those found on the Internet and partly because the study of free associations has yielded different results for these two words than obtained in corpus analysis. Further research into the psychological perception of *helikopter* and *śmigłowiec* is planned within the APPROVAL project with the intention to confirm or refute the conjectures made on the basis of corpus data. Whatever the results of the research, there is no doubt a difference between *helikopter* and *śmigłowiec* in their semantic content, if only non-designative components of the word meaning are allowed.

4.3 eksplozja and wybuch, kuriozalny and osobliwy

Eksplozja 'explosion' and *wybuch* 'explosion, outbreak, outburst' may refer to the same kind of events, but the phenomena referred to by the former word are perceived as stronger and more violent. The difference is so distinct that it has even been noted in the definitions for these two words in some dictionaries. There are more synonym pairs in which the referents of a loanword seem larger and more powerful than the referents of its native synonym, cf. *dewastować* and *niszczyć*, both meaning 'destroy'. Here the first element, cognate with English *devastate*, denotes a purposeful or mindless activity, especially against public property or natural environment.

However, sometimes the difference between a loanword and its native synonym lies not in the referents themselves, but in the way they are talked about, e.g. in the values conveyed. The adjective *kuriozalny* 'peculiar, bizarre', related to Latin *curiosum*, is half as frequent in Polish as its native counterpart *osobliwy* 'peculiar', but in parliamentary reports the former word prevails overwhelmingly. A closer inspection shows that Polish MPs need it to criticize their political opponents, e.g. *Pana poglądy są kuriozalne, panie pośle* 'Your views are bizarre, Mr. X'.

The tendency to use difficult and erudite words for hyperbolic effects, whether to make a phenomenon look more powerful or just to convey negative attitudes, can be well explained in the context of Song and Schwarz's (2010) experiments discussed above.

4.4 dealer and diler

The last of our examples is different from the previous ones, because it is not concerned with a synonym pair. It deals with a pair of spelling variants of the same word, a recently new Polish borrowing from English. However, the situation is much the same as before, because one element of the pair is foreign while the other one is 'nativised' (rather than native), and the two elements therefore exhibit the same 'unfamiliar – familiar' opposition as in the case of synchronic loans and their native synonyms. Thus, the necessary conditions are met for different associations to evolve around the different words.

As all variants, *dealer* and *diler* have the same designative meaning, but their stylistic distribution and the areas of their application are not identical. In the press, *dealer* is almost twice as frequent as *diler* and used mainly with reference to car vendors, whereas *diler* has equal frequency in automobile and drug-related contexts. In literary texts, on the other hand, *diler* is twice as frequent as *dealer* and applied usually to drug sellers, whereas *dealer* is equally often used in automobile contexts. Apparently, the foreign variant is more prestigious and better suited to name the job of authorized vendors in car showrooms; the domestic variant, on the other hand, is unpretentious and corresponds well with the dubious job of drug peddling. The difference can well be observed in the Google image galleries, too, which once more confirms the usefulness of Google images in linguistic analysis (Bańko 2013b).

5 Conclusions for practical lexicography

Our intention was not to question the need for normative assessments in dictionaries, nor to claim that cumulative synonym dictionaries have no raison d'être. The conclusions from our work are relevant to the theoretical foundations of lexicography, but also to lexicographic practice. It is important for lexicographers to be aware that distinctions often may lie where similarity seemingly prevails (in a way, all of language is based on distinctions, and here we are in complete agreement with de Saussure). Furthermore, it is important to realise that in the adaptation of lexical loans, form and meaning are interdependent: the form of a word can affect its meaning, but it can also be can influenced by it. In many cases, a tendency to maintain harmony between a word form and word meaning can be observed in the process of loanword adaptation.

Better recognition of differences between near synonyms is essential for both monolingual and bilingual lexicography. Adequate definitions, especially in production dictionaries, should explain differences between synonymous words. Adequate translation equivalents are dependent, among other things, on how well near synonyms of the source language and the target language are discriminated. This is not to say that each dictionary ought to be equally specific in its treatment of word meanings. For example, decoding dictionaries need not focus so much on distinctions between words as may be expected from encoding dictionaries. Lexicographers should consider for themselves to what extent the observations made in this paper may be useful in their work. In any event, more caution is advised before assessing a loanword as unnecessary and more thoroughness is needed in how new words borrowed from other languages are treated. It is not enough to blame those who use them of snobbery.

A separate question, not to be dealt with here, is how our findings can be incorporated in what are called distinctive synonym dictionaries, which are intended to account for differences among synonyms rather than to gather as many words close in meaning as possible (cf. *Dystynktywny słownik synonimów* by Nagórko et al. 2004 as an example of works of this type). Another area where it is more important to show differences between words than to identify similarities is the so called synonym discussions, known from some dictionaries (cf. special paragraphs, headed *Synonyms*, in *The American Heritage Dictionary*).

6 References

Adamska-Sałaciak, A. (2001). Linde's Dictionary: A landmark in Polish lexicography. In *Historiographia Linguistica*, 28(1-2), pp. 65-83.

American Heritage Dictionary. Second College Edition. Boston: Houghton Mifflin Company, 1982.

Bańko, M. (2013a). Normatywista na rozdrożu. Dwugłos w sprawie tzw. kryterium narodowego. In J. Migdał, A. Piotrowska-Wojaczyk (eds.) *Cum reverentia, gratia, amicitia... Księga jubileuszowa dedykowana Profesorowi Bogdanowi Walczakowi*, vol. 1. Poznań: Wydawnictwo Rys, pp. 141-148.

Bańko, M. (2013b). Obrazy Google jako źródło informacji lingwistycznej. In. W. Chlebda (ed.) *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole: Wydawnictwo Uniwersytetu Opolskiego, pp. 73-84.

Bańko, M. & Hebal-Jezierska, M. (2012). Proč *jazz*, nikoliv *džez*? Harmonie grafické podoby lexému a obsahu – jako jeden z činitelů ovlivňujících adaptaci cizojazyčných přejímek? In S. Čmejrková, J. Hoffmannová, J. Klímová (eds.) *Čeština v pohledu synchronním a diachronním*. Praha: Karolinum, pp. 371-375.

Czesak, A. (2007). *Oczyszciciel mowy polskiej* E. S. Kortowicza, Poznań 1891 – idee i zawartość. In J. Kamper-Warejko, I. Kaproń-Charzyńska (eds.) *Z zagadnień leksykologii i leksykografii języków słowiańskich*. Toruń: UMK, pp. 79-85.

Engelhardt, G. (2001). Český a německý purismus z konce 19. století. In *Naše řeč* 84(5), pp. 235-242.

Korpus Języka Polskiego PWN [PWN Corpus of Polish]. Online at http://korpus.pwn.pl.

Kortowicz, E. S. (1891). Oczyszciciel mowy polskiej, czyli Słownik obcosłów, składający się z blisko 10,000 wyrazów i wyrażeń z obcych mów utworzonych a w piśmie i w mowie polskiéj niepotrzebnie używanych, oraz z wyrazów gminnych, przestarzałych i ziemszczyn w różnych okolicach Polski używanych z wysłowieniem i objaśnieniem polskiem. Poznań: czcionkami drukarni „Dziennika Poznańskiego".

Král, J. (1917). Naše brusy I. In *Naše řeč*, 1(4). Online at http://nase-rec.ujc.cas.cz/archiv.php?art=64 [05/04/2014]

Leszczyński, Z. (2000). Krótka relacja o puryście sprzed wieku. In *Prace Filologiczne*, XLV, pp. 347-352.

Linde, S. B. (1807-1814). *Słownik języka polskiego*, 6 vols. Warszawa.

Lipczuk, R. (2007). Geschichte und Gegenwart des Fremdwortpurismus in Deutschland und Polen. Frankfurt am Main: Peter Lang.

Lipczuk, R. (2011). O słownikach wyrazów obcych, słownikach zniemczających i spolszczających. In B. Afeltowicz, J. Ignatowicz-Skowrońska, P. Wojdak (eds.) *In silva verborum. Prace dedykowane Profesor Ewie Pajewskiej z okazji 300-lecia pracy zawodowej*. Szczecin: Volumina.pl, pp. 205-216.

Markowski, A. (1999) (ed.). *Nowy słownik poprawnej polszczyzny*. Warszawa: PWN.

Markowski, A. (2005). Kultura języka polskiego. Teoria. Zagadnienia leksykalne. Warszawa: PWN.

Nagórko, A., Łaziński, M. & Burkhardt, H. (2004). *Dystynktywny słownik synonimów*. Kraków: Universitas.

Narodowy Korpus Języka Polskiego [National Corpus of Polish]. Online at http://nkjp.pl.

Rączaszek, J. (2013). Studying the semantics of loanwords which have near synonyms in the host language: Psycholinguistic and multidimensional corpus representation methods. Accessed at http://www.approval.uw.edu.
pl/en_GB/publikacje [05/04/2014].

Słownik wyrazów obcych. 25.000 wyrazów, wyrażeń, zwrotów i przysłów cudzoziemskich, używanych w mowie potocznej i w prasie polskiej, 9th ed. Warszawa: Wydawnictwo M. Arcta, 1921.

Song, H. & Schwarz, N. (2010). If it's easy to read, it's easy to do, pretty good, and true. In *The Psychologist*, 23(2). Accessed at http://www.thepsychologist.org.uk/archive/archive_home.cfm?volumeID=23&editionID=185&
ArticleID=1629 [05/04/2014].

Svobodová, D. (2009). Aspekty hodnocení cizojazyčných přejímek: mezi módností a standardem. Ostrava: Universitas Ostraviensis.

Ullmann, S. (1964). Semantics. An Introduction to the Science of Meaning. Oxford: Basil Blackwell.

Urdang, L. (1978). Introduction. In J. I. Rodale (ed.) *The Synonym Finder*. Emmaus, Pa.: Rodale Press.

Acknowledgements

This project has been supported by a grant from the National Science Centre, Poland, no DEC-2011/03/B/HS2/02279.

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

Hanno Biber
Austrian Academy of Sciences
hanno.biber@oeaw.ac.at

Abstract

Satirical literary texts have certain properties that are highly interesting for the study of pejorative language use. The language of the satirical journal "Die Fackel" published and almost entirely written by Karl Kraus is the text basis for a text-lexicographic exploration into the field of pejorative language and its specific lexicographic units. The „Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift „Die Fackel". Alphabetisches, Chronologisches, Explikatives" was published in 2008. The three volumes document the usage of invectives in the journal, in alphabetical and in chronological order, and in explicative form explained through the example of the last article of the journal. The alphabetical part consists of 2,775 examples of pejorative phrases and related indices. The chronological part presents 555 of these pejorative phrases arranged in chronological order providing expanded contexts. The third volume contains explicatory texts as well as "Wichtiges von Wichten", the final article of "Die Fackel", where pejorative phrases were marked up and accompanied by commentaries. This source is representing a literary genre that offers a variety of different forms of pejorative language to be studied from various perspectives. The lexicographic insights offered by the text dictionary into the use of pejoration by Karl Kraus will be presented in this paper.

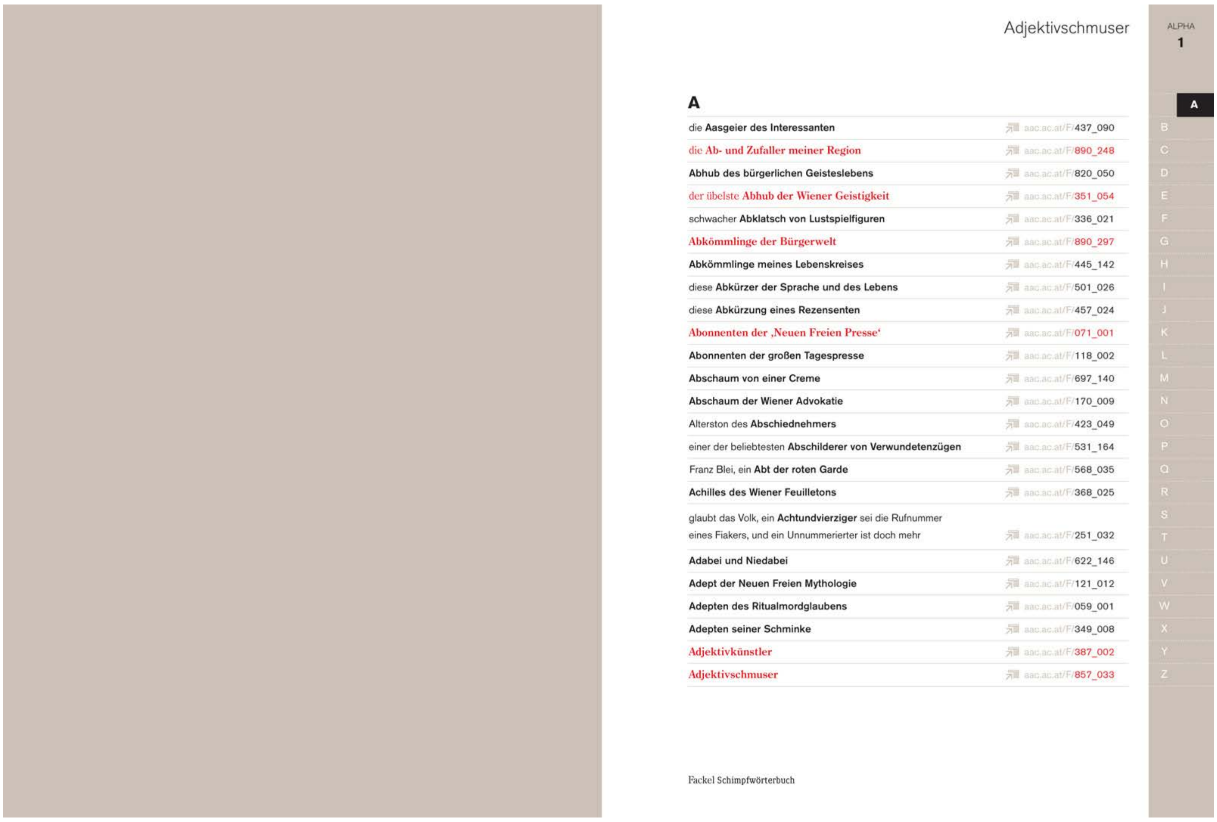
Keywords: text lexicography; literary studies; pejorative language

1 Text Dictionary of "Die Fackel"

"Die Fackel" ("The Torch") is the name of the satirical magazine of 22.586 pages which was published by Karl Kraus in 922 issues in Vienna from 1 April 1899 until February 1936. The work of the satirist, language critic and social critic Karl Kraus, who was born in Bohemia in 1874 and died in Vienna in 1936, is an abundant and highly interesting source not only for the history of his time, but for the language spoken and written at the time and above all for the moral transgressions which the satirist observed by interpreting the words and phrases of his time and which he pointed out in his satirical and polemical texts. His very influential literary journal comprises a great variety of articles, essays, glosses, notes, commentaries, aphorisms, poems, songs, advertisements, and many other literary forms. The main method of his satire is the method of quotation, whereby Karl Kraus wittingly comments upon the quotations he finds in the newspapers, journals and magazines as well as in the literature and in the political speeches of his time. He critizises in numerous satirical and polemical articles of his magazine the acts and the words of his contemporaries who were active in various intellectual fields, not only in the media, but also in the theatre, the university, the church, in politics, the economy, the military, and so on. Karl Kraus covers in his typical style in thousands of texts the themes of journalism and war, of politics and corruption, of literature and lying. The language of the satirical journal "Die Fackel" - for several decades since its start in 1899 almost entirely written by Karl Kraus, and from 1911 on without external contributions, - is the text basis for a unique text lexicographic exploration which has been carried out at the Austrian Academy of Sciences for several years. Nowhere else in German literature, to mention but one aspect, is there such an extensive documentation of the socio-political idiom of the time as there is in "Die Fackel" by Karl Kraus. The idea of compiling a text-dictionary of the "Fackel" derives from our interest in language and how it is used in "Die Fackel". The Fackel-Dictionary is a selective text dictionary research project initiated by Werner Welzig. The original plan was to develop three different types of dictionary. First, a "Dictionary of Idioms", the *Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift 'Die Fackel'* (Welzig 1999), published on the occasion of the 100th anniversary of "Die Fackel" in 1999, a monumental scholarly publication that has won several international and national prizes, among them the Prix Logos by the French association of linguists and the Golden Letter at the Leipzig book fair as the most beautiful book of the year 2000 for its design worked out by the designer Anne Burdick in cooperation with the Fackellex working group (Hanno Biber, Evelyn Breiteneder, Susanne Buchner, Heinrich Kabas, Karlheinz Mörth, Christiane Pabst, Franco Schedl, Adriana Vignazia, Werner Welzig). In order to thoroughly analyse and interpret the more than 9000 idiomatic units of the text dictionary, of which 144 idioms are described in great detail in individual entries, which are longer than common dictionary entries are, it has proved reasonable to search large volumes of texts for comparison, a procedure that has also been made use of for the dictionary that followed. It is a lexicographic necessity to have large text corpora available, in particular when searching for lexical units, for example for idioms or for pejorative phrases, so that these text dictionary projects can be regarded as examples and applications of corpus based textual studies. The second example within this context of text lexicography is the "Dictionary of Insults and Invectives", *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift 'Die Fackel'* (Welzig 2008), which was also worked out by the Fackellex working group (Hanno Biber, Evelyn Breiteneder, Gerald Krieghofer, Karlheinz Mörth) and will be introduced in this short presentation. Third, an "Ideological Dictionary" had been originally planned, which later in the course of the development of the program plans has been decided to be transformed into a special edition of the posthumously published text "Third Walpurgis Night" (*Dritte Walpurgisnacht*) by Karl Kraus, written in May 1933, constituting a manifestation the most important contemporary text of German literature dealing with the early time of National Sozialism and the issue how the intellectuals reacted when this most violent regime came to power.

2 Pejorative Language Use

In this paper a short presentation of the main aspects of the second text dictionary, the "Dictionary of Insults and Invectives", *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift 'Die Fackel'* (Welzig 2008) will be given, offering an exploration into the field of pejorative language use and its specific lexicographic units as selected for this dictionary. "Die Fackel" can be regarded as an ideal text basis for such a dictionary in that it has no equal in the German literature of the twentieth century either in terms of form and content or in the use of language. The *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‚Die Fackel'* (Welzig 2008) published in 2008 consists of three parts: „Alphabetisches", „Chronologisches", „Explikatives". The three volumes document the usage of the invectives and pejorative phrases in the journal, in alphabetical order (ALPHA), in chronological order (CHRONO), and explained through the example of the last article of the journal in a volume (EXPLICA), that shows by thoroughly analysing one short, but semantically very dense text, the full potential of a text lexicographic documentation of its pejorative forms.



Adjektivschmuser

ALPHA
1

A

die **Aasgeier des Interessanten**	aac.ac.at/F/437_090
die **Ab- und Zufaller meiner Region**	aac.ac.at/F/890_248
Abhub des bürgerlichen Geisteslebens	aac.ac.at/F/820_050
der übelste **Abhub der Wiener Geistigkeit**	aac.ac.at/F/351_054
schwacher **Abklatsch von Lustspielfiguren**	aac.ac.at/F/336_021
Abkömmlinge der Bürgerwelt	aac.ac.at/F/890_297
Abkömmlinge meines Lebenskreises	aac.ac.at/F/445_142
diese **Abkürzer der Sprache und des Lebens**	aac.ac.at/F/501_026
diese **Abkürzung eines Rezensenten**	aac.ac.at/F/457_024
Abonnenten der ‚Neuen Freien Presse'	aac.ac.at/F/071_001
Abonnenten der großen Tagespresse	aac.ac.at/F/118_002
Abschaum von einer Creme	aac.ac.at/F/697_140
Abschaum der Wiener Advokatie	aac.ac.at/F/170_009
Alterston des **Abschiednehmers**	aac.ac.at/F/423_049
einer der beliebtesten **Abschilderer von Verwundetenzügen**	aac.ac.at/F/531_164
Franz Blei, ein **Abt der roten Garde**	aac.ac.at/F/568_035
Achilles des Wiener Feuilletons	aac.ac.at/F/368_025
glaubt das Volk, ein **Achtundvierziger** sei die Rufnummer eines Fiakers, und ein Unnummerierter ist doch mehr	aac.ac.at/F/251_032
Adabei und Niedabei	aac.ac.at/F/622_146
Adept der Neuen Freien Mythologie	aac.ac.at/F/121_012
Adepten des Ritualmordglaubens	aac.ac.at/F/059_001
Adepten seiner Schminke	aac.ac.at/F/349_008
Adjektivkünstler	aac.ac.at/F/387_002
Adjektivschmuser	aac.ac.at/F/857_033

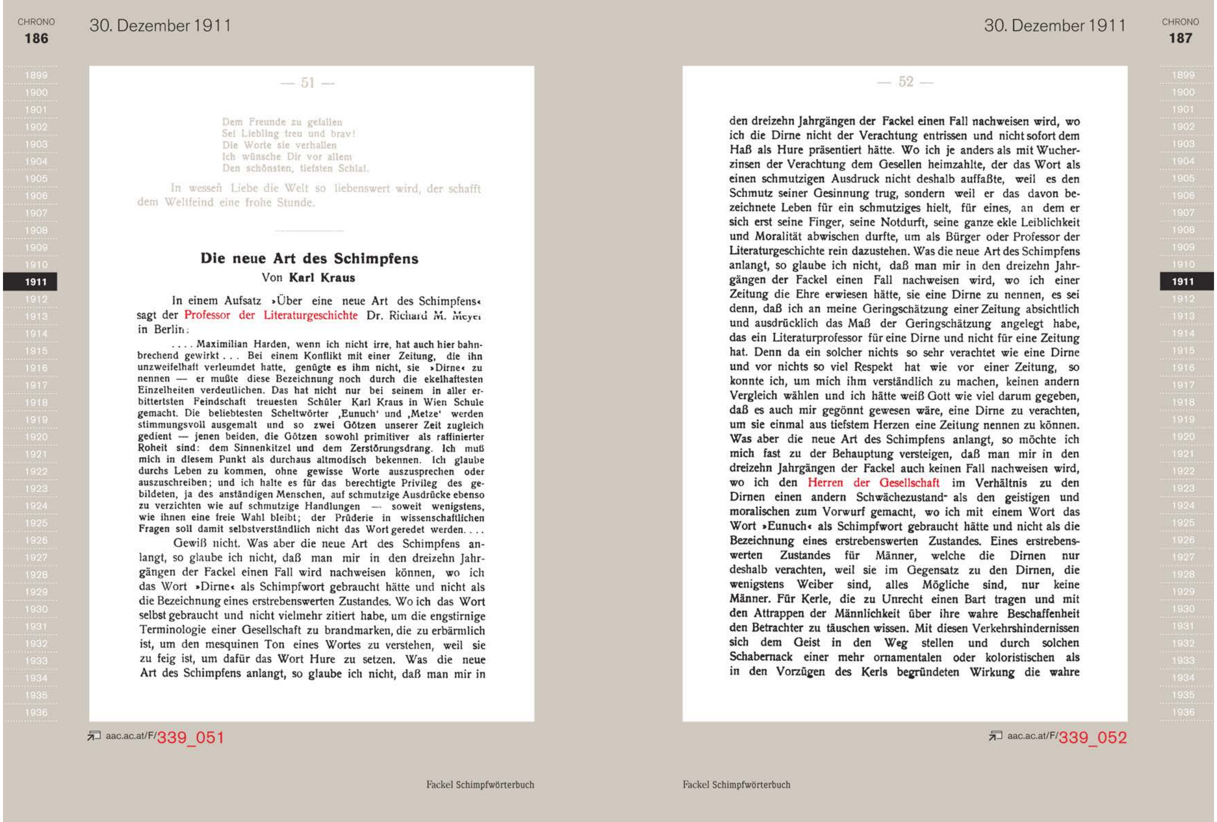
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Fackel Schimpfwörterbuch

Figure 1: ALPHA.

The alphabetical volume of the text dictionary (ALPHA) consists of 2,775 examples of pejorative phrases and several related indices. In this alphabetical part the lemmatized entries of the pejorative forms are marked and each of the entries is given a short context. Only one reference is given with one entry and those entries which are represented in full detail of the page in the chronological volume of the text dictionary are printed in red color. The indices at the end of the alphabetical part, referring to the pages and the lines indicated by the marginal letters, comprise a complete index of the documented word forms, second a selective index of inverted word forms and also a few indices of names (an index of personal names, placenames and other names).

The chronological volume (CHRONO) of the *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift 'Die Fackel'* (Welzig 2008) presents 555 selected examples of the overall 2,775 selected pejorative phrases arranged in chronological order as they appear in the magazine, thereby providing expanded contexts by printing the full page of the original journal in a graphically transformed facsimile where the quoted passage is highlighted. In all three volumes of the text dictionary the references to the digital edition of the journal, the AAC-Fackel - published within the framework of the corpus research enterprise "AAC – Austrian Academy Corpus" (*AAC-FACKEL* 2007) - are given by means of short URLs of the individual pages, so that the user of the text dictionary has always the full context available and can at the same time make use of the print representations of the highlighted pejorative expressions. The larger context given in the chronological volume of the text dictionary allows the reader to evaluate the textual dynamics and the effects how the pejorative expression is constituted, which in many cases is gradually intensified and accompanied by other related expressions in the context of the satirical or polemical text of "Die Fackel". In many cases the pejorative intensifications are made possible by word formation processes or syntactical effects, which is one of the reasons why the particular use of compounds is documented to a larger extent as well as certain significant pejorative collocations are taken into consideration for this text dictionary of insults and invectives. In many cases words that are not commonly used pejoratively are used in this way by the satirist, who is also reflecting upon this particular satirical procedure in his texts.



30. Dezember 1911

— 51 —

Dem Freunde zu gefallen
Sei Liebling treu und brav!
Die Worte sie verhallen
Ich wünsche Dir vor allem
Den schönsten, tiefsten Schlaf.

In wessen Liebe die Welt so liebenswert wird, der schafft dem Weltfeind eine frohe Stunde.

Die neue Art des Schimpfens

Von **Karl Kraus**

In einem Aufsatz »Über eine neue Art des Schimpfens« sagt der Professor der Literaturgeschichte Dr. Richard M. Meyer in Berlin:

. . . . Maximilian Harden, wenn ich nicht irre, hat auch hier bahnbrechend gewirkt . . . Bei einem Konflikt mit einer Zeitung, die ihn unzweifelhaft verleumdet hatte, genügte es ihm nicht, sie »Dirne« zu nennen — er mußte diese Bezeichnung noch durch die ekelhaftesten Einzelheiten verdeutlichen. Das hat nicht nur bei seinem in aller erbittertsten Feindschaft treuesten Schüler Karl Kraus in Wien Schule gemacht. Die beliebtesten Scheltwörter ‚Eunuch' und ‚Metze' werden stimmungsvoll ausgemalt und so zwei Götzen unserer Zeit zugleich gedient — jenen beiden, die Götzen sowohl primitiver als raffinierter Roheit sind: dem Sinnenkitzel und dem Zerstörungsdrang. Ich muß mich in diesem Punkt als durchaus altmodisch bekennen. Ich glaube durchs Leben zu kommen, ohne gewisse Worte auszusprechen oder auszuschreiben; und ich halte es für das berechtigte Privileg des gebildeten, ja des anständigen Menschen, auf schmutzige Ausdrücke ebenso zu verzichten wie auf schmutzige Handlungen — soweit wenigstens, wie ihnen eine freie Wahl bleibt; der Prüderie in wissenschaftlichen Fragen soll damit selbstverständlich nicht das Wort geredet werden. . . .

Gewiß nicht. Was aber die neue Art des Schimpfens anlangt, so glaube ich nicht, daß man mir in den dreizehn Jahrgängen der Fackel einen Fall wird nachweisen können, wo ich das Wort »Dirne« als Schimpfwort gebraucht hätte und nicht als die Bezeichnung eines erstrebenswerten Zustandes. Wo ich das Wort selbst gebraucht und nicht vielmehr zitiert habe, um die engstirnige Terminologie einer Gesellschaft zu brandmarken, die zu erbärmlich ist, um den mesquinen Ton eines Wortes zu verstehen, weil sie zu feig ist, um dafür das Wort Hure zu setzen. Was die neue Art des Schimpfens anlangt, so glaube ich nicht, daß man mir in

aac.ac.at/F/339_051

Fackel Schimpfwörterbuch

30. Dezember 1911

— 52 —

den dreizehn Jahrgängen der Fackel einen Fall nachweisen wird, wo ich die Dirne nicht der Verachtung entrissen und nicht sofort dem Haß als Hure präsentiert hätte. Wo ich je anders als mit Wucherzinsen der Verachtung dem Gesellen heimzahlte, der das Wort als einen schmutzigen Ausdruck nicht deshalb auffaßte, weil es den Schmutz seiner Gesinnung trug, sondern weil er das davon bezeichnete Leben für ein schmutziges hielt, für eines, an dem er sich erst seine Finger, seine Notdurft, seine ganze ekle Leiblichkeit und Moralität abwischen durfte, um als Bürger oder Professor der Literaturgeschichte rein dazustehen. Was die neue Art des Schimpfens anlangt, so glaube ich nicht, daß man mir in den dreizehn Jahrgängen der Fackel einen Fall nachweisen wird, wo ich einer Zeitung die Ehre erwiesen hätte, sie eine Dirne zu nennen, es sei denn, daß ich an meine Geringschätzung einer Zeitung absichtlich und ausdrücklich das Maß der Geringschätzung angelegt habe, das ein Literaturprofessor für eine Dirne und nicht für eine Zeitung hat. Denn da ein solcher nichts so sehr verachtet wie eine Dirne und vor nichts so viel Respekt hat wie vor einer Zeitung, so konnte ich, um mich ihm verständlich zu machen, keinen andern Vergleich wählen und ich hätte weiß Gott wie viel darum gegeben, daß es auch mir gegönnt gewesen wäre, eine Dirne zu verachten, um sie einmal aus tiefstem Herzen eine Zeitung nennen zu können. Was aber die neue Art des Schimpfens anlangt, so möchte ich mich fast zu der Behauptung versteigen, daß man mir in den dreizehn Jahrgängen der Fackel auch keinen Fall nachweisen wird, wo ich den Herren der Gesellschaft im Verhältnis zu den Dirnen einen andern Schwächezustand- als den geistigen und moralischen zum Vorwurf gemacht, wo ich mit einem Wort das Wort »Eunuch« als Schimpfwort gebraucht hätte und nicht als die Bezeichnung eines erstrebenswerten Zustandes. Eines erstrebenswerten Zustandes für Männer, welche die Dirnen nur deshalb verachten, weil sie im Gegensatz zu den Dirnen, die wenigstens Weiber sind, alles Mögliche sind, nur keine Männer. Für Kerle, die zu Unrecht einen Bart tragen und mit den Attrappen der Männlichkeit über ihre wahre Beschaffenheit den Betrachter zu täuschen wissen. Mit diesen Verkehrshindernissen sich dem Geist in den Weg stellen und durch solchen Schabernack einer mehr ornamentalen oder koloristischen als in den Vorzügen des Kerls begründeten Wirkung die wahre

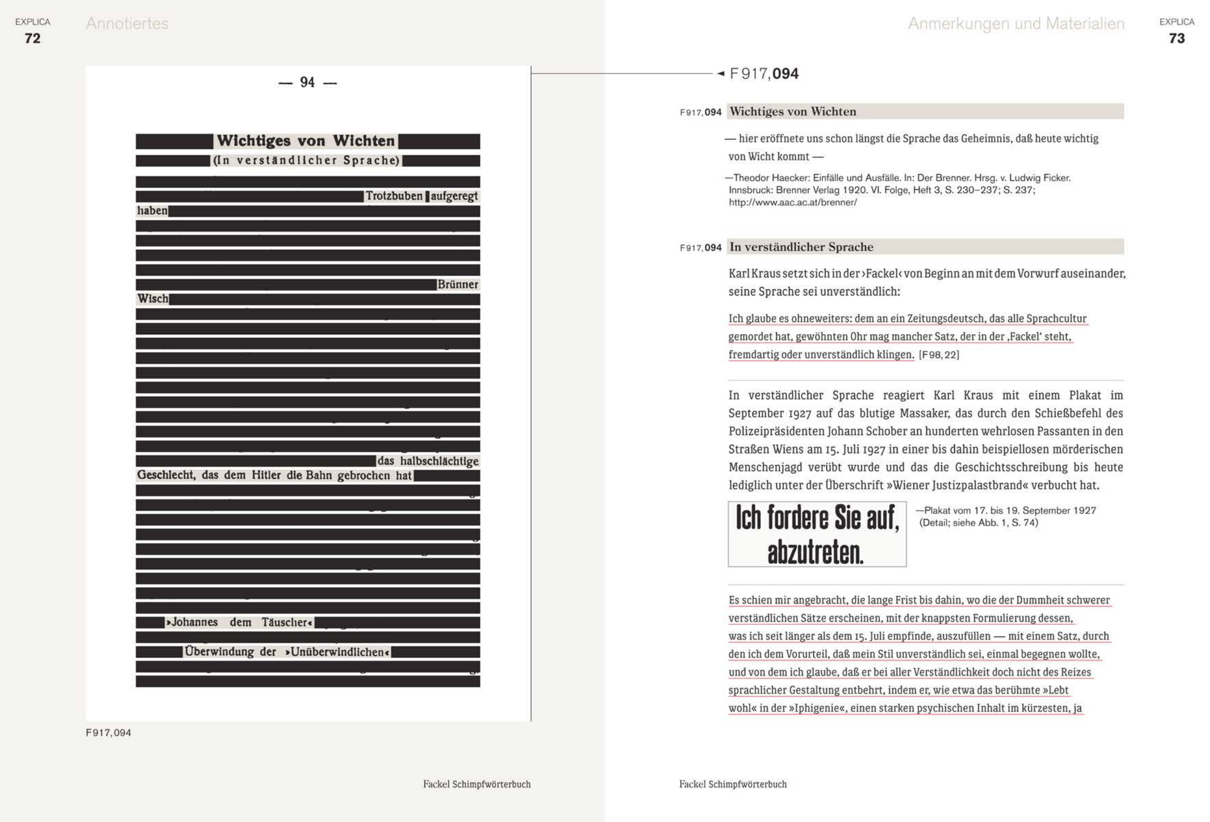
aac.ac.at/F/339_052

Fackel Schimpfwörterbuch

Figure 2: CHRONO.

The third volume (EXPLICA) contains explicatory texts as well as "Wichtiges von Wichten", the final article of "Die Fackel", where pejorative phrases were marked up in this specific text and have been accompanied by detailed commentaries. This last text published in "Die Fackel" in February 1936 is treated as a source text for the analysis of pejorative terms in order to exemplify the difficult and ambitious task of getting to terms with the high level of pejoration in the satirical and polemical texts written by Karl Kraus. The politically interesting polemical note "Wichtiges von Wichten" from 1936, referring to the political situation at the time and how to react to it by means of writing, is reproduced in the explicatory volume of the dictionary in a plain form first, giving the readers an uninterrupted chance to read a large piece of textual evidence first and then it is given in a form in which, according to the text lexicographers' interpretations, pejoratively used expressions are highlighted in the text and provided with their alphabetical list. In a third part of this volume the same text is reproduced again, this time dashed out with only those expressions left visible which are commented upon by the editors and compilers of the text dictionary, documenting the intensive need and necessity for detailed commentary in order to understand and fully assess the pejorative qualities of the expressions selected.

This source text in particular as well as the whole journal in general is representing a literary genre that offers a variety of different forms of pejorative language to be studied from various perspectives. The various use of the pejorative expressions not only in the last text, but above all in all texts chosen are, as has been observed, dominated to a large extent by the creative transformations and configurations performed by the writer. These creative adaptations and modifications are not only important for the documentation and the analysis of the idiomatic expressions as represented in the text lexicographic project of the "Dictionary of Idioms", but also to a large extent these creative adaptations are most relevant for the way, in which pejorative expressions are formed, which can be studies in great detail in the "Dictionary of Insults and Invectives". In the case of the idioms, a more or less normal form is creatively transformed into others forms, which cannot easily be detected by standard automated corpus query systems, only systematic annotation and possible semi-automatic methods could provide the scholar with reasonable results to be gained from larger corpora. In the case of the "Dictionary of Insults and Invectives" the corpus of the whole text has been made use of for purposes of systematization. This extensive literary text is full of a great variety of satirical forms in the context of pejorative expressions. For this reason it can be used as a research object following a certain combined interest in the study of pejorative language and in phenomena related to the more methodological questions of satirical and polemical language as a research topic for text lexicography as well as for corpus research.



EXPLICA
72

Annotiertes

— 94 —

[redacted] Wichtiges von Wichten [redacted]
[redacted] (In verständlicher Sprache) [redacted]

[redacted] Trotzbuben [redacted] aufgeregt
haben [redacted]

[redacted] Brünner
Wisch [redacted]

[redacted] das halbschlächtige
Geschlecht, das dem Hitler die Bahn gebrochen hat [redacted]

[redacted] »Johannes dem Täuscher« [redacted]

[redacted] Überwindung der »Unüberwindlichen« [redacted]

F917,094

Fackel Schimpfwörterbuch

Anmerkungen und Materialien

EXPLICA
73

◂ F917,**094**

F917,**094** **Wichtiges von Wichten**

— hier eröffnete uns schon längst die Sprache das Geheimnis, daß heute wichtig von Wicht kommt —

—Theodor Haecker: Einfälle und Ausfälle. In: Der Brenner. Hrsg. v. Ludwig Ficker. Innsbruck: Brenner Verlag 1920. VI. Folge, Heft 3, S. 230–237; S. 237; http://www.aac.ac.at/brenner/

F917,**094** **In verständlicher Sprache**

Karl Kraus setzt sich in der ›Fackel‹ von Beginn an mit dem Vorwurf auseinander, seine Sprache sei unverständlich:

Ich glaube es ohneweiters: dem an ein Zeitungsdeutsch, das alle Sprachcultur gemordet hat, gewöhnten Ohr mag mancher Satz, der in der ‚Fackel' steht, fremdartig oder unverständlich klingen. [F98, 22]

In verständlicher Sprache reagiert Karl Kraus mit einem Plakat im September 1927 auf das blutige Massaker, das durch den Schießbefehl des Polizeipräsidenten Johann Schober an hunderten wehrlosen Passanten in den Straßen Wiens am 15. Juli 1927 in einer bis dahin beispiellosen mörderischen Menschenjagd verübt wurde und das die Geschichtsschreibung bis heute lediglich unter der Überschrift »Wiener Justizpalastbrand« verbucht hat.

Ich fordere Sie auf,
abzutreten.

—Plakat vom 17. bis 19. September 1927
(Detail; siehe Abb. 1, S. 74)

Es schien mir angebracht, die lange Frist bis dahin, wo die der Dummheit schwerer verständlichen Sätze erscheinen, mit der knappsten Formulierung dessen, was ich seit länger als dem 15. Juli empfinde, auszufüllen — mit einem Satz, durch den ich dem Vorurteil, daß mein Stil unverständlich sei, einmal begegnen wollte, und von dem ich glaube, daß er bei aller Verständlichkeit doch nicht des Reizes sprachlicher Gestaltung entbehrt, indem er, wie etwa das berühmte »Lebt wohl« in der »Iphigenie«, einen starken psychischen Inhalt im kürzesten, ja

Fackel Schimpfwörterbuch

Figure 3: EXPLICA.

The research initiatives concerning the magazine published by Karl Kraus from April 1899 until February 1936 has offered the lexicographers at the Austrian Academy of Sciences in Vienna a unique opportunity to study and to document the language of this important writer in great detail. No author of the 19th and 20th centuries has thought or written about the language of his contemporaries in Vienna, Berlin, or Prague as precisely, as continuously and as passionately as Karl Kraus. No other author is showing such a productive and such an effective use of pejorative expressions as Karl Kraus in his satirical and polemical texts in "Die Fackel".

3 References

AAC - Austrian Academy Corpus: AAC-FACKEL, Online Version: Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936. AAC Digital Edition No 1 (ed. Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörth), 2007, Accessed at: http://www.aac.ac.at/fackel [01/01/2007].

Welzig, W. (ed.) (1999). Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‚Die Fackel'. Austrian Academy Press, Vienna.

Welzig, W. (ed.) (2008). *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‚Die Fackel'* (3 volumes: Alphabetisches, Chronologisches, Explikatives). Austrian Academy Press, Vienna.

The Presence of Gender Issues in Spanish Dictionaries

Ana Costa Pérez
Universidad Carlos III de Madrid
Acosta.hum@uc3m.es

Abstract

Dictionaries are ideological creations as they are but a reflection of society itself. A dictionary sets a standard for language; makes an authority, a cultural product, and builds a lexical *encyclopaedia* and a social reference. This analysis is built upon the idea of anundeniable overlap between ideology and dictionary and the role of the latter as a mechanism to transmit the limited sights of everything around us; talking is a world-defining act by individuals who forces themselves to adapt to a code which is seemingly open and closed at the same time; a code imposed by the society to which they belong and which will be enforced on future generations. Therefore, the present work will highlight the catalogue of definitions that challenge the descriptive neutrality of current lexicographical work, turned into dictionaries which should mirror an equal society, without discrimination.

This requires defining the concept from different scopes: linguistic, anthropological, sociolinguistic, philosophical or cognitive. We aim at showing how grammar, with its two (and even up to three) *genera,* provides us with the perfect field to focus on sexes, at both biological and social senses, on Nature and Culture, without favouring the existence of two different sexes nor any individual powers and decisions. The *Academia* notes that the Spanish language foresees the possibility to refer to mixed groups through the grammatical masculine gender, possibility in which there is no discriminatory intent, but the application of the linguistic law based in the expressive economy. Only when the opposition of the sexes is a relevant factor in the context, the Academy considers necessary the explicit presence of both genders

Keywords: Dictionaries; Gender; Ideology

1 Preliminal Issues

Gender is linguistically reserved for words in which sex evidences a sexed condition of human beings. But if such a distinction is so shockingly clear, why does it continue to emerge almost cyclically? why the same arguments about this pair of concepts in specialized areas such as grammar or lexicography?

Taking a diachronic perspective, we can check how talking about gender from a feminist perspective, rather than a more objective one, sets a more objective scope, that of women-centered movements.

The objective, as such, arises when we try to return gender to unmarked meanings, free from criticism, politics or claims.

When defining one or several related categories based on what has been historically common to them, it is essential to analyze the gaps where they cease to be common. For example, categorically man or woman can not conceive based on specific characteristics, but the problem is that some concepts contain a complex network of variables that unravel some concepts of our language, such as gender, mother, sex, marriage, education, fatherhood, female, manhood, religion or science. Though taking the form of descriptive definitions, they are actually conditioned acts. This is due to an attempt to isolate, to "numb" the emotional charge that for years has been lodged in certain socially marked terms.

That is why a big part of what has been traditionally charged to both man and woman have depended on an intervened meaning. Therefore, we have chosen a descriptive approach based on the analysis of the definitions provided by the academic dictionary related to various *nuclea* of meaning: the major professions, body differences, adjectives related to personality or maternity / paternity will be the main ones. Through comparative analysis (masculine and feminine terms will be opposed), we will try to highlight the ideology underlying the definitions of the Spanish Academy dictionary. (Figure1)

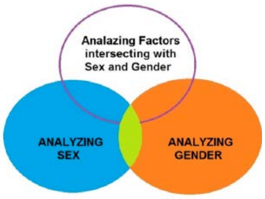


Figure 1: Analizing Gender.

2 Methodology

For the last 20 years, treating languages by meanings of databases (linguistic *corpora*) has turned into a challenge that has been imposed to all the linguists and lexicographers. The structure of information in IT bases presents big advantages. The constant number of systems of meaning (semantic fields) allows describing the lexical elements in an unified and coherent way. Thus, it is guaranteed, for all the lexical units of a certain type, a common methodological response. The comparisons among related terms are possible and objectivable and we can proceed to controls which guarantee coherence, regularity and uniformity.

Corpus linguistics stands as a method for carrying out linguistic analyses. As it can be used for researching many kinds of linguistic questions and, as it has been shown to have the potential to yield highly interesting and new insights about language and relationships between men and women, it has become one of the most widespread methods of linguistic research in the last years.

In order to elaborate complete and systematic definitions, I hereby make a proposal to establish a standard of definition for every category included in meaning systems (1 to 6). Then, we will decide the characteristics that must appear in the concept definitions belonging to the systems that I specify in the appendix. At present, different strategies or methods of definition coexist, and then the lexicographer chooses, yet never openly opting for a single one, combining different methods.

For these definitions we have to bear in mind the following:

(1) Context: general information of the word in question in relation to its frequency of use.

(2) Situation: how people deal with, define or perceive the term. This point is based on the topics the study is based upon.

(3) Perspectives: possible ways of defining a term (end)

(4) I process (try): it (he, she) sequences of semes, flow of information, changes of meaning in the time.

(5) Activities and events: difficulties to find definitions related the obsolescence or frequency of the different meanings.

(6) Strategies: ways of mana ging information.

(7) Relations and social structure: ideology presents in the standards of definitions.

After extracting the terms from the CREA (Sincronic Spanish Database) and CORDE (Diacronic Spanish Database) databases, we have proceeded to arrange the results from the definitions of the dictionaries. They are generally accessible *corpora*, accessible to everyone *via* Spanish Academy Web (http://www.rae.es).

Coming across the definitions and contexts, we will be able to observe the features describing men and women

The definitions which dictionaries offer about correlative terms (in masculine and feminine) do not coincide at informative levels, thus offering a diverse kind of *seme* in every case.

The aim of this project is to analyze the definitions that refer to terms related to women and men in different semantics systems and them to arrange when recounted to men and to analyze possible faults so much to macrostructure as well as microstructure level. We pretend to give a fully readable account of how dictionaries represent women and men.

The result of this study goes towards a standard of definition for the concepts included in the systems listed in the appendix.

3 Discussion

Gender, as we have seen, is globally understood as the set of beliefs, prescriptions and attributions that are socially constructed taking sexual difference as a base. This social construction works sometimes as a kind of cultural "filter", one through reality is interpreted, due to that tendency of every society to define what is proper for women and what is proper for men, and above this cultural framework, setting out the obligations of each sex.

From childhood we perceive representations of what is proper to each sex through language, and the materiality of culture (objects, images, etc.). It has been that children between two and three years old, know how to refer to themselves in feminine or masculine, but do not necessarily have a clear notion of the actual biological differences.

Today, the notion of ideology, born linked to *bourgeois* society in which a set of values and ideals, driven by political and social pluralism, led to today's modern society. Social representations are symbolic constructions that give powers to the objective and subjective behavior of people. The social environment is more than a territory, a symbolic space defined by the imagination, and decisive in the construction of each person self-image, consciousness is inhabited by social discourse. Lucien Goldmann states in this regard that

> "The overall vision of human relations between man and the universe implies, this type of collective consciousness, the possibility, and often the actual presence of a ideal man and this leads us to differentiate the type of collective consciousness that we call ideology, called view of the world "(Goldman, 1969: 210).

Sociolinguistics: Although the behavioral differences between men and women are explained in a general way as a product of sex (vocal cords, tonal range), gender has been linked to the position in society and the complex network of relationships that are developed in within it.

Anthropological: The word gender attempts to rebuild each and every one of the areas of significance that have been superimposed for decades, unraveling the network of relationships and social interactions that are constructed from the symbolic division into sexes. In the psychological field, we also emphasize the neither natural nor spontaneous character of the categories male and female,

Linguistic: We aim at showing how grammar, with its two (and even up to 3) genera, provides us with the perfect field to focus on sexes, at both biological and social senses, on Nature and Culture, without favoring the existence of two different sexes nor any individual powers and decisions. The Academy notes that the Spanish language foresees the possibility to refer to mixed groups through the grammatical masculine gender, possibility in which there is no discriminatory intent, but the application of the linguistic law based in the expressive economy. Only when the opposition of the sexes is a relevant factor in the context, the Academy considers necessary the explicit presence of both genders.

4 Conclusion

Human beings symbolize a basic material, which is identical in every society: bodily difference, specifically sex. Although apparently biology shows that human beings are in both sexes, more combinations arise from the five physiological areas. Of these five areas depends on what, in general terms and in a very simplistic way, has been called the "biological sex" of a person: genes, hormones, gonads, internal reproductive organs and external reproductive organs (genitals).

Although the multitude of cultural representations of biological facts is very large and has varying degrees of complexity, sexual difference has some basic persistence and is the source of our image of the world, as opposed to some other. The body is the first uncontrollable evidence of human difference. The culture marks human beings with gender and gender marks perception of everything else: social, political, religious and quotidian.

The dictionary is an ideological creation. It reflects the society and the dominant ideology. As indisputable authority, as a cultural tool, the dictionary acts as a fixing element and intends to the conservation, not only of language but also the attitudes and ideology behind it.

5 References

Arias Barredo, A. (1995): De feminismo, machismo y género gramatical, Valladolid, Universidad.

Blecua, J.M. (1990): «Análisis provisional de una muestra aleatoria en el DRAE», en El vocabulari i l'escrit, Barcelona: Universidad de Barcelona, 1990.

Calonge, J. (1981): "Implicaciones del género en otras categorías gramaticales". Logos Semantikos, Studia Linguistica in honorem Eugenio Coseriu, Madrid, Gredos, IV, 1991, pp. 19-28.

Casares, J. (1992 [1950]): Introducción a la lexicografía moderna, Madrid, CSIC, 1992.

Demonte, V. (1982a): Lenguaje y sexo, ideología y papeles sociales, Madrid, Akal.

Goldman, Lucien. 1969. The Human Sciences and Philosophy. London: Cape.

López García, Á. y Morant, R. (1991): Gramática femenina, Madrid, Cátedra.

López García, Á. (1992): Lenguaje y discrimación sexista en los libros escolares, Murcia, Universidad

Lozano Domingo, I. (2005): Lenguaje femenino, lenguaje masculino, Minerva: Madrid.

Pascual J.A. y Olaguíbel, M.C. (1992): «Ideología y diccionario», en I. Ahumada Lara (ed.): Diccionarios españoles: contenido y aplicaciones. Lecciones del I Seminario de Lexicografía Hispánica, Facultad de Humanidades, Jaén, 21 al 24 de enero de 1991, El Estudiante Facultad de Humanidades, Jaén, 1992, pp. 73- 89.

5.1 Dictionaries

Alvar Ezquerra, Manuel (dir.) (2000): Diccionario para la enseñanza de lengua española, Barcelona, Bibliograf / Universidad de Alcalá de Henares.

Gutiérrez Cuadrado, Juan (dir.) (1996): Diccionario Salamanca de la lengua española, Madrid, Santularia/ Universidad de Salamanca.

Moliner, María (1998): Diccionario de uso del español, Madrid, Gredos.

Real Academia Española (2001): Diccionario de la lengua española, Madrid, Espasa- Calpe, vigésima segunda edición.

Appendix

Corpora: List of Words used in the research (Spanish Language).

System 1.Women social respect

Distinguished: dama, damisela, dona, dueña, gran señora, madama, madamisela, madona, maestresa, matrona, ricadueña, ricahembra, señora, señorita, señora principal, señorita.

Esteem: mujercilla, mujeruca, mujerzuela, mujeruca, pingo, prójima

System 2: Women relationships

Lawful: dama, barragana, cara mitad, conyuge, consorte, costilla, esposa, desposada, media naranja, mujer, mujer velada, mujer de bendición, oponente, pareja, señora.

Unlawful: barragan, coima, combleza, compañera, concubina, daifa, entretenida, manceba, pretendida

System 3. Women appearance

Seme +beuty: Beldad, belleza, bombón, gachí, gachona, hembra, hembra de bandera, hermosura, monumento, preciosidad, sílfide, venus.

Seme stocky: buena moza, moza, mujerona, real moza

Seme masculinity : machota, machirulo, marimacho, marota, varona, varonesa, virago.

System 4: Women procreation

seme sterile: horra,mañera, machorra

seme fertile: descinta, embarazada, empreñada, encinta, gestante, gravida, madre, malparida, multiparia, mulipara, parida, paridera, paridora, parturienta, parturiente, preñada, primeraza, primípara, puérpera, recién parida.

System 5: Women sexuality

seme virginity: doncella,doncellueca, escosa, entera, poncela, prematura, virgen

seme homosexuality: bollera, fricadora, lesbiana, tortillera, tribada

seme “sexual desires”: cachonda, salida, ninfómana.

Seme “ sex trade”: prostituta, andorra, ave nocturna, bagasa, baldonada, bacanera, burraca, buscona…

System 6: Women personality

Seme “dishonest” corralera, escaldada, facilona, farota, galante, mujer fatal, piruja, tigresa, tragona, vampiresa, ventanera,

Seme “gossip”: alcahueta, celestina, comadre, lagarta, pécora, tercera, trotaconventos, víbora.

Seme“bad temper”: arpía, mujerota, sargenta, sargentona

Seme “boastful”: bachillera, coqueta, lechugina, marisabidilla, petimetra.

Reflexive Verbs in a Valency Lexicon: The Case of Czech Reflexive Morphemes

Václava Kettnerová, Markéta Lopatková
Charles University in Prague
kettnerova@ufal.mff.cuni.cz, lopatkova@ufal.mff.cuni.cz

Abstract

In this paper, we deal with Czech reflexive verbs from the lexicographic point of view. We show that the Czech reflexive morphemes *se* and *si* constitute different linguistic meanings: either they are formal means of the word formation process of the so called reflexivization, or they are associated with the syntactic phenomena of reflexivity, reciprocity, and diatheses.

All of these processes are associated with changes in the valency structure of verbs. We formulate a proposal for their lexicographic representation for the valency lexicon of Czech verbs, VALLEX. We make use of the division of the lexicon into a data component and a grammar component which represents a part of the overall Czech grammar. The data component stores information on valency structure of verbs in unmarked (active) structures. The grammar component consists of formal rules describing regular changes in the valency structure of verbs; these rules allow for the derivation of valency frames underlying the usages of verbs in marked structures (reflexive, reciprocal, deagentive and dispositional) from the valency frames corresponding to unmarked structures (non-reflexive, unreciprocal, and active).

Czech reflexive verbs thus represent an illustrative example of the lexical-grammar interplay: we demonstrate that a close interaction between the lexicon and the grammar is necessary for a representation of these verbs and they both are indispensable if such a representation is to be adequate and economical.

Keywords: reflexive verb; reflexive morpheme; valency lexicon; Czech

1 Introduction

In this paper, the possibility of a lexicographic representation of the reflexivity of Czech verbs is described in detail. In Czech, the reflexives *se* and *si* are on the one hand formal means of word formation process of so called reflexivization; on the other hand, they are associated with syntactic phenomena of reflexivity (in the narrow sense, also called "true reflexives"), reciprocity, and diatheses. According to their function, the reflexives *se* and *si* represent clitic morphemes corresponding either (ia) to components of verb lemmas (*se* and *si*), or (ib) to a component of verb form (only *se*), or (ii) to the personal pronoun *se* (with its inflected variant *si* and corresponding non-clitic variants *sebe* and *sobě*, respectively). As examples (1), (2), (3) and (4) with the verb *zabít* "to kill" show, both types of reflexives can occur with a single verb, constituting different linguistic meanings: in example (1), *se* is a component of the verb lemma *zabít se* "to kill (oneself)" (type (ia)); in examples (2) and (3), *se* is interpreted as the reflexive personal pronoun (however, expressing different meanings, true reflexivity and reciprocity, respectively, type (ii)); and example (4) illustrates *se* as a component of a verb form of the verb *zabít* "to kill" (type (ib)).

(1) *Zabil se pádem ze střechy.* (CNC, SYN2006pub)
"killed – SE_{morph} – by falling – from roof."
Eng. He killed himself (unintentionally) by falling from the roof.

(2a) *Zabil se vlastní zbraní …* (CNC, SYN2006pub)
"killed – SE_{pron} – own – weapon …"
Eng. He killed himself with his own weapon …

(2b) *Zabil sebe vlastní zbraní.*
"killed – $\text{SEBE}_{\text{pron}}$ – own – weapon …"
Eng. He killed himself with his own weapon …

(3a) *Zabili se navzájem.*
"killed – SE_{pron} – each other."
Eng. They killed each other.

(3b) *Zabili sebe navzájem.*
"killed – $\text{SEBE}_{\text{pron}}$ – each other."
Eng. They killed each other.

(4) *… zabila se dvě vykrmená prasata a pečínka provoněla celý dům.* (CNC, SYN2005)
"… killed – SE_{morph} – two – fattened – pigs – and – roast meat – scented – the whole – house."
Eng. Two fattened pigs were killed and roast meat scented the whole house.

We show that the reflexives *se* and *si* belong to several different language phenomena which involve specific changes in the valency structure of verbs. As a consequence, they require to be represented in a lexicon in different ways. Here we describe the representation of these phenomena in the Valency Lexicon of Czech Verbs, VALLEX.

1.1 Related Work

Reflexivity has been extensively studied in the theoretical linguistics since the 1980s. The research has focused on linguistic means encoding reflexivity, their interpretations and ambiguities in individual languages. Recently, this linguistic phenomenon has received considerable attention even from the cross-linguistic perspective (König & Gast, 2008), (Nedjalkov, 2007). Numerous analyses show that linguistic means expressing reflexivity are usually ambiguous as they fulfill diverse functions in individual languages and that drawing clear distinctions between these functions represents a tricky task. For these reasons, developing a satisfactory lexicographic representation of reflexivity – despite being highly beneficial esp. for natural language processing and foreign learners – remains rather challenging (Renau & Battaner, 2012).

In Czech, reflexivity encoded by the reflexives *se* and *si* represent widely debated phenomenon from both theoretical (Oliva, 2001; Panevová, 1999, 2007) and computational point of view (Oliva, 2003; Petkevič, 2013). Two primary functions of the Czech reflexives are determined: (i) the reflexives as components of verb lemmas or verb forms and (ii) the reflexives as the personal pronoun, see Section 1. However, despite the plenitude of studies focused on Czech reflexivity, testable criteria for their distinction have not yet been established. In most cases, the substitutability of the reflexives *se* and *si* with *sebe* and *sobě*, respectively, can be applied as an operational test for distinguishing the reflexive personal pronoun from *se* and *si* as the components of verb lemmas or verb forms. However, this test can fail esp. in cases where the substitution leads to stylistically unacceptable sentences or in cases of haplology, see Section 2.1. In such cases, we adopt solutions taking economy and systematicity of the lexicographic representation into account.

As we attempt a lexicographic representation of reflexivity in a lexicon, let us introduce several lexical resources providing the information on these phenomena. First, *LexIt*, a large-scale lexical resource providing the automatically derived information on subcategorization and semantic properties of Italian verbs, nouns and adjectives stores the information on reflexivity as well (Lenci et al., 2012). Second, this type of information is also covered in *Diccionario de ensěnanza del espãnol como lengua extranjera, DAELE*, a Spanish learner's dictionary (Renau & Battaner, 2012). Third, *FrameNet* records the information on reciprocity of frame elements (linguistic phenomenon closely related to reflexivity) by adding special semantic frames indicating reciprocity (Ruppenhofer, et al., 2010).

For Czech, *PDT-VALLEX*, a valency lexicon linked with word occurrences in the Prague Dependency Treebank 2.0 (PDT) (Hajič, et al., 2006), provides the information on valency behavior of verbs, nouns, adjectives and adverbs (Hajič, et al., 2003). Although the information on reflexivity and reciprocity of verbs is not explicitly recorded in this lexicon, it can be easily extracted from PDT (if reflexive or reciprocal usages appear in the corpus). In addition, a fully automatically derived *Czech Syntactic Lexicon* (which is however not publically available) was designed, providing the information on possible reciprocity, reflexivity and diatheses of verbs (Skoumalová, 2001).

1.2 VALLEX

The valency lexicon of Czech verbs, *VALLEX*,[1] is a collection of linguistically annotated data and documentation (Žabokrtský & Lopatková, 2007; Lopatková et al., 2008). It provides the information on valency structure of Czech verbs in their particular meanings / senses, possible morphological forms of

[1] http://ufal.mff.cuni.cz/vallex/

their valency complementations and additional syntactic information accompanied with glosses and examples. In VALLEX, version 2, there are roughly 2,730 lexeme entries containing together around 6,460 lexical units ('senses'). Verb lexemes were selected according to their frequency in the Czech National Corpus.[2] The lexicon has been developed for both human users and NLP applications, and is therefore in three different formats: HTML, XML and printable versions.

In VALLEX, the valency theory developed within the theoretical framework of the Functional Generative Description (henceforth FGD) is used as the theoretical background for the description of valency of verbs, see esp. (Sgall et al., 1986), (Panevová, 1994). According to this theory, valency complementations are divided into arguments (inner participants) and free modifications (adjuncts). They both can be obligatory or optional. The types of (verbal) arguments are distinguished mainly on the basis of the syntactic behavior of verbs. Five types of arguments have been determined – 'Actor' (ACTor, label ACT), 'Patient' (PATient, PAT), 'Addressee' (ADDRessee, ADDR), 'Origin' (ORIGin, ORIG), and 'Effect' (EFFect, EFF). In contrast to the arguments, free modifications are semantically distinctive, being identified on the basis of their syntactico-semantic functions.

In VALLEX, the key information on the valency structure of a given lexical unit is encoded in the form of valency frames. A valency frame is formed as a sequence of slots; each slot stands for one valency complementation and consists of its type ('ACTor', 'ADDRessee', etc.), possible morphemic forms and its obligatoriness (obligatory or optional). Further, each lexical unit can be characterized by additional syntactic information on, e.g., syntactico-semantic class membership, diatheses, reciprocity of valency complementations, reflexivity. This information is provided in special attributes attached to individual lexical units.

The lexicon is divided into the data and the grammar component; the latter stores rules describing regular syntactic properties of verbs and it represents a part of the overall grammar of Czech, see esp. (Kettnerová et al., 2012a). The close interplay of these two parts of the lexicon is demonstrated on the representation of the reflexives *se* and *si* in the following sections. First, the reflexives *se* and *si* as components of verb lemmas (i.e., as formal means of the word formation process of reflexivization) are discussed in detail, esp. the syntactic properties of reflexivization are described and their lexicographic description is outlined in Section 2. Second, the reflexive pronoun *se* as a formal means of reflexivity (in the narrow sense) and reciprocity is surveyed and its representation in the lexicon is introduced in Section 3. Finally, the reflexive *se* as a component of reflexive verb forms that is involved in two types of Czech diatheses is debated in Section 4. In conclusion, the lexical entry of the verb *zabít* "to kill" (illustrated in Section 1) is displayed.

[2] http://www.korpus.cz/

2 Czech Morphemes *se* and *si* as Components of Verb Lemmas

In Czech, the reflexive morphemes *se* and *si* can represent *freestanding components of verb lemmas* of the so called *reflexive verbs*. In Czech, there are two types of reflexive verbs: (i) reflexive tantum verbs (Section 2.1) and (ii) derived reflexive verbs (Section 2.2).

3 Reflexive Tantum Verbs

The first type is represented by the so-called reflexive tantum verbs, i.e., the verbs that have no non-reflexive counterparts, e.g., *bát se* "to be afraid", *smát se* "to laugh", *stěžovat si* "to complain", *zapamatovat si* "to remember", *domnívat se* "to assume", *chlubit se* "to boast", *líbit se* "to like", *ptát se* "to ask", *zamilovat se* "to fall in love", see examples (5) and (6). In the case of reflexive tantum verbs, the reflexive morpheme *se* or *si* is a part of the verb lemma representing the respective verb lexeme in the data component of the lexicon.

Moreover, there are cases in Czech where a reflexive verb seemingly has a non-reflexive counterpart but these reflexive and non-reflexive verbs are not related by any derivational relation: the lexical meanings of these verbs are completely different, e.g. *dít se* "to happen" vs. *dít* "to tell", *dopustit se* "to commit" vs. *dopustit* "to fill (with water)", *hodit se* "to match" vs. *hodit* "to throw", see examples (7)-(8). These are represented as separate verb lexemes (in separate lexical entries) in the valency lexicon.[3]

(5a) *Jan se bojí zkoušky.*
"John – SE$_{\text{morph}}$ – is afraid – of the exam."
Eng. John is afraid of the exam.

(5b) **Jan bojí zkoušky.*
"John – is afraid – of the exam."

(6a) *Hosté si stěžovali na špatnou stravu v hotelu.*
"guests – SI$_{\text{morph}}$ – complained – of bad food – at the hotel"
Eng. The guests complained about bad food at the hotel.

(6b) *Hosté stěžovali na špatnou stravu v hotelu.
"guests – complained – of bad food – at the hotel"

(7) *Co se děje?*
"what – SE$_{\text{morph}}$ – happens."
Eng. What is happening?

[3] In case where a sentence contains more than one reflexive tantum verbs, the reflexive *se* can be subject to haplology: a single occurrence of the reflexive can be associated with two verbs, see the following example where both the verb *pokusit se* "to try" and *usmát se* "to smile" are reflexive tantum verbs:
Jan se pokusil usmát.
"John – SE$_{\text{morph}}$ – tried – smile."
Eng. John tried to smile.

(8) *"Pravdu díš," odvětil Petr.*

"the truth – you are telling – replied – Peter."

Eng. "You are telling the truth," replied Peter.

4 Derived Reflexive Verbs

Reflexive verbs of the second type are derived from non-reflexive verbs by adding the freestanding morpheme *se* or *si.* This process is called reflexivization, see esp. (Dokulil, 1986). In Czech, the reflexivization is a productive word formation process, which is largely syntactically motivated. Basically, two types of changes in valency structure of verbs are associated with this process (Sections 2.2.1 and 2.2.2). Further, in rare cases, reflexivization does not involve any change in valency structure (Section 2.2.3).

4.1 Reflexivization Applied to Transitive Verbs Resulting in Reflexive Intransitive Verbs

When reflexivization is applied to transitive verbs, it results in reflexive intransitive verbs associated with specific shifts in the lexical meaning of verbs: whereas non-reflexive transitive verbs express intentional acts (9a), reflexive intransitive verbs prototypically indicate non-intentional acts (9b):[4] the argument corresponding to the direct object (expressed by the accusative) of the transitive non-reflexive verb maps onto the subject (expressed by nominative) of the derived intransitive reflexive verb.

(9a) *Maminka vaří brambory.*

"mother – cooks – potatoes$_{\text{Dobj-acc}}$."

Eng. The mother is cooking potatoes.

(9b) *Brambory se vaří.*

"potatoes$_{\text{Subj-nom}}$ – SE$_{\text{morph}}$ – cook."

Eng. Potatoes are cooking.

4.2 Reflexivization Applied to Verbs Implying Reciprocity

Reflexivization is also involved in the derivation of reflexive reciprocal verbs, i.e., verbs indicating reciprocity in their lexical meaning (Panevová & Mikulová, 2007). These verbs are derived by the reflexive morphemes *se* or *si* from verbs that imply (at least two) semantically homogeneous arguments, typically structured as ACTor (in nominative) and PATient or ADDRessee (expressed either by the

[4] The act expressed by verbs denoting movement can be conceived as intentionally or unintentionally performed, e.g., *Petr opřel kolo o zeď.* Eng. Peter leaned the bike against the wall. (intentional act) – *Petr se opřel o zeď.* Eng. Peter leaned against the wall. (un/intentional act).

accusative or by the dative), see examples (10a) and (11a), respectively. (As for reflexive pronouns in reciprocal constructions, see esp. Section 3.2.)

Reflexive verbs indicating reciprocity are associated with specific changes in their valency structure: the argument of the non-reflexive verb that is expressed in the accusative or dative is expressed by a prepositional group with the reflexive verb indicating reciprocity, see examples (10b) and (11b), respectively.

(10a) *Petr potkal Marii.*
“Peter – met – Mary$_{\text{PAT-acc}}$.”
Eng. Peter met Mary.

(10b) *Petr se potkal s Marií.*
“Peter – SE$_{\text{morph}}$ – met – with Mary$_{\text{PAT-s+instr}}$.”
Eng. Peter met with Mary.

(11a) *Dědeček vypráví dětem pohádky.*
“the grandpa – tells – the children$_{\text{ADDR-dat}}$ – fairy tales.”
Eng. The grandpa is telling the children fairy tales.

(11b) *Dědeček si vypráví s dětmi pohádky.*
“the grandpa – SI$_{\text{morph}}$ – tells – with the children$_{\text{ADDR-s+instr}}$ – fairy tales”
Eng. The grandpa and the children are telling each other fairy tales.

4.3 Reflexivization without Changes in Valency Structure

For a limited number of verbs, reflexivization does not result in any changes in either the valency structure or the meaning. The derivation by the morphemes *se* or *si* without clear syntactic or semantic motivation can be illustrated by the following examples (12) and (13).

(12a) *Myslím, že je to dobře.*
“I think – that – is – it – good.
Eng. I think that it is good.

(12b) *Myslím si, že je to dobře.*
“I think – SI$_{\text{morph}}$ – that – is – it – good.”
Eng. I think that it is good.

(13a) *Zítra začíná ve městě festival vína.*
“tomorrow – starts – in the town – a festival of wine.”
Eng. Tomorrow a wine festival starts in the town.

(13b) *Zítra se ve městě začíná festival vína.*
“tomorrow – SE$_{\text{morph}}$ – in the town – starts – a festival of wine.”
Eng. Tomorrow a wine festival starts in the town.

5 Representation of Reflexive Tantum and Derived Reflexive Verbs in the Lexicon

In the case of both *reflexive tantum verbs* and *derived reflexive verbs*, the reflexive morphemes *se* and *si* are represented *in the data component* of the lexicon as a part of their verb lemmas (Section 2.1, 2.2.1 and 2.2.2, respectively). Derived reflexive verbs and their non-reflexive counterparts are recorded as separate verb lemmas (and thus separate lexical entries). Only derived reflexive verbs without syntactic changes (Section 2.2.3) are handled as variants of the respective non-reflexive verbs.

6 Czech Morphemes *se* and *si* as a Reflexive Pronoun

The reflexive *se* can also represent a *personal pronoun* (with the morphemic form *se* for accusative and *si* for dative, and their non-clitic variants *sebe* and *sobě,* respectively). The reflexive pronoun expresses reflexivity (in the narrow sense, Section 3.1) and reciprocity (Section 3.2).

6.1 Reflexivity

In cases where ACTor performs an action that is focused on himself/herself (also called "true reflexivity"), the reflexive pronoun *se* is used in Czech as a formal means of grammatical coreference, see esp. (Hajičová, et al., 1985, 1986, 1987): in these cases, the reflexive pronoun *se* stands for an argument of the verb that is referentially identical with ACTor in the subject position, examples (14) and (15). The form of the reflexive pronoun depends on the morphemic case of the argument (*se* in the accusative and *si* in the dative). In the case of reflexivity, the clitic forms of the reflexive pronoun *se/si* can be replaced by their non-clitic variants *sebe/sobě*:[5]

(14a) *Petr se myje.*
"Peter$_{\text{ACT-Subj}}$ – SE$_{\text{pron-acc}}$ – washes."
Eng. Peter is washing himself.

(14b) *Petr myje sebe (ale ne dítě).*
"Peter$_{\text{ACT-Subj}}$ – washes – SEBE$_{\text{pron-acc}}$ – (but not the child)."
Eng. Peter is washing himself (but not the child).

(15a) *Marie si koupila k obědu sendvič.*
"Marie$_{\text{ACT-Subj}}$ – SI$_{\text{pron-dat}}$ – bought – for lunch – a sandwich."
Eng. Mary bought herself a sandwich for lunch.

[5] The use of clitic and non-clitic variants of the reflexive pronoun is affected esp. by the topic-focus articulation – thus the possibility to replace clitic forms by non-clitic forms of the reflexive pronoun in a sentence is often conditioned by changes in word order; however, this issue is not addressed in this paper as it goes beyond its scope.

(15b) *Sobě k obědu Marie koupila sendvič, dětem hranolky.*
“SOBĚ$_{\text{pron-dat}}$ – for lunch – Marie$_{\text{ACT-Subj}}$ – bought – a sandwich, – to the children – French fries”
Eng. Mary bought a sandwich to herself and French fries to children for lunch.

Reflexivity is represented in the lexicon by a special attribute -rfl attached to relevant lexical units. In this attribute, the information about the possibility of the reflexive usage of some arguments is provided by the value cor3 (for arguments in the dative, example (15)) and cor4 (for arguments in the accusative), example (14)). Other forms (e.g., prepositional groups) are not explicitly marked in the lexicon as they are expressed only by long variants of the reflexive personal pronoun (which are not ambiguous).

6.2 Reciprocity

Further, the reflexive pronoun *se* can express reciprocity. Reciprocalization is a syntactic operation on two (or three) arguments of a verb which puts the involved arguments in the symmetry. The main conditions imposed on such arguments are (i) their semantic homogeneity and (ii) same status with respect to topic-focus articulation. Reciprocalization leads to specific changes in the valency structure of a verb: the involved argument expressed in a less prominent surface syntactic position is shifted to the more significant position (subject or direct object) of the other symmetrically used argument, see (Panevová, 1999, 2007) and (Panevová & Mikulová, 2007). The resulting surface syntactic structure is characterized by a “multiplied” subject (or direct object) which is filled by a coordination, example (16), morphological, example (17), or semantic plural (e.g., the collective noun in example (18)). The syntactic position of the shifted (less significant) argument is typically formally filled by the reflexive pronoun *se* (expressed in the appropriate case), see below.

(16a) *Petr a Pavel se bijí.*
“Peter – and – Paul – SE$_{\text{pron-acc}}$ – beat.” “
Eng. Peter and Paul are beating each other.

(16b) *Petr a Pavel bijí sebe navzájem.*
“Peter – and – Paul – beat – SEBE$_{\text{pron-acc}}$ – each other.”
Eng. Peter and Paul are beating each other.

(17a) *Děti se bijí.*
“children – SE$_{\text{pron-acc}}$ – beat.”
Eng. Children are beating each other.

(17b) *Děti bijí sebe navzájem.*
“children – beat – SEBE$_{\text{pron-acc}}$ – each other.”
Eng. Children are beating each other.

(18a) *Celá rodina si pomáhá.*
“whole – family – SI$_{\text{pron-dat}}$ – help

Eng. The whole family helps each other.

(18b) *Rodina pomáhá sobě (navzájem – a ne jim).*

“family – helps – SOBĚ$_{\text{pron-dat}}$ (each other – and not them).”

Eng. The family helps each other.

In Czech, reciprocal constructions are created by two different types of verbs, by non-reciprocal verbs, i.e., by verbs that do not imply reciprocity in their lexical meaning (Section 3.2.1), and by inherently reciprocal verbs (Section 3.2.2).

6.2.1 Verbs Not Implying Reciprocity

Many verbs in Czech can potentially express reciprocity although reciprocity is not implied in their lexical meaning,[6] e.g., *děkovat* “to thank”, *obviňovat* “to accuse”*, hrozit* “to threaten”, *pomáhat* “to help”, *vydírat* “to blackmail”, examples (19) and (20). In such cases, the reciprocal constructions (as described above) are optionally accompanied with the lexical expressions *vzájemně, navzájem, jeden druhý* “each other”, and *spolu* “together”, emphasizing the reciprocal meaning.

(19) *Manželé se (vzájemně) obviňují z nevěry.*

“husband and wife – SE$_{\text{pron-acc}}$ – (each other) – accuse – of infidelity.”

Eng. Hausband and wife accuse each other of infidelity.

(20) *Otec a syn si (vzájemně) lhali, aby si neublížili.*

“father and son – SI$_{\text{pron-dat}}$ – (each other) – lied – in order to – SI$_{\text{pron-dat}}$ – not-to-hurt .”

Eng. Father and son lied (to each other) in order not to hurt each other.

The lexical expressions (explicitly) indicating reciprocity are, however, obligatory in reciprocal constructions created by reflexive tantum verbs that do not imply reciprocity. For instance, although the verbs *smát se* and *vysmívat se* “to laugh at” do not imply reciprocity, their ACTor and ADDRessee can be put in the symmetrical relation. In reciprocal constructions with these verbs, *se* represents the morpheme that is a component of the verb lemmas, not the reflexive pronoun (as it cannot be substituted with the non-clitic form *sebe*). The reciprocity therefore must be expressed by lexical means, see example (21a) and (22a). If no such lexical means is present, the construction is either not reciprocal (21b), or even not grammatical (22b):

(21a) *Petr a Pavel se smáli jeden druhému.*

“Peter and Paul – SE$_{\text{morph}}$ – laughed – at each other.”

Eng. Peter and Paul were laughing at each other.

(21b) *Petr a Pavel se smáli.*

“Peter and Paul – SE$_{\text{morph}}$ – laughed.”

Eng. Peter and Paul were laughing.

[6] Compare also with Section 2.2.2 describing derived reflexive verbs that imply reciprocity in their lexical meaning.

(22a) *Petr a Pavel se vysmívali jeden druhému.*
“Peter and Paul – SE$_{\text{morph}}$ – laughed – at each other.”
Eng. Peter and Paul were laughing at each other.

(22b) **Petr a Pavel se vysmívali.*
“Peter and Paul – SE$_{\text{morph}}$ – laughed.”

6.2.2 Verbs Implying Reciprocity

In addition, reciprocalization can be also applied to inherently reciprocal verbs. There are basically two types of such verbs: (A) *verbs indicating reciprocity in their lexical meaning* that might undergo the derivation of reflexive verbs indicating reciprocity (see Section 2.2.2) and (B) *non-reflexive verbs that imply reciprocity* of some of their arguments *in their lexical meaning* (e.g., *soupeřit* “to fight”, *sousedit* “to neighbor”).

In the case of (A), in addition to the derivation of reflexive reciprocal verbs (with the change of an accusative or dative complement into a prepositional group (see Section 2.2.2 and example (10), here repeated as (23)), the verbs can undergo “standard” reciprocal derivation, as in example (24a).

Here the question arises whether reciprocal constructions are derived from the non-reflexive *potkat* “to meet” (23a) or the reflexive verb *potkat se* “to meet” (23b). From the theoretical point of view, this question remains still open. The reflexive *se* in these constructions can be seen as the reflexive pronoun (as it can be replaced by its non-clitic variant *sebe* (24b)); however, the possible haplology of the morpheme *se* (Petkevič, 2013) makes the interpretation complicated and the theoretical interpretation of the derivation of these reciprocal constructions can differ, see also (Panevová, 2007).

For the practical implementation in the lexicon, we propose to derive the reciprocal constructions as in (24a) from the non-reflexive verbs (*potkat* for this case), not from its reflexive counterpart (*potkat se*) since this proposal allows us to use a single derivational rule for both types of verbs allowing for reciprocity (Section 3.2.1 and 3.2.2A).

(23a) *Petr potkal Marii.*
“Peter – met – Mary$_{\text{Pat-acc}}$.”
Eng. Peter met Mary.

(23b) *Petr se potkal s Marií.*
“Peter – SE$_{\text{morph}}$ – met – with Mary$_{\text{PAT-s+instr}}$.”
Eng. Peter met with Mary.

(24a) *Petr a Marie se potkali (navzájem).*
“Peter and Mary – SE$_{\text{pron-acc}}$ – met (each other)”
Eng. Peter and Mary met (each other).

(24b) *Petr a Marie potkali sebe navzájem i další přátele.*
“Peter and Mary met – SEBE$_{\text{pron-acc}}$ – each other – and other friends.”
Eng. Peter and Mary met (each other) as well as other friends.

In the case of (B), *non-reflexive verbs*, changes in the valency structure (i.e., the "multiplication" of subject) are sufficient markers of reciprocity, and the reflexive pronoun is not present in the reciprocal structures (25).

(25) *Týmy ČR soupeří o postup do finále. (= tým s týmem // mezi sebou/spolu)*
Eng. The teams of the Czech Republic fight for the finals. (= each team with other teams)

6.3 Reciprocity in the Lexicon

Reciprocity of arguments is described in the data component of the lexicon in a special attribute -rcp providing the list of the arguments that can enter the symmetrical relation. Changes in the valency structure of verbs (including the use of lexical means for expressing reciprocity) are regular enough to be captured by formal rules. These rules are stored in the grammar component of the lexicon and they make it possible to automatically derive valency frames underlying reciprocal constructions, see (Kettnerová et al., 2012b).

7 Czech Morpheme *se* as a Component of a Reflexive Verb Form

Finally, the reflexive *se* can represent a freestanding *morpheme* that *is a component of a reflexive verb form* in two types of diatheses in Czech: (i) the deagentive diathesis (Section 4.1) and (ii) the dispositional diathesis (Section 4.2). Diatheses are relations between syntactic structures of a verb which differ in the grammatical category of voice, i.e., they are associated with specific morphological meanings of a verb.

In Czech, five specific morphological meanings are determined: passive, deagentive, resultative, dispositional, and recipient-passive meanings (Panevová et al., in print). The surface structure of a verb with active voice is considered to be the unmarked member of a diathesis, whereas the structure with the given verb characterized by some of the five above given meanings constitutes its marked member. Whereas the passive, resultative and recipient-passive meanings of verbs are formed by the auxiliary verbs *být* (passive and resultative d.), *mít* (resultative d.), and *dostat* (recipient-passive d.), respectively, plus past participle of a lexical verb, the deagentive and dispositional diatheses are associated with the reflexive verb form. This form is constituted by the active form of a verb and the freestanding morpheme *se*.

8 Deagentive Diathesis

Marked members of the deagentive diathesis prototypically imply agentive ACTor of the event expressed by the verb; however, the ACTor is never expressed in the surface structure. The use of the deagentive meaning of a verb results in specific changes in its valency structure. In Czech, deagentive meaning can be applied to both transitive and intransitive verbs; the reflexive verb form is limited to 3rd person.[7] For a transitive verb, (i) ACTor is shifted from the subject position (expressed in the nominative) and (ii) the subject is filled with the argument of the verb corresponding to the direct object in the unmarked construction (prototypically an accusative object), as in (26). For an intransitive verb, the shift of ACTor away from the subject position results in a subjectless surface structure in which the verb has prototypical form of 3rd sg neutrum, as in (27).

(26a) *Dělníci opravují silnici.*
“workers$_{\text{ACT-Subj-nom}}$ – repair – the road$_{\text{PAT-Dobj-acc}}$.”
Eng. Workers repair the road.

(26b) *Silnice se opravuje.*
“the road$_{\text{PAT-Subj-nom}}$ – SE$_{\text{morph}}$ – repairs.”
Eng. The road is being repaired.

(27a) *Lidé na večírku tančili.*
“People$_{\text{ACT-Subj-nom}}$ – at the party – danced.”
Eng. People danced at the party.

(27b) *Na večírku se tančilo.*
“at the party – SE$_{\text{morph}}$ – danced$_{\text{3rd-sg-neutr}}$.”
Eng. At the party, there was some dancing.

9 Dispositional Diathesis

As in the case of the deagentive diathesis, the marked members of the dispositional diathesis indicate human ACTor that is shifted from the subject position (in the nominative). This position is filled by the argument corresponding to the direct object position of a transitive verb, see example (28); in the case of an intransitive verb, dispositional meaning of the verb results in a subjectless surface structure (with the 3rd sg neutrum verb form), see example (29). In contrast to the deagentive diathesis, ACTor can be optionally expressed as an indirect object expressed by the dative. The marked members of

[7] The status of several constructions with 2nd person (i) and 1st person (ii) is rather unclear, as they can be interpreted as either deagentive (with the grammatical morpheme *se*) (i), or reflexive (with the reflexive pronoun *se*) (ii).

(i) *Odsuzujete se k pěti letům vězení.*
“Sentence$_{\text{2nd-pl-masc/fem}}$ _ SE$_{\text{morph/pron}}$ _ to five year‘s imprisonment.”
Eng. You are sentenced to five year‘s imprisonment. // You sentence yourself to five year‘s imprisonment.

(ii) *Léčím se u doktora Nováka.*
“Treat$_{\text{1st-sg}}$ _ SE$_{\text{morph/pron}}$ -_ at the doctor Novák.”

dispositional diathesis are characterized by the presence of evaluative adverbs; thus if ACTor is expressed in the surface structure, it can be interpreted as an evaluator, see examples (28) and (29).

(28a) *Petr četl tuto knihu.*

“Peter$_{\text{ACT-Subj-nom}}$ – read – this book$_{\text{PAT-Obj-acc}}$.”

Eng. Peter read this book.

(28b) *Tato kniha se (Petrovi) dobře četla.*

“this book$_{\text{PAT-Subj-nom}}$ – SE$_{\text{morph}}$ – (Peter$_{\text{ACT-IObj-dat}}$) – well – read.”

Eng. This book read well.

(29a) *Já jsem tam spal.*

“I$_{\text{ACT-Subj-nom}}$ – am – there – slept.”

Eng. I slept there.

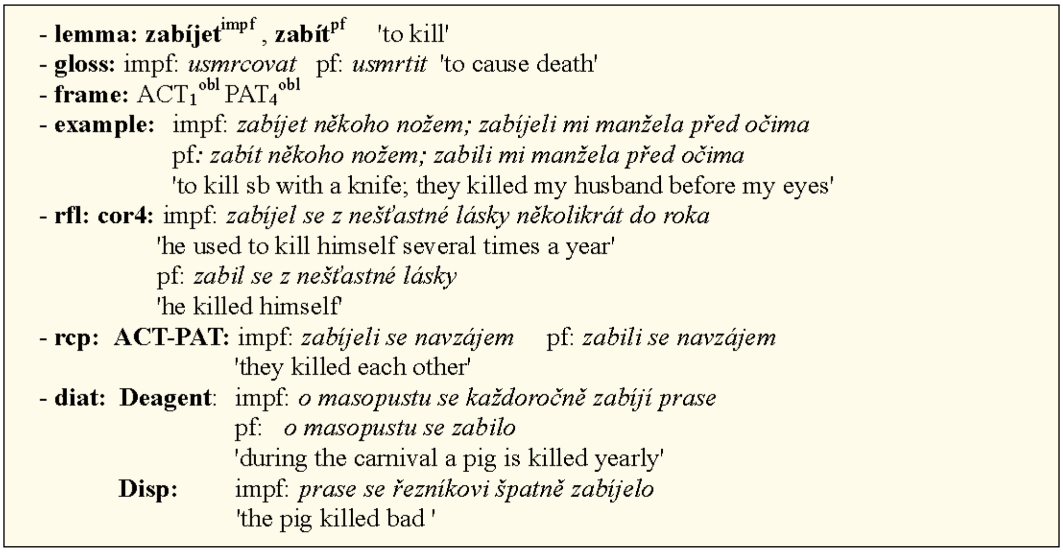
(29b) *Spalo se (mi) tam dobře.*

“slept$_{\text{3rd-sg-neutr}}$ – SE$_{\text{morph}}$ – (me$_{\text{ACT-IObj-dat}}$) – there – well.”

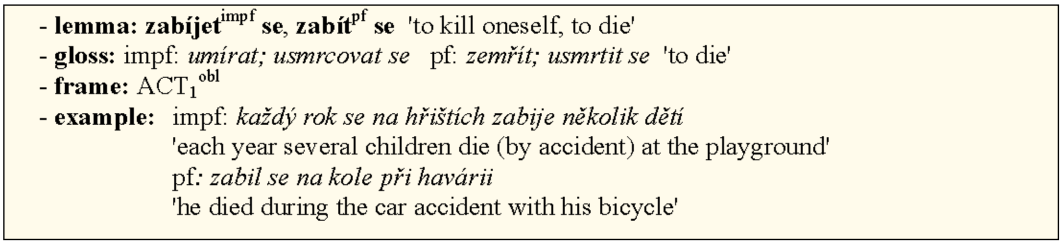
Eng. I slept well there.

10 Representation in the Lexicon

The changes in the valency structure of verbs in both deagentive and dispositional diatheses involve changes in morphemic forms of the arguments affected by surface shifts. These changes are regular enough to be captured by formal rules, which are stored in the grammar component of the lexicon. In the data component, only valency frames corresponding to the unmarked (active) uses of a verb are recorded. The optional attribute -diat is attached to each relevant valency frame in the data component (see Figure 1); it provides the information on applicability of the specific morphological meaning(s). On the basis of the formal rules (see Figure 2), the valency frames corresponding to marked structures of diatheses can be automatically derived (Kettnerová et al., 2012b).



- **lemma: zabíjet**^{impf} , **zabít**^{pf} 'to kill'
- **gloss:** impf: *usmrcovat* pf: *usmrtit* 'to cause death'
- **frame:** ACT$_1^{obl}$ PAT$_4^{obl}$
- **example:** impf: *zabíjet někoho nožem; zabíjeli mi manžela před očima*
  pf*: zabít někoho nožem; zabili mi manžela před očima*
  'to kill sb with a knife; they killed my husband before my eyes'
- **rfl: cor4:** impf: *zabíjel se z nešťastné lásky několikrát do roka*
  'he used to kill himself several times a year'
  pf: *zabil se z nešťastné lásky*
  'he killed himself'
- **rcp: ACT-PAT:** impf: *zabíjeli se navzájem* pf: *zabili se navzájem*
  'they killed each other'
- **diat: Deagent**: impf: *o masopustu se každoročně zabíjí prase*
  pf: *o masopustu se zabilo*
  'during the carnival a pig is killed yearly'
  Disp: impf: *prase se řezníkovi špatně zabíjelo*
  'the pig killed bad '

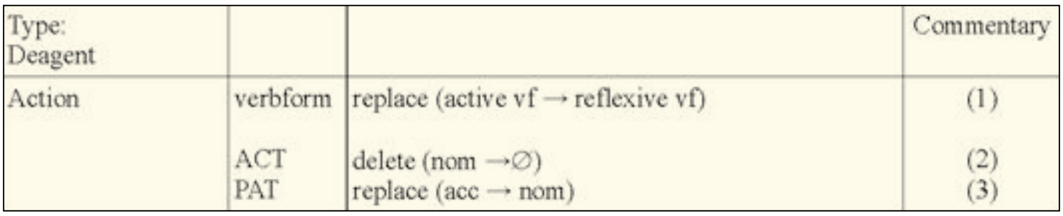


- **lemma: zabíjet^{impf} se, zabít^{pf} se** 'to kill oneself, to die'
- **gloss:** impf: *umírat; usmrcovat se* pf: *zemřít; usmrtit se* 'to die'
- **frame:** ACT_1^{obl}
- **example:** impf: *každý rok se na hřištích zabije několik dětí*
  'each year several children die (by accident) at the playground'
  pf*: zabil se na kole při havárii*
  'he died during the car accident with his bicycle'

Figure 1: Example of two lexical units of two lexemes *zabíjet*^{impf} , *zabít*^{pf} “to kill” and *zabíjet*^{impf} *se, zabít*^{pf} se “to kill oneself, to die” in the data component of the Valency Lexicon of Czech Verbs, VALLEX.

11 Conclusion

We have discussed the possibility of a lexicographic representation of Czech reflexive verbs. We have shown that reflexivity should be described in different ways, depending on the function of the reflexive morphemes *se* and *si*: (i) Reflexive tantum verbs and derived reflexive verbs (where *se*/*si* is a part of a verb lemma) are stored as separate verb lexemes (represented by separate verb lemmas) in the data component of the lexicon. (ii) The possibility of a verb to be used in reflexive constructions (in the narrow sense) and in reciprocal constructions (where *se* is the personal pronoun coreferring to the subject) is marked in the special attributes -rfl and -rcp, respectively, assigned to relevant lexical units in the data component. For reciprocal constructions, formal rules stored in the grammar component make it possible to automatically derive respective valency frames. (iii) Similarly, the possibility of a verb to undergo deagentive or dispositional diathesis (where *se* is part of the verb form) is marked in the attribute -diat assigned to each relevant lexical unit in the data component; formal rules stored in the grammar component enable the derivation of the valency frames underlying the marked members of the diatheses.



Type: Deagent			Commentary
Action	verbform	replace (active vf → reflexive vf)	(1)
	ACT	delete (nom →∅)	(2)
	PAT	replace (acc → nom)	(3)

Commentary:

(1) The verb form changes from active to reflexive (adding the reflexive *se*).

(2) ACTor cannot be expressed in the surface syntactic structure.

(3) The morphemic expression of PATient changes from the accusative into thenominative (and shifts to the subject position).

Figure 2: Example of a formal rule in the grammar component of the Valency Lexicon of Czech Verbs, VALLEX – the rule describing the deagentive diathesis.

12 References

Dokulil, M. (1986). *Mluvnice češtiny I.* Praha: Academia.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V. & Pajas, P. (2003). PDT-VALLEX: Creating a Large-Coverage Valency Lexicon for Treebank Annotation. In Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 15-15 November. Vaxjö, Sweden, pp. 57-68.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J. & Mikulová, M. (2006). Prague Dependency Treebank 2.0. Philadelphia, PA: Linguistic Data Consortium. LDC2006T01.

Hajičová, E., Panevová, J. & Sgall, P. (1985,1986,1987). Coreference in the Grammar and in the Text. Part I. In *The Prague Bulletin of Mathematical Linguistics*, 44, pp. 3-22. Part II. In *The Prague Bulletin of Mathematical Linguistics*, 46, pp. 1-11. Part III. In *The Prague Bulletin of Mathematical Linguistics,* 48, pp. 3-12.

Kettnerová, V., Lopatková, M. & Bejček, E. (2012a). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In *Proceedings of the 15th Euralex International Congress 2012, 7-11 August 2012.* University of Oslo, Norway, pp. 434-443.

Kettnerová, V., Lopatková, M. & Urešová, Z. (2012b). The Rule-Based Approach to Czech Grammaticalized Alternations. In *Proceedings of the 15th International Conference Text, Speech, Dialogue 2012, 3-7 September 2012.* Masaryk University, Czech republic, pp. 158-165.

Lenci, A., Lapesa, G. & Bonansinga, G. (2012) LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of LREC 2012*, pp. 3712- 3718.

Lopatková, M., Žabokrtský, Z., & Kettnerová, V. (2008). *Valenční slovník českých sloves.* Praha: Nakladatelství Karolinum.

König, E., Gast, V. (2008). Reciprocals and Reflexives: Theoretical and Typological Explorations. Berlin, New York: Mouton de Gruyter.

Nedjalkov, V. (2007). *Typology of Reciprocal Constructions.* Amsterdam: Benjamins.

Oliva, K. (2001). Reflexe reflexivity reflexive. In *Slovo a slovesnost*, 57, pp. 200-207.

Oliva. K. (2003). Linguistics-based PoS-tagging of Czech: disambiguation of *se* as a test. In *Contributions of the 4th European Conference on Formal Description of Slavic Languages, 28-30 November 2001.* Postdam University, Germany, pp. 299-314.

Panevová, J. (1994). Valency Frames and the Meaning of the Sentence. In P.A. Luelsdorff (ed.) *The Prague School of Structural and Functional Linguistics.* Amsterdam, Philadelphia: John Benjamins Publishing Company, pp. 223-243.

Panevová, J. (1999). Česká reciproční zájmena a slovesná valence. In *Slovo a slovesnost*, 60, pp. 269-275.

Panevová, J. (2007). Znovu o reciprocitě. In *Slovo a slovesnost*, 68, pp. 91-100.

Panevová, J., Mikulová, M. (2007). On Reciprocity. In *The Prague Bulletin of Mathematical Linguistics*, 87, pp. 27-40.

Panevová et al. (in print). *Mluvnice současné češtiny. Část 2: Syntax češtiny na základě anotovaného korpusu.* Praha: Nakladatelství Karolinum.

Petkevič, V. (2013). Formal (Morpho)Syntactic Properties of Reflexive Particles *se, si* as Free Morphemes in Contemporary Czech. In *Proceedings of the 7th International Conference 2013, 13-15 November 2013.* Slovenská akadémia vied, Slovakia, pp. 206-216.

Renau, I., Battaner, P. (2012). Using CPA to Represent Spanish pronominal Verbs in a Learner's Dictionary. In *Proceedings of the 15th Euralex International Congress 2012, 7-11 August 2012.* University of Oslo, Norway, pp. 350-361.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, Ch.R. & Scheffczyk, J. (2010) *FrameNet II: Extended Theory and Practice.* https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

Sgall, P., Hajičová, E. & Panevová, J. (1986). *The Meaning of the Sentence in Its Pragmatic and Semantic Aspects.* Dordrecht: Reidel.

Skoumalová, H. (2001). *Czech syntactic lexicon.* PhD thesis. Charles University, Prague, Czech Republic.

Žabokrtský, Z., Lopatková, M. (2007) Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. In *The Prague Bulletin of Mathematical Linguistics*, 87, pp. 41-60.

Acknowledgements

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. P406/12/0557. This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

Polysemous Models of Words and Their Representation in a Dictionary Entry

Tinatin Margalitadze
Lexicographic Centre at Ivane Javakhishvili Tbilisi State University
tinatin@margaliti.ge

Abstract

The paper deals with one of the universal models of polysemous adjectives and verbs, namely one-dimensional model and examines the ways of its representation in a dictionary entry. Polysemy is connected with the human perception and cognition of the world. It is determined by the process of perceiving not only particular objects and phenomena, but also the similarities existing between them or seen as such by members of the given language community. It is also connected with the ability of the language to reflect the new, yet un-cognized objects and phenomena by means of their associations and relations with already known, cognized objects and phenomena, i.e. to translate diversity of the world into linguistic unity. This makes polysemy an extremely interesting linguistic phenomenon but also leads to controversies concerning the interpretation of different issues connected with it. The paper touches upon some debatable issues connected with polysemy, such as: boundaries between senses, meaning and context, the role of context in the process of realization of meanings of polysemous words, meanings and sub-meanings, sense-numbering in a dictionary entry, etc. The paper also discusses some peculiarities of lexical meanings of adjectives and verbs.

Keywords: one-dimensional; general semantic component; subsume

1 Introduction

As early as in the 1st century AD, Marcus Fabius Quintilian, a theoretician in oratorical skills, explains in his textbook on rhetoric *Institutio Oratoria* the concepts of metaphor and metonymy, which are important mechanisms of semantic changes and development of transferred meanings of words. The compilation of comprehensive explanatory dictionaries in the 17th century, first in Italy (1612) and then in France (1694), in the 18th century's England (*A Dictionary of the English Language* by Samuel Johnson, 1755), made a significant contribution to the description and study of semantics as a field of knowledge, and particularly to that of polysemy. In the early 20th century, a German linguist Hermann Paul distinguishes between usual (*usuelle Bedeutung*) and occasional meanings (*okkasionelle Bedeutung*), drawing the attention of linguists to context as an important tool for the realization of polysemous meanings of a word (Paul 1920: 75). Since then, much has been written on polysemy and context and the issue is still under discussion. Although it is not the aim of the present paper to discuss all problems connected with polysemy, still we cannot avoid touching upon some important issues, namely: What is the meaning of a polysemous word? Does it exist on the systemic level of language, or the meaning of a polysemous word is entirely determined by the context in which the word is used? Does a polysemous word have one abstract / general meaning, activated differently in different contexts, or does its structure represent a set of interconnected lexical units (LU) arranged into a single whole by means of various semantic relationships?

These issues aroused bitter controversy not only in 1950s and 1960s (Firth 1968; Antal 1963; Ulmann 1964 and others). Over the last 10-15 years, many scholars yet again discuss these issues and, as it seems, the answer is still not unequivocal (Stock 1984; Atkins 1993; Kilgariff 1997; Hanks 2000; Rundell 2002; Kosem 2008; Trap-Jensen 2010 and others). "I don't believe in word senses", states Sue Atkins (Atkins 1993). According to P. Hanks, a word does not have separate meanings, but rather a set of meaning potentials, which may be activated in a particular context (Hanks 2000). M. Rundell distinguishes between senses of polysemous words with clearly distinct meanings which, in this respect, "conform quite well to the conventional dictionary model, and much fuzzier *meaning-clusters*, where a basic semantic core is elaborated, in real text, in a variety of ways" (Rundell 2002: 148). More and more questions arise on how to present word meanings? How to find boundaries between senses? Lumping or Splitting? How to deal with the issue of meaning clines? How are meanings of polysemous words activated in a context?

Such discussions led some scholars even to the questioning of lexicographic practice of division of words' meanings into senses, particularly for NLP purposes (Kilgariff 1997; Hanks 2000).

The study of above issues is important not only from the theoretical point of view, in order better to perceive the phenomenon of polysemy, but also from the viewpoint of representing a polysemous word in a dictionary entry, also for NLP purposes. The present paper expounds on one of the universal models of polysemous words studied in adjectives (Margalitadze 1982), later in verbs and nouns (Margalitadze 2006), namely the one-dimensional model. The study of this model shed some light on specific debatable issues of polysemy.

2 General Semantic Component and Subseme

As mentioned above, one-dimensional model is characteristic of adjectives and verbs. For the description of the present model two semantic components of a word's lexical meaning are to be introduced: general semantic component, and subseme. 'General semantic component' (GS) denotes that semantic component of word, which serves as the basis for the development of a number of LUs of polysemous adjectives and verbs.

By the term 'subseme' we denote the semantic component which serves to differentiate LUs of a polysemous word. The subseme concretizes the abstract meaning of GS in each particular LU, thus activating meanings of polysemous adjectives and verbs[1].

In order to illustrate the GS (marked in blue colour in examples given below) and the subseme (marked in red colour in examples given below), let us examine the adjective STRAIGHT. LUs of the adjective concerned are based on the GS – being free from deviation / bending /.

(1) STRAIGHT

1. Direct, not crooked (a straight street, a straight edge, a straight railway line) –

   being free from deviation / bending / *in direction*

2. Erect, not crooked or stooping (a straight back) –

   being free from deviation / bending / *in deportment*

3. Direct, shortest, uninterrupted (a straight flight, a straight road, a straight path) –

   being free from deviation *in course*

4. Straightforward, frank, open (a straight answer, a straight question, straight talks) –

   being free from deviation *in truth, openness, frankness*

5. Fair, virtuous, honest (a straight woman) –

   being free from deviation *in dealings / rectitude /*

6. Consistent, logical, clear (a straight thinker) –

   being free from deviation *in some method*

7. Conventional; respectable (she looked straight, a straight play) –

   being free from deviation *from conventional, accepted, traditional behaviour / norms / views*

and so on.

2.1 Systemic Context and Subseme

The first example represents the following meanings of the adjective straight:

1. Direct, not crooked;
2. Erect, not crooked or stooping;
3. Direct; shortest; uninterrupted;
4. Straightforward, frank, open;
5. Fair, virtuous, honest;
6. Consistent, logical, clear;
7. Conventional; respectable

[1] The term 'general semantic component' is not an adequate translation of the Georgian name given to this semantic component, which is well rendered by its Russian equivalent *сквозная сема* 'skvoznaia sema' (literally 'through-going semantic component'). *Сквозная сема* 'skvoznaia sema' has its origin in the theatrical term introduced by Constantin Stanislavski – "through-going action" (*сквозное действие*). In fact, the term aptly expresses the essence of the semantic component identified in the semantic structure of verbs and adjectives.

and so on.

Adjectives and verbs denote an important logical category of 'feature': a feature, quality of an object – in adjectives, and a feature of an action or a state in verbs. The feature denoted by these words is singled out from different classes of objects or actions / states. Accordingly, adjectives and verbs contain in their meanings various expressions of this or that feature in objects or actions of different classes. As a result of this, along with each meaning of an adjective or a verb, there is constantly implied a systemic context denoting components of one and the same category of the objective reality.

Systemic context of the first meaning of the adjective STRAIGHT are the nouns denoting objects having linear shape (*a street, an edge, a railway line, etc*). Out of this group of nouns, the GS 'being free from deviation / bending /' singles out the common seme '*direction*', which becomes included in the semantic structure of the given meaning of the adjective as its component, and concretizes the general meaning of the GS 'being free from deviation / bending /' – 'being free from deviation / bending / in *direction*', thus enabling the realization of the meaning 'straight'.

Systemic context of the second meaning of STRAIGHT are the nouns denoting back, shoulders, *etc.* Out of this group of nouns, the GS 'being free from deviation / bending /' selects the common seme '*deportment*' which becomes included in the semantic structure of the given meaning of the adjective as its component, and concretizes the general meaning of the GS – 'being free from deviation / bending / in *deportment*', thus enabling the realization of the meaning 'erect, not crooked or stooping'.

Systemic context of the third meaning of STRAIGHT are the nouns denoting flight, road, way, *etc*. Out of this group of nouns, the GS 'being free from deviation' selects the common seme 'course', which concretizes the general meaning of the GS – 'being free from deviation in *course*', thus enabling the realization of the meaning 'direct, shortest'.

Systemic context of the fourth meaning of STRAIGHT are the nouns denoting conversation, question, answer, *etc*. Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*truth, openness*', which concretizes the general meaning of the GS – 'being free from deviation in *truth, openness*', thus enabling the realization of the meaning 'straightforward, frank, open'.

Systemic context of the fifth meaning of STRAIGHT are the nouns denoting human beings. Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*dealings, rectitude*', which concretizes the general meaning of the GS – 'being free from deviation in *dealings, rectitude*', thus enabling the realization of the meaning 'fair, honest'.

Systemic context of the sixth meaning of STRAIGHT are the nouns denoting thinking, thinker, *etc*. Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*method*', which concretizes the general meaning of the GS – 'being free from deviation in *some method*', thus enabling the realization of the meaning 'consistent, logical', and so on.

Thus GS 'being free from deviation / bending /' singles out the common semes – subsemes from systemic contexts: '*direction*' from nouns, denoting objects with linear shape; '*deportment*', from nouns denoting back, shoulders, *etc.*; '*course*' from nouns denoting flight, road, way, and so on. These subsemes enter the semantic structure of LUs of the adjective STRAIGHT as their component, and concretize the abstract meaning of the GS – 'being free from deviation / bending / in *direction*', in *deportment*', in *course*', etc, thus activating the realization of LUs: 'straight', 'erect, not crooked or stooping', 'direct, shortest' and so on.

GS and subseme can be illustrated on the example of other adjectives and verbs.

(2) LUs of the adjective CROOKED are based on the GS – having deviation in / from /.

1. not straight, bent, twisted (crooked streets, a crooked road, a crooked blade) –

having deviation *in direction*

2. deformed; bent (an aged man with a crooked frame, yellow and crooked teeth) –

having deviation *from normal form*

3. dishonest, not straightforward (crooked politicians, crooked dealings) –

having deviation *in rectitude*

4. fraudulent; illegal (crooked business, crooked business deal) –

having deviation *from legal frame*

and so on.

(3) LUs of the adjective LOW are based on the GS – being below the average level.

1. of small upward extent (a low wall, a low hill) –

being below the average level *in upward extension*

2. not elevated in position (low bridges, Low Countries) –

being below the average level *in elevation from the ground or some other downward limit*

3. not tall, short (a low man, a man of low stature) –

being below the average level *in statute*

4. not high in amount (low price, low wages) –

being below the average level *in amount*

5. deficient in degree of intensity (low redness, low colour) –

being below the average level *in degree of intensity*

6. not loud (low voice, low laugh) –

being below the average level *in volume*

7. of humble rank, position (low birth, low life) –

being below the average level *in social rank*

8. wanting in elevation, of inferior quality (low art, low standard) –

being below the average level *in quality*

9. wanting in decent breeding, vulgar, coarse (low person, low company) –

being below the average level *in social "respectability"*

and so on.

(4) LUs of the verb ESCAPE are based on the GS – breaking / getting / away from.

1. to get away, to get free (to escape from prison, to escape from the army) –

breaking / getting / away from *(physical) confinement*

2. to avoid or retreat from the realities of life (to escape reality) –

breaking / getting / away from *unpleasant realities of life*

3. to avoid or elude an evil that threatens (to escape poverty, to escape punishment) –

breaking / getting / away from *misfortune of any kind*

4. to avoid psychological problems (to escape television addiction) –

breaking / getting / away from *mental / psychological / problems*

5. to elude notice or recollection (to escape one's mind, to escape smb.'s eyes) –

breaking / getting / away from *notice / mental grasp /*

6. to leak from a container (of a gas, liquid, etc) –

(as if) breaking / getting / away from *some confining envelope or enclosure*

and so on.

2.2 Mechanism of the Interrelation between Adjective / Verb and Noun

Figure 1 demonstrates the underlying mechanism of the interrelation between adjective or verb and noun in their semi-automated syntagms, where dotted line represents GS, ellipse represents a LU of a polysemous adjective or verb, and a circle – the systemic context of the given LU. GS acts from adjective or verb to noun (A / V → N). GS determines the choice of nouns, from them selects common feature – subseme, which concretizes its abstract meaning. Whereas subseme acts in the opposite direction, from noun to adjective / verb (N → A / V). Subseme enters the semantic structure of adjectives or verbs, concretizes meaning of GS and activates individual LUs of polysemous adjectives and verbs.

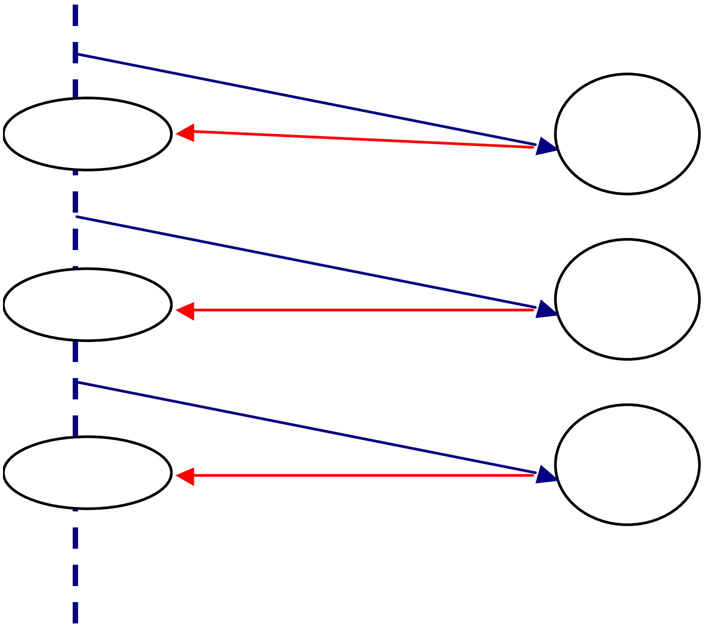


Figure 1: Semi-automated Syntagms of Adjectives / Verbs and Nouns.

- GS is a generative semantic component, providing the basis for the development of a number of LUs;
- GS is a common semantic component, a kind of semantic "thread" uniting several LUs of a polysemous word;
- GS is an integrating seme, by means of which a polysemous verb and adjective can function as one word;
- GS is generated in paradigmatic, that is, vertical section. Its existence is revealed through the comparison of several meanings of a polysemous word;
- By its structural and semantic status, GS is more abstract, than differential and potential semes making up the lexical meaning of a word, as far as it governs several LUs of verb and adjective;
- GS inherently implies the idea of the classes of objects, which may be characterized by a given verb or adjective. Accordingly, it motivates or blocks the selection of a systemic context, wherewith a given verb or adjective may liaise.

GS differs from common semantic component, archeseme or hyperseme by being a generative semantic component. Not only is it the common semantic component of several LUs, but it is also the basis of the generation of polysemous meanings of verb and adjective and does govern them.

- Subseme is a differential seme, on the basis of which a concrete LU of verb and adjective is generated. Like GS, we regard subseme as a generative seme. While GS generates several LUs of polysemous verb and adjective, subseme serves as the basis for the creation of one specific LU;
- Subseme is singled out from an entire class of objects, which is represented by a definite group of nouns. Consequently, it implies the idea of the given class of objects and, accordingly, that of the definite area of denotation;
- Upon the syntagmatic axis, subseme is generated in the course of interrelationship between verb and adjective on the one hand, and semantic structures of noun on the other hand;
- As a result of the existence of subseme, for each LU in the semantic structure of verb and adjective there is generated a systemic context united by the given subseme.

3 One-Dimensional Model

'One-dimensional' are termed such polysemous verbs and adjectives, all LUs of which are generated on the basis of a single GS (see Figure 2).

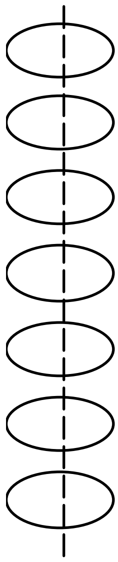


Figure 2: One-Dimensional Model.

All the examples discussed above represent one-dimensional models. One-dimensional is not the only model of polysemous adjectives and verbs. There are more models described for these parts of speech (Margalitadze 1982; 2006) but this model is quite universal and many adjectives and verbs develop their LUs on the basis of one GS. Below are given more examples of one-dimensional adjectives and verbs.

(5) Polysemous meanings of the English verb 'to break' are based on the GS – 'destroying / violating / the completeness, wholeness, continuity'. Its subsemes in different LUs may be *a bone, a plate, a surface, skin, a performance, a lecture, spiritual, moral or financial state, silence*, etc.

(6) The verb 'to kill' has the GS – 'depriving of some essential quality', which is concretized by different subsemes in different LUs: *life, vitality, activity, feeling, desire,* etc.

(7) The GS of the verb 'to erupt' is – 'bursting forth from natural or artificial limits', concretized by the following subsemes: *volcano, water, fire, air, soldiers,* etc.

(8) The one-dimensional adjective 'dull' has its polysemous meanings generated on the basis of the GS – 'wanting some essential quality'. Its subsemes in different LUs are: *wit, sensibility or keenness of perception, motion or action, vivacity or cheerfulness, colour, intensity*, etc.

(9) All polysemous meanings of the adjective 'small' are developed on the basis of the GS – 'being less than average' (being less than average in *size*, in *statute*, in *number*, in *duration*, in *importance*, in *amount*, in *rank or condition*, in *scale*, etc).

(10) All polysemous meanings of the adjective 'great' are developed on the basis of the GS – 'being more than average' (being more than average in *size*, in *number*, in *duration*, in *importance*, in *rank or condition*, in *scale,* etc).

(11) All polysemous meanings of the adjective 'high' are developed on the basis of the GS – 'being above the average level' (being above the average level in *upward extension*, in *elevation from the ground or some other downward limit*, in *stature*, in *amount*, in *social rank*, etc).

4 Discussion

As it has been shown by the above analysis, one-dimensional words have meanings of equal status. What may seem an abstract / general meaning, meaning potential or a semantic core activated differently in real contexts, is in fact a semantic component, the GS which is very general, very abstract, as far as it contains in itself the idea of those classes of words wherefrom it is singled out. Whenever the GS is concretized by a subseme, there appears an individual meaning of a polysemous word, an individual LU. Subseme shows the boundaries between senses of a polysemous adjective and verb. Thus, one-dimensional model has an extremely abstract GS and meanings of equal status.

The role of context must be mentioned specifically. Context is always necessary for the actualization of the meanings of a polysemous word, but the GS, as we have seen, is an active semantic component that can select or block nouns in question. Context can not trigger any meaning of one-dimensional adjective or verb which is not present in their semantic structure. Consequently, context reveals what is already present in the semantic structure of adjective or verb, it does not motivate meaning, it actualizes existing meanings. This shows the relative independence of adjectives and verbs within the language system.

Feature does not exist independently in the objective reality. It is present inside object and is unimaginable without the latter. However, the feature translated into a linguistic category appears as a separate category, as that of adjective and verb, thus acquiring a different linguistic status. Within the system of language, feature acquires relative autonomy, which results in complex interrelations between adjective and verb and their systemic context in their lexical syntagms. The role of systemic context in the process of realization of meanings of verb and adjective consists in conveying particular information in the form of subseme to the semantic structure of verb and adjective. This information concretizes general meaning of the GS and breaks it up into concrete variants, thus enabling the differentiation and realization of separate LUs of verb and adjective. On the other hand, verb and adjective, as independent parts of speech, contain such semes within the semantic structure of their meaning, which not only generate a number of LUs, but also determine the selection of a definite, rather than any noun. Within the language system, the interaction between objects / actions and their features is formed on a completely different level of generalization.

5 One-Dimensional Words in a Dictionary Entry

Dictionaries of the English language give different interpretation to the polysemous words of the described model and represent them accordingly in a dictionary entry. E.g. MEDAL, Oxford Dictionary of English treat the following meaning of 'escape' – leak from a container (of a gas, liquid, etc) – as a full-fledged meaning of this verb, while OED, Shorter Oxford English Dictionary, Webster's Third New International Dictionary interpret it as a sub-meaning of 'escape'. Likewise, LU – avoid capture, punishment, or something unwelcome – are meanings according to Shorter Oxford English Dictionary, Webster's Third New International Dictionary, while MEDAL, Oxford Dictionary of English treat them as sub-meanings, meaning-clusters. Such examples may be sited *ad infinitum*.

LUs of one-dimensional model have equal status. Semantic relationships between LUs, generated on the basis of GS, are equipollent and not that of dependence. Consequently, LUs should be represented as full-fledged meanings in a dictionary entry and they should be numbered in the same manner by Arabic numerals. GS may be given at the beginning of an entry as a general description of the feature. Each LU should be supplied with its systemic context, i.e. with nouns denoting one and the same category of the objective reality. Below are given some examples of entries.

(12) Straight *adjective*

[being free from deviation / bending]

1. Direct, not crooked (*used with nouns denoting objects, having linear shape*);

    a straight street, a straight edge, a straight railway line;

2. Erect, not crooked or stooping (*used with nouns back, shoulders, etc*)

    a straight back;

3. Direct, shortest, uninterrupted (*used with nouns denoting travelling on foot or by other means*)

    a straight flight, a straight road, a straight path;

4. Straightforward, frank, open (*used with nouns denoting talking*)

    a straight answer, a straight question, straight talks;

5. Fair, virtuous, honest (*used with nouns denoting human beings*)

    a straight woman;

6. Consistent, logical, clear (*used with nouns denoting thinking*)

    a straight thinker;

and so on.

(13) Escape *verb*

[breaking / getting / away from usually smth. unpleasant]

1. to get away, to get free (*used with nouns denoting places of physical confinement*)

    to escape from prison, to escape from the army;

2. to avoid or retreat from the realities of life (*used with nouns denoting unpleasant realities of life*)

    to escape reality;

3. to avoid or elude an evil that threatens (*used with nouns denoting any misfortune*)

    to escape poverty, to escape punishment;

4. to avoid psychological problems (*used with nouns denoting different addictions*)

    to escape television addiction;

5. to leak from a container (*used with nouns denoting gas, liquid, etc*)

and so on.

Another alternative of a dictionary entry may be GS+subseme descriptions in each LU of the one-dimensional adjective or verb (see example 14).

(14) Straight *adjective*

[being free from deviation / bending]

1. Direct, not crooked (*used with nouns denoting objects, having linear shape*);

    a straight street, a straight edge, a straight railway line;

[being free from deviation / bending / in direction]

2. Erect, not crooked or stooping (*used with nouns back, shoulders, etc*)

    a straight back;

[being free from deviation / bending / in deportment]

3. Direct, shortest, uninterrupted (*used with nouns denoting travelling on foot or by other means*)

    a straight flight, a straight road, a straight path;

[being free from deviation / bending / in course]

4. Straightforward, frank, open (*used with nouns denoting talking*)

    a straight answer, a straight question, straight talks;

[being free from deviation / bending / in truth, frankness]

5. Fair, virtuous, honest (*used with nouns denoting human beings*)

    a straight woman;

[being free from deviation / bending / in dealings / rectitude /]

6. Consistent, logical, clear (*used with nouns denoting thinking*)

    a straight thinker;

[being free from deviation / bending / in some method]

and so on.

6 Conclusion

The study of the deep structure of interrelation between adjectives / verbs and nouns in their semi-automated syntagms has revealed the active generating semantic component – GS in the semantic structure of polysemous adjectives and verbs. On the one hand, GS generates several LUs and governs them, on the other hand, GS inherently has the knowledge of the classes of objects wherefrom it is singled out, thus motivating or blocking the selection of nouns wherewith adjectives and verbs may liaise. GS selects the subseme from the systemic context, which enters the semantic structure of LU and is present there. As a result of this, along with each meaning of adjective or verb, there is constantly implied a systemic context denoting components of one and the same category of the objective reality.

The interrelation between GS and subseme and the presence of subseme in the semantic structure of LU indicates that context reveals existing meaning of adjective and verb and does not motivate it. Subsemes mark the boundaries between senses of polysemous adjectives and verbs.

One-dimensional adjectives and verbs have meanings of equal status, which should be numbered in the same manner in a dictionary entry, as full-fledged meanings.

GS may be given in a dictionary entry, as a general description of the feature, expressed by adjective and verb (see examples 12, 13).

Each LU should be supplied with its systemic context, specifying the group of nouns used with the respective LU (see examples 12, 13).

Unlike identifying words such as nouns which, depicting objects and phenomena, comprise multiple semantic components in the semantic structure of their meanings, verbs and adjectives denote feature. Accordingly, their lexical meaning is "scarce" of semantic components and thus it is natural that polysemous structure of adjectives and verbs should be characterized by linear development and one feature, one semantic component should become the basis for the formation of multiple meanings.

7 References

Antal, L. (1963) *Questions of Meaning*. The Hague: Mouton Co.

Atkins, B.T.S. (1993) "Theoretical Lexicography and its relation to Dictionary-making". *Dictionaries* 14:4-43.
Firth, J.R. (1958) *Papers in Linguistics*. Oxford University Press.

Hanks, P. (2000) Do Word Meanings Exist? *Computers and the Humanities* 34: 205-215.

Kilgariff, A. (1997) I Don't Believe in Word Senses. *Computers and the Humanities* 31(2): 91-113.

Kosem, I. (2008). Dictionaries for University Students: A Real Deal or Merely a Marketing Ploy?
Proceedings of the XIII EURALEX International Congress. Barcelona.

Margalitadze, T. (1982) Strukturno-semanticheskaia Kharakteristika Mnogoznachnykh Prilagatel'nykh, kak
Nominativnykh Edinits v Sovremennom Angliiskom Iazyke. Candidate's Thesis. Tbilisi : Tbilisi University Press.

Margalitadze, T. (1982) The Main Models of the Semantic Structure of Adjectives in Modern English. In: *Bulletin of the*
Academy of Sciences of the Georgian SSR. 105, 3 : 181 – 184.

Margalitadze, T. (2006) *Meaning of a Word and Methods of its Research*. Tbilisi State University.

Paul, H. (1920) Prinzipien der Sprachgeschichte. Halle: Niemeyer.

Rundell, M. (2002) Good Old-fashioned Lexicography: Human Judgement and the Limits of Automation. *Lexicography*
and Natural Language Processing. EURALEX : 138-155.

Stock, P. (1984) Polysemy. *Lexeter '83 Proceedings.* Tübingen: Max Niemeyer.

Trap-Jensen, L. (2010) One, Two, Many: Customization and User Profiles in Internet Dictionaries. *Proceedings of the*
XIV Euralex International Congress. Fryske Akademy, Leeuwarden.

Ullmann, St. (1964) Semantics. An Introduction to the Science of Meaning. Oxford: Basil Blackwell.

Dictionaries:

Hanks P. et al (2005). *Oxford Dictionary of English.* Second Edition, Revised. Oxford University Press.

Macmillan English Dictionary for Advanced Learners (MEDAL). Accessed at: http://www.macmillandictionary.com [05.09.2013]

Oxford English Dictionary on Historical Principles (1989). Second edition on CD-ROM. Version 2.0. Oxford University Press.

Stevenson A. et al (2007). *Shorter Oxford English Dictionary*. Sixth Edition (SOED). Oxford University Press.

Webster's Third New International Dictionary (Unabridged). Merriam Webster Inc., 1981.

One Lexicological Theory, two Lexicographical Models and the Pragmatemes

Lena Papadopoulou
Hellenic Open University
papadopoulou.lena@gmail.com

Abstract

Generally little attention has been paid to pragmatics in most dictionaries. The present paper focuses on the concept of pragmatemes and their lexicographical treatment within the frame of Explanatory Combinatorial Lexicology. First, necessary preliminary notions closely related with pragmatemes are considered, by briefly reviewing the mel'čukian global model of human linguistic behaviour, linguistic sign and phrasemes typology. Second, the definition of pragmatemes, that is of phrasemes used in given extralinguistic situations, and the central acting part of the conceptual representation of the communicative situation are presented. Following, the structure of *Explanatory Combinatorial Dictionary* (ECD) and *PragμatLex* are outlined and an illustration of the Greek pragmatemes *Συγχαρητήρια* 'Congratulations' and *Συλλυπητήρια* 'Condolences' is provided within the simplified versions of ECD - *Dictionary of Collocations* and *Lexique actif du français*- and the lexicographical model for pragmatemes *PragμatLex*. Finally, we conclude our paper with a brief discussion on which lexicographical approach is preferable.

Keywords: Meaning Text Theory; phraseology; pragmatemes

1 Introduction

Pragmatics generally is an area of great importance. However, it is relatively poorly treated in the majority of dictionaries, so there is scope for work on this subject. The concept of pragmatemes, that is expressions that are used in specific extralinguistic situations, and the developed models for their lexicographical treatment represent a significant step towards addressing that challenge.

This paper aims to present two lexicographical models in which pragmatemes can be processed; the ECD and the PragμatLex. To do so, first our theoretical framework (Meaning⇔Text Theory) will be set, then the definition of pragmatemes will be provided and, following, the dictionaries' structure will be described and illustrated by processing the Greek pragmateme *Συγχαρητήρια* 'Congratulations' and its antonym *Συλλυπητήρια* 'Condolences'. Finally, criteria for model selection will be proposed.

2 Preliminary notions

Our work is framed within Meaning⇔Text Theory (MTT), the main aim of which is to build models of natural languages (among others, Mel'čuk 1988a; Mel'čuk 1997; Mel'čuk 2001b; Polguère 1998; Milićević 2001; Milićević 2006; Kahane 2001).

Concept-Sound Model (CSM) is the global model of human linguistic behavior which is developed within MTT:

{WORLD}⇔{SemR$_i$}⇔{SPhonR$_j$}⇔{LINGUISTIC SOUNDS}

Figure 1: Concept-Sound Model (Mel'čuk 2012: 170-181).

Conceptics, Meaning-Text Model (MTM) and Phonetics/Graphics are the three major models of CSM which represent the production of an utterance. First, Conceptics model captures the construction of a semantic representation (SemR) based on the conceptual representation (ConceptR) of the given extralinguistic situation (SIT). Second, MTM describes the construction of the Phonological Representation (PhonR) of the given SemR. The third model - Phonetics/Graphics- represents the construction of the corresponding sound/letter string for the given PhonR.

A typology of linguistic signs has been established within MTT based on the transitions between these three models and the applied restrictions. A simple linguistic sign within MTT corresponds to the triplet of X= <'X'; /X/; ΣX>, where 'X' is the signified, /X/ the signifier and ΣX the combinatorial properties of the linguistic sign X. Simple linguistic signs are combined into complex linguistic signs. The notions of unrestrictedness and regularity are implied by such combination, that is freedom in the selection of meanings and lexical units and the compositionality, respectively.

Free phrases are complex linguistic signs whose signified and signifier are constructed both unrestrictedly and regularly, while on the contrary phrasemes, or non free phrases, are not. Phrasemes are classified into semantic phrasemes and pragmatic phrasemes, or pragmetemes, (Mel'čuk 1995; Mel'čuk 1998). On the one hand, pragmatemes are restrictedly constructed by the ConceptR(SIT). On the other hand, the signified 'X' of a semantic phraseme is unrestrictedly constructed by the ConceptR(SIT) but its signifier /X/ is constrained by the selected SemR. Semantic phrasemes are non compositional and they are categorized into three types on the basis of their semantic opacity: (i) full idioms, (ii) semi-idioms, or collocations, and (iii) quasi-idioms.

3 Pragmatemes

Pragmatemes are compositional phrasemes whose signified is restrictedly constructed by the Conceptual Representation of the given extralinguistic situation (Mel'čuk 1998). Blanco (to appear) points out that this definition concerns the prototypical pragmatemes. On the one hand, a pragmateme can

be both constrained by the ConceptR and the SemR, i.e. the idiom/pragmateme break a leg [to wish good luck to actors and musicians before they go on stage to perform]. On the other hand, a lexeme whose signified is bound by the Concept(SIT) is considered as a pragmateme, i.e. Congratulations.
The ConceptR(SIT) plays the lead role in pragmatemes definition Although, Mel'čuk recognizes the inherent difficulties in defining the extralinguistic reality, he proposes that ConceptR is based on three main models (2001a, p. 90): (i) the speaker´s model, (ii) the speaker´s model of the addressee and (iii) the situation´s model. The ConceptR(SIT) of the pragmateme break a leg will be based on that 'I am addressing to an actor which is going on stage to perform. I wish (s)he will have a successful presentation. If I were (s)he I would like to be encouraged. I will wish him/her good luck, as I should do.' (speaker´s model), '(S)he is thinking that (s)he is going on stage to perform, that (s)he is stressed, that (s)he expects to be encouraged' (speaker´s model of the addressee) and on that 'the speaker wants to encourage a performer before going on the stage by wishing him good luck ' (situation´s model).

4 Lexicological processing of pragmatemes

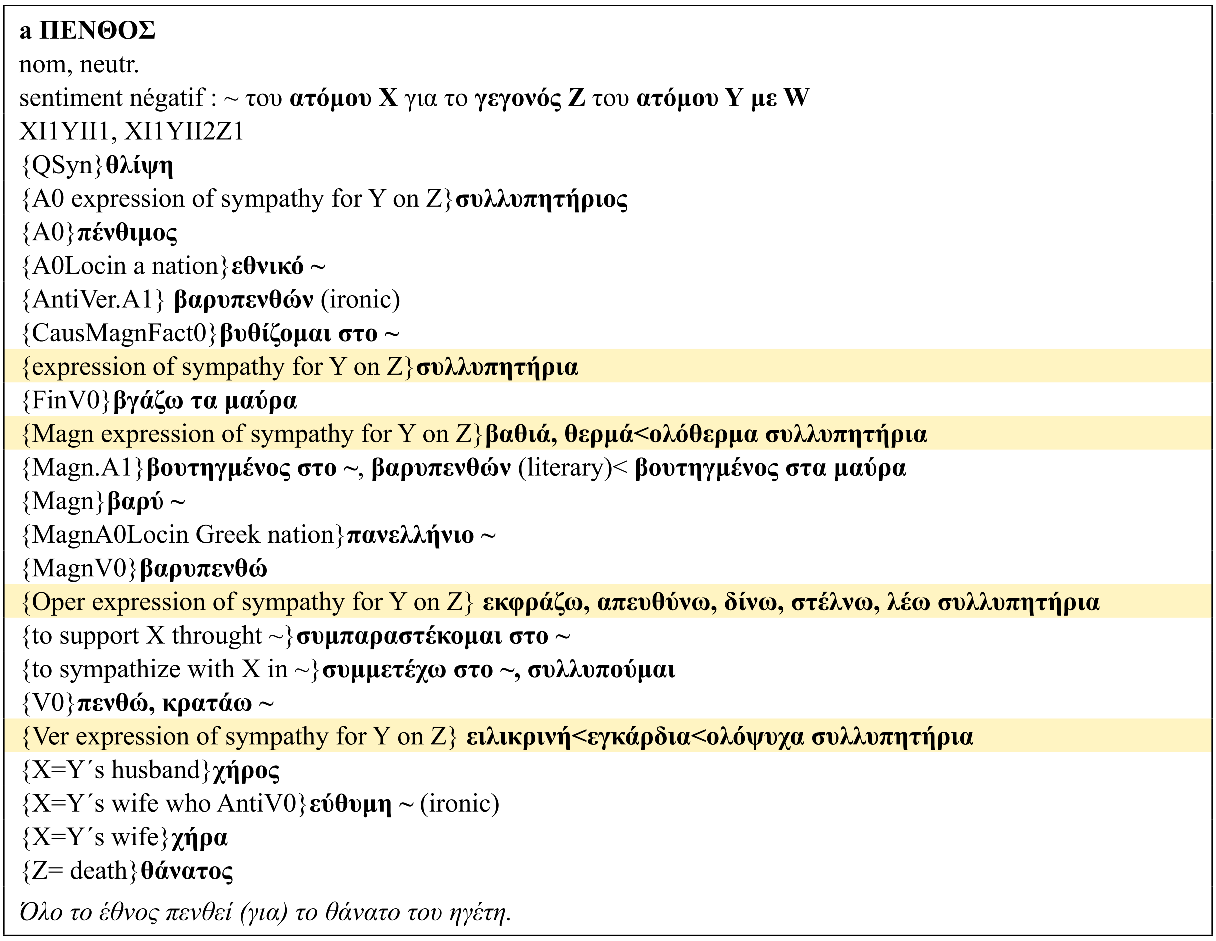
Once the pragmatemes have been defined and before moving to the presentation of the two lexicological models for pragmatemes, which are both framed within Explanatory Combinatorial Lexicology, some preliminary remarks upon pragmatemes have to been made. Although pragmatemes are linguistic signs, they are not considered to be LUs, because they dispose of an internal argumental structure, so as they are ordered within the keyword(s) that phraseologically bind(s) them, that is within the LU(s) that can define the SIT of the pragmateme. It has to been also pointed out that pragmatemes and specifically their SIT is described by non standard lexical functions (LFs) (Mel'čuk, 1995); (Blanco, 2010).

4.1 Explanatory Combinatorial Dictionaries

Explanatory Combinatorial Lexicology is developed within the MTT (among others, Mel'čuk & Zholkovsky 1984; Mel'čuk 1988b; Mel'čuk 1995; Mel'čuk, Clas, & Polguère 1995; Mel'čuk 2006b; Mel'čuk & Polguère 2007) and Explanatory Combinatorial dictionaries (ECD) are compiled within it.
ECDs are highly formal theoretical lexicons, whose entries are exhaustively described on the basis of explicitness and consistency. The macrostructure of an ECD is structured by super-entries, entries and sub-entries. Vocables constitute the super-entries, which are sets of lexical units (LUs) that share the same signifier and they are linked by a semantic bridge, LUs are the entries, which can correspond to lexemes, idioms or quasi-idioms, and collocations and pragmatemes are considered to be sub-entries. As far is microstructure is concerned, it is structured in four zones: (i) the semantic, (ii) the phonological/graphematic zone, the (iii) syntactics zone and (iv) illustrative zone.

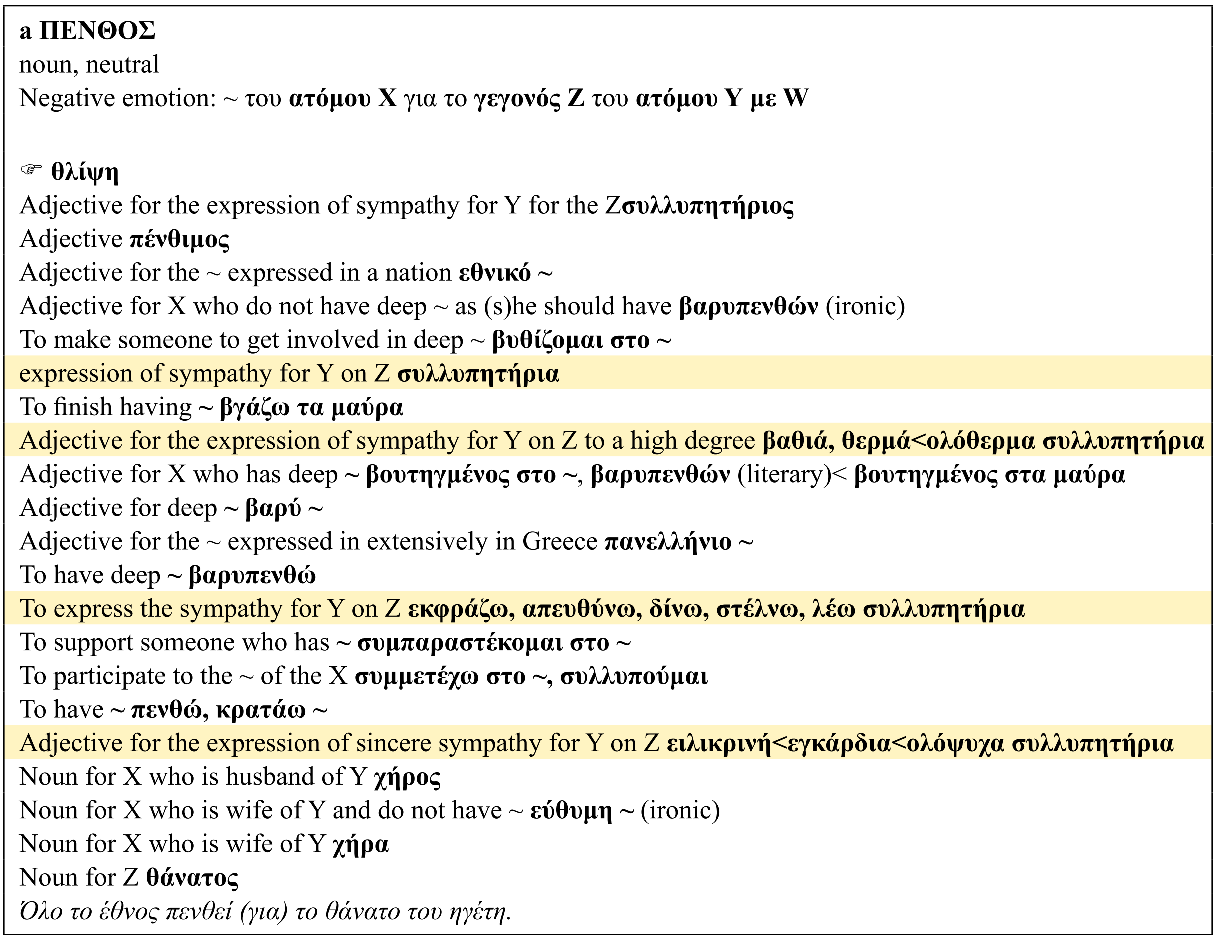
Due to the theoretical basis of ECD and the its subsequent high lengthiness, the Dictionary of Collocations (DiCo) and Lexique actif du français (LAF) have been developed as simplified versions. DiCo (Dictionary of Collocations) is the formalized version of the purely "theoretical" ECD. DiCo is sort of a "simplified" and more formalized ECD and in which the lexical units are structured as a series of eight main fields: (i) Name of the unit, (ii) grammatical properties, (iii) semantic formula, (iv) government pattern, (v) synonyms, (vi) semantic derivations and collocations, (vii) examples and (viii) full idioms that include the LU (Polguère 2000: 519) and LAF is the "popularized" version of the ECD which attempt to bridge the gap between "theoretical" and "commercial" lexicography with regard to explanatory combinatorial lexicology in order to be as much as possible accessible to a public of non-specialists (Polguère 2000: 522-3).

Following an illustrative example of processing the Greek pragmateme *Συλλυπητήρια* 'Condolences' (Figure 2 and 3) (Papadopoulou to appear) within the keyword-LU, respectively:



a ΠΕΝΘΟΣ
nom, neutr.
sentiment négatif : ~ του **ατόμου X** για το **γεγονός Z** του **ατόμου Y με W**
XI1YII1, XI1YII2Z1
{QSyn}**θλίψη**
{A0 expression of sympathy for Y on Z}**συλλυπητήριος**
{A0}**πένθιμος**
{A0Locin a nation}**εθνικό ~**
{AntiVer.A1} **βαρυπενθών** (ironic)
{CausMagnFact0}**βυθίζομαι στο ~**
{expression of sympathy for Y on Z}**συλλυπητήρια**
{FinV0}**βγάζω τα μαύρα**
{Magn expression of sympathy for Y on Z}**βαθιά, θερμά<ολόθερμα συλλυπητήρια**
{Magn.A1}**βουτηγμένος στο ~**, **βαρυπενθών** (literary)< **βουτηγμένος στα μαύρα**
{Magn}**βαρύ ~**
{MagnA0Locin Greek nation}**πανελλήνιο ~**
{MagnV0}**βαρυπενθώ**
{Oper expression of sympathy for Y on Z} **εκφράζω, απευθύνω, δίνω, στέλνω, λέω συλλυπητήρια**
{to support X throught ~}**συμπαραστέκομαι στο ~**
{to sympathize with X in ~}**συμμετέχω στο ~, συλλυπούμαι**
{V0}**πενθώ, κρατάω ~**
{Ver expression of sympathy for Y on Z} **ειλικρινή<εγκάρδια<ολόψυχα συλλυπητήρια**
{X=Y´s husband}**χήρος**
{X=Y´s wife who AntiV0}**εύθυμη ~** (ironic)
{X=Y´s wife}**χήρα**
{Z= death}**θάνατος**
Όλο το έθνος πενθεί (για) το θάνατο του ηγέτη.

Figure 2 LU a ΠΕΝΘΟΣ in DiCo (Papadopoulou to appear).



a ΠΕΝΘΟΣ

noun, neutral

Negative emotion: ~ του **ατόμου X** για το **γεγονός Z** του **ατόμου Y με W**

☞ **θλίψη**

Adjective for the expression of sympathy for Y for the Z**συλλυπητήριος**

Adjective **πένθιμος**

Adjective for the ~ expressed in a nation **εθνικό ~**

Adjective for X who do not have deep ~ as (s)he should have **βαρυπενθών** (ironic)

To make someone to get involved in deep ~ **βυθίζομαι στο ~**

expression of sympathy for Y on Z **συλλυπητήρια**

To finish having **~ βγάζω τα μαύρα**

Adjective for the expression of sympathy for Y on Z to a high degree **βαθιά, θερμά<ολόθερμα συλλυπητήρια**

Adjective for X who has deep **~ βουτηγμένος στο ~**, **βαρυπενθών** (literary)< **βουτηγμένος στα μαύρα**

Adjective for deep **~ βαρύ ~**

Adjective for the ~ expressed in extensively in Greece **πανελλήνιο ~**

To have deep **~ βαρυπενθώ**

To express the sympathy for Y on Z **εκφράζω, απευθύνω, δίνω, στέλνω, λέω συλλυπητήρια**

To support someone who has **~ συμπαραστέκομαι στο ~**

To participate to the ~ of the X **συμμετέχω στο ~, συλλυπούμαι**

To have **~ πενθώ, κρατάω ~**

Adjective for the expression of sincere sympathy for Y on Z **ειλικρινή<εγκάρδια<ολόψυχα συλλυπητήρια**

Noun for X who is husband of Y **χήρος**

Noun for X who is wife of Y and do not have ~ **εύθυμη ~** (ironic)

Noun for X who is wife of Y **χήρα**

Noun for Z **θάνατος**

Όλο το έθνος πενθεί (για) το θάνατο του ηγέτη.

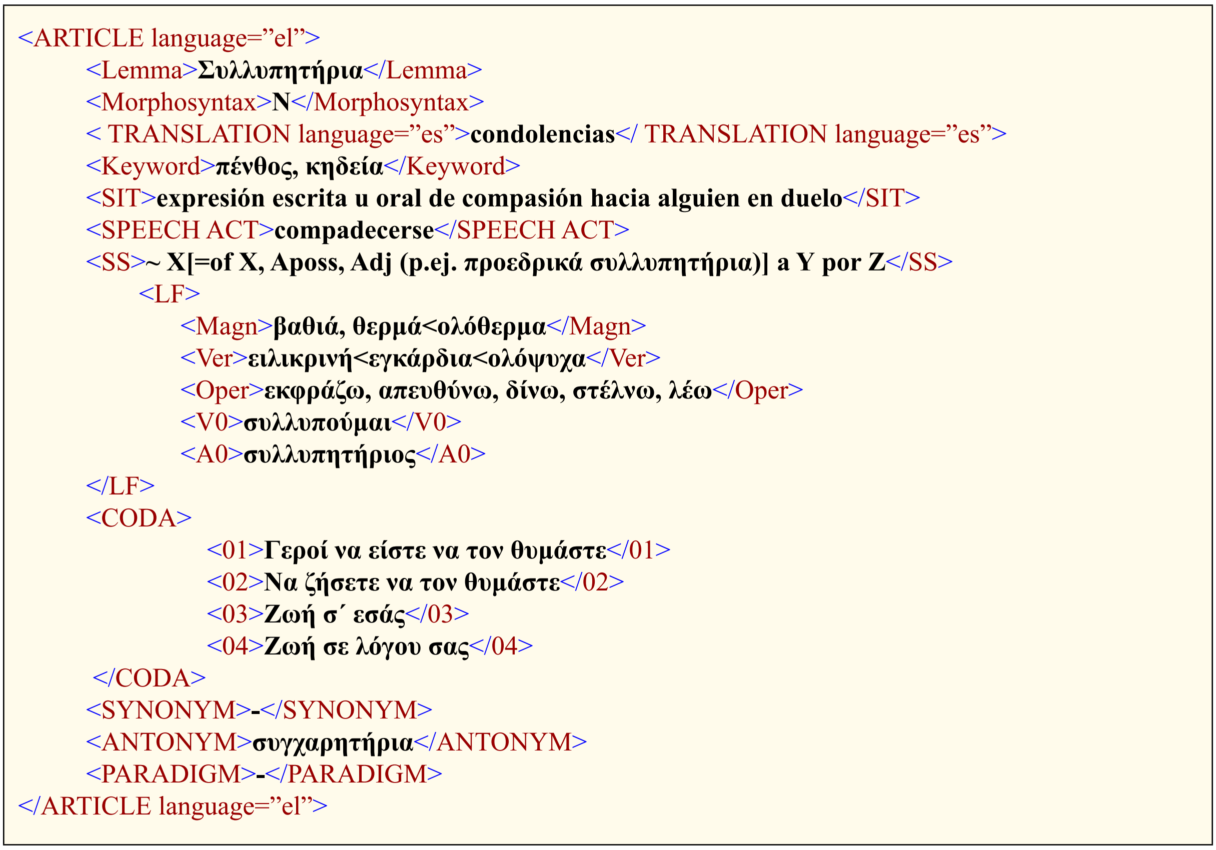
Figure 3 LU aΠΕΝΘΟΣ in LAF (Papadopoulou to appear).

4.2 PragμatLex

Please note that there must always be at least two level 2 and level 3 headings if you need to use these in your paper (e.g. at least 4.1 and 4.2 Blanco (2010; to appear$_a$; to appear$_b$) obviously based on MTT and recognizing the lack of dictionaries of ECD type in the majority of languages proposed the PragμatLex, which is designated as a lexicographical model for the processing of pragmatemes. PragμatLex is highly formal and it provides an exhaustive description for each pragmateme, which is structured in thirteen fields. It is worth pointing out that PragμatLex is written in XML in order to be applicable to NLP systems.

PragμatLex is a dictionary of monolingual coordinated dictionary type (Blanco 2001), considering that the translation equivalence of each pragmateme is provided linearly according to the overall microstructure information, so as the description of pragmatemes is enterassigned within the language indication: (<ARTICLE language=" ">**description of pragmatemes**</ARTICLE language =" ">. First, the canonical form of the pragmateme is indicated Lemma>**canonical form of the pragmateme**</Lemma>. Second, the morphosyntax of the pragmateme is annotated based on the six deep-syntactic parts of speech (Mel'čuk 2006a), Third, the translation equivalence is provided in the target language according to the corresponding structure of the L2 PragμatLex. Following, the LU-keyword(s), the definition of the SIT, the performing Speech act, the semantic structure and the lexical functions of the pragmateme are indicated. Afterwards, the coda, that is pragmatemes extensions which with no semantic addition complement the pragmatemes, the synonyms and the antonyms of the pragmateme and, finally, the decomposition of the local grammar that may the lemma disposes.

In the following figures the pragmatemes *Συλλυπητήρια-Condolencias* 'Condolences' (Papadopoulou to appear) and *Συγχαρητήρια-Felicidades* 'Congratulations' are shown within PragμatLex in Greek and Spanish language:



```
<ARTICLE language=”el”>
        <Lemma>Συλλυπητήρια</Lemma>
        <Morphosyntax>N</Morphosyntax>
        < TRANSLATION language=”es”>condolencias</ TRANSLATION language=”es”>
        <Keyword>πένθος, κηδεία</Keyword>
        <SIT>expresión escrita u oral de compasión hacia alguien en duelo</SIT>
        <SPEECH ACT>compadecerse</SPEECH ACT>
        <SS>~ X[=of X, Aposs, Adj (p.ej. προεδρικά συλλυπητήρια)] a Y por Z</SS>
              <LF>
                   <Magn>βαθιά, θερμά<ολόθερμα</Magn>
                   <Ver>ειλικρινή<εγκάρδια<ολόψυχα</Ver>
                   <Oper>εκφράζω, απευθύνω, δίνω, στέλνω, λέω</Oper>
                   <V0>συλλυπούμαι</V0>
                   <A0>συλλυπητήριος</A0>
        </LF>
        <CODA>
                      <01>Γεροί να είστε να τον θυμάστε</01>
                      <02>Να ζήσετε να τον θυμάστε</02>
                      <03>Ζωή σ΄ εσάς</03>
                      <04>Ζωή σε λόγου σας</04>
         </CODA>
        <SYNONYM>-</SYNONYM>
        <ANTONYM>συγχαρητήρια</ANTONYM>
        <PARADIGM>-</PARADIGM>
</ARTICLE language=”el”>
```

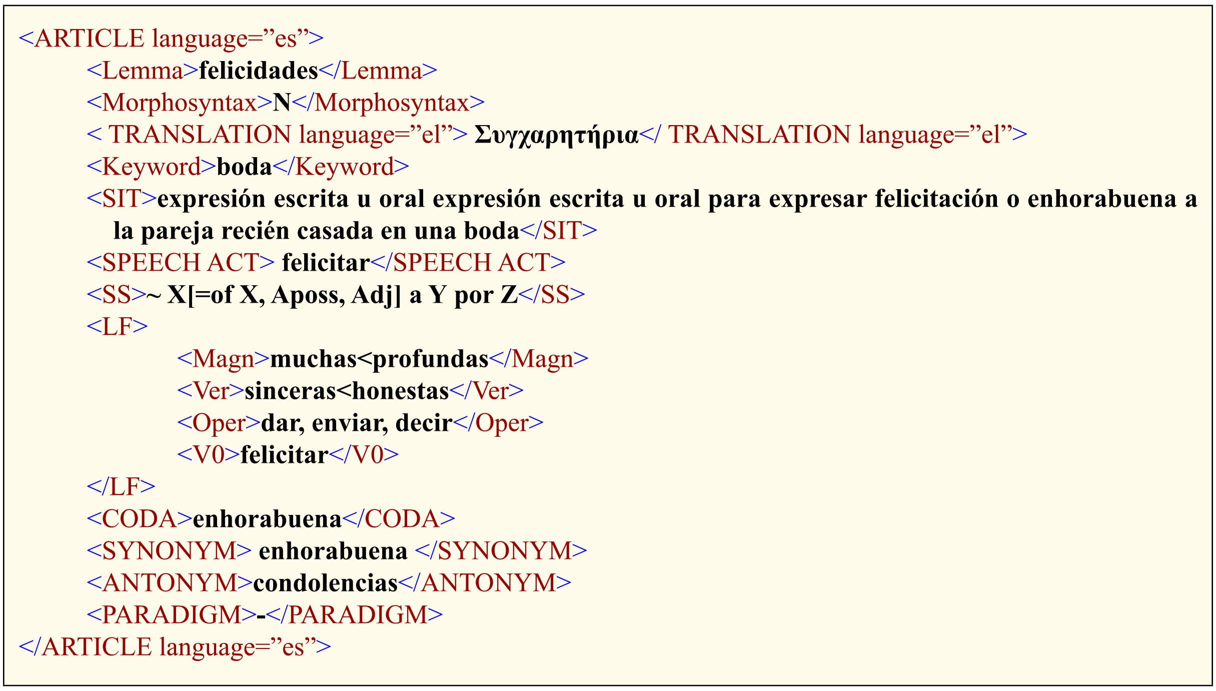
Figure 4: *Συλλυπητήρια* in PragμatLex (Papadopoulou to appear).

```
<ARTICLE language=”es”>
        <Lemma>condolencias</Lemma>
        <Morphosyntax>N</Morphosyntax>
        < TRANSLATION language=”el”> συλλυπητήρια</ TRANSLATION language=”el”>
        <Keyword>duelo, funeral</Keyword>
        <SIT>expresión escrita u oral de compasión hacia alguien en duelo</SIT>
        <SPEECH ACT>compadecerse</SPEECH ACT>
        <SS>~ X[=of X, Aposs, Adj (p.ej. Condolencias presidenciales)] a Y por Z</SS>
        <LF>
                  <Magn>mayores<profundas</Magn>
                  <Ver>sentidas<sinceras<cordiales</Ver>
                  <Oper>expresar, dar, manifestar, enviar,</Oper>
                  <V0>condoler</V0>
        </LF>
        <CODA>Siempre lo recordaremos</CODA>
        <SYNONYM>pésame</SYNONYM>
        <ANTONYM>congratulaciones, felicitaciones</ANTONYM>
        <PARADIGM>-</PARADIGM>
</ARTICLE language=”es”>
```

Figure 5: ***Condolencias*** **in PragμatLex (Papadopoulou to appear).**

```
<ARTICLE language=”el”>
        <Lemma> Συγχαρητήρια</Lemma>
        <Morphosyntax>N</Morphosyntax>
        < TRANSLATION language=”es”>felicidades</ TRANSLATION language=”es”>
        <Keyword>γάμος</Keyword>
        <SIT>expresión escrita u oral para expresar felicitación o enhorabuena a la pareja recién casada
          en una boda </SIT>
        <SPEECH ACT>felicitar</SPEECH ACT>
        <SS>~ X[=of X, Aposs, Adj] a Y por Z</SS>
             <LF>
                  <Magn> πολλά<θερμά<ολόθερμα</Magn>
                  <Ver>ειλικρινή<εγκάρδια<ολόψυχα</Ver>
                  <Oper>εκφράζω, απευθύνω, δίνω, στέλνω, λέω</Oper>
                  <V0>συγχαίρω</V0>
                  <A0>συγχαρητήριος</A0>
        </LF>
        <CODA>
                     <01>να ζήσετε</01>
                     <02>και καλούς απογόνους</02>
        </CODA>
        <SYNONYM>-</SYNONYM>
        <ANTONYM>συλλυπητήρια</ANTONYM>
        <PARADIGM>-</PARADIGM>
</ARTICLE language=”el”>
```

Figure 6: ***Συγχαρητήρια*** **in PragμatLex.**



```
<ARTICLE language=”es”>
        <Lemma>felicidades</Lemma>
        <Morphosyntax>N</Morphosyntax>
        < TRANSLATION language=”el”> Συγχαρητήρια</ TRANSLATION language=”el”>
        <Keyword>boda</Keyword>
        <SIT>expresión escrita u oral expresión escrita u oral para expresar felicitación o enhorabuena a
           la pareja recién casada en una boda</SIT>
        <SPEECH ACT> felicitar</SPEECH ACT>
        <SS>~ X[=of X, Aposs, Adj] a Y por Z</SS>
        <LF>
                   <Magn>muchas<profundas</Magn>
                   <Ver>sinceras<honestas</Ver>
                   <Oper>dar, enviar, decir</Oper>
                   <V0>felicitar</V0>
        </LF>
        <CODA>enhorabuena</CODA>
        <SYNONYM> enhorabuena </SYNONYM>
        <ANTONYM>condolencias</ANTONYM>
        <PARADIGM>-</PARADIGM>
</ARTICLE language=”es”>
```

Figure 7: *Felicidades* in PragμatLex.

5 ECD or PragμatLex?

ECD, or PragμatLex, that is NOT the question, as two different types of dictionaries are concerned, which are based on the same lexicological theory, yet; ECD is a dictionary of lexical units and PragμatLex is a dictionary of pragmatemes. However, we could answer the question in three different rounds from three different points of view.

First, the ideal lexicographical treatment of pragmatemes is within ECD, given that ECD’s structure provides a global description of pragmatemes within their semantic frame (lexical units’ links). However, there are no available complete ECD dictionaries for all languages. Second, PragμatLex’ structure is proper for pragmatemes processing, as it focuses only on pragmatemes. Third, we propose a parallel processing of ECD and PragμatLex, that is the lexicographer elaborates pragmatemes which are associated with a keyword within PragμatLex and (s)he incorporates these data into the structure of ECD, i.e. the information of pragmateme *condolences* can be introduced as subentries into the structure of the lexical unit MOURNING (Papadopoulou, to appear) or *congratulations* into WEDDING.

6 References

Blanco, X. (2001). Dictionnaires électroniques et traduction automatique espagnol-français. In *Langages 143* (pp. 49-70).

Blanco, X. (to appearb). Équivalents de traduction pour les pragmatèmes dans la lexicographie bilingue Français-Espagnol.

Blanco, X. (2010). Los frasemas composicionales pragmáticos. In S. Mejri, & P. Mogorrón, *Opacité, Idiomaticité, Traduction.* Universitat d'Alacant.

Blanco, X. (to appear$_{a}$). Microstructure Évolutive pour un Dictionnaire de Pragmatemes. In *Actes des JournéesLexicologie, Lexicographie et Traduction.* Paris: AGence Universitaire de la Francophonie.

Kahane, S. (2001). *Grammaires de dépendance formelles et théorie Sens-Texte.* Tours: Tutoriel, TALN 2001.

Mel'čuk, I. (2006a). *Aspects of the Theory of Morphology.* Berlin/New York: Mouton de Gruyter.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A. P. Cowie, *Phraseology: Theory, Analysis, and Applications* (pp. 23-53). Oxford: Clarendon Press.

Mel'čuk, I. (2001a). *Communicative Organization in Natural Language. The Semantic-Communicative Structure of Sentences.* Amsterdam/Philadelphia: Benjamins.

Mel'čuk, I. (1993). *Cours de morphologie générale.* (Montréal — Paris: Les Presses de l'Université de Montréal — CNRS.

Mel'čuk, I. (1988a). *Dependency syntax: theory and practice.* Albany, NY: State Univ. of New York Press.

Mel'čuk, I. (2006b). Explanatory Combinatorial Dictionary. In G. Sica, *Open Problems in Linguistics and Lexicography* (pp. 225-355). Monza: Polimetrica.

Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner, *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37-102). Amsterdam/Philadelphia: Benjamins.

Mel'čuk, I. (2006a). Parties du discours et locutions. *Bulletin de la Société de Linguistique de Paris 101:1* , págs. 29-65.

Mel'čuk, I. (2006c). Parties du discours et locutions. In *Bulletin de la Société de Linguistique de Paris 101:1* (pp. 29-65). Paris.

Mel'čuk, I. (1995). Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder, *Idioms. Structural and Psychological Perspectives* (pp. 167-232). Hillsdale, N.J.—Hove, UK: Lawrence Erlbaum Associates.

Mel'čuk, I. (2013). Phraseology:its place in the language, in the dictionary, and in natural language processing. In Z. Gavriilidou, A. Efthymioy, E. Thomadaki, & P. Kambakis-Vougiouklis, *Selected Papers of the 10th I.C.G.L.* (pp. 62-67).

Mel'čuk, I. (1988b). Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria. In *International Journal of Lexicography, 1 : 3* (pp. 165-188).

Mel'čuk, I. (2001b). *Semantics and the Lexicon in Modern Linguistics.* Unpublished Article.

Mel'čuk, I. (2012). *Semantics. From Meaning to Text.* Amsterdam/Philadelphia: Benjamins Publishing Company.

Mel'čuk, I. (to appear). Tout ce que nous voulions savoir sur les phrasèmes, mais .... In *Cahiers de lexicologie, revue internationale de lexicologie et de lexicographie.* Paris: Classiques Garnier.

Mel'čuk, I. (1982). *Towards the Language of Linguistics.* München: Wilhelm Fink.

Mel'čuk, I. (1997). *Vers une linguistique Sens-Texte. Leçon inaugurale (given on Friday January 10th 1997).* Collège de France, Chaire internationale.

Mel'čuk, I., & Polguère, A. (2007). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations semantiques et collocations du français.* Bruxelles: De Boeck.

Mel'čuk, I., & Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian.* Vienna: Wiener Slawistischer Almanach.

Mel’čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., et al. (1988). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II.* Montréal: Les Presses de l’Université de Montréal.

Mel’čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., & Lessard, A. (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I.* Montréal: Les Presses de l’Université de Montréal.

Mel’čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., & Mantha, S. (1992). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III.* Montréal: Les Presses de l’Université de Montréal.

Mel’čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S., & Polguère, A. (1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV.* Montréal: Les Presses de l’Université de Montréal.

Mel’čuk, I., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Louvain-la-Neuve: Duculot.

Melchuk, I., & Zholkovsky, A. (1984). Explanatory Combinatorial Dictionary of Modern Russian. In *Sonderband 14.* Vienna: Wiener Slawistischer Almanach.

Milićević, J. (2001). A short guide to the Meaning-Text linguistic theory. In A. Gelbukh, *Intelligent Text Processing and Computational Linguistics.* Mexico: Colección en Ciencias de Computación, Fondo de Cultura Económica - IPN - UNAM.

Milićević, J. (2006). A short Guide to the Meaning-Text Linguistic Theory. *Journal of Koralex* (Vol. 8), pp. 187-233.

Papadopoulou, L. (to appear). “My deepest condolences”: Lexical functions of Greek pragmatemes [in a funeral]. In *Proceedings of the 11th International Conference on Greek Linguistics, 26 - 29 September 2013, Rhodes, Greece.*

Polguère, A. (1998). La théorie Sens-Texte. *Dialangue , Vol. 8-9,* pp. 9-30.

Polguère, A. (2000). Towards a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French. In *Proceedings of EURALEX 2000* (pp. 517-527). Stuttgart.

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

Anna Riccio
University of Naples "L'Orientale"
ariccio@unior.it

Abstract

The aim of this study is to investigate the semantics and syntax of verbs in French, Italian, and English medical discourse by exploring the relationship between verb semantics and argument realization. The verbs under consideration are common lexical units which have acquired the status of a term through their specialization of meaning, such as *affect*, *involve*, etc. Unlike terminological verbs (e.g. *keratinize* or *lyophilize*), they have a lower level of technicalness, and co-occur with arguments (usually terms) in syntagmatic units. The data are extracted from the parallel EMEA corpus including documents published by the European Medicines Agency. The description of the verbs is based on the theoretical model of Frame Semantics (Fillmore1977a-b, 1982, 1985; Fillmore and Atkins 1992) and on the FrameNet methodology (Ruppenhofer et al. 2010). The resultant analysis of the collected data reveals a sentence-level scenario (i.e., the Damaging frame) which groups together verbal forms which share similar syntactic and semantic valence patterns, both within and across languages.

Keywords: specialized verbs; medical domain; parallel corpus; intra-/cross-linguistic equivalents; Frame Semantics; FrameNet methodology

1 Introduction

The special status of the verb in terminological resources, rather than the noun, is an issue which has been widely discussed in literature over the last fifteen years. See, among others, Picht (1987), L'Homme (1995, 1998), Lorente and Bevilacqua (2000), Valente (2000), Costa and Silva (2004), and more recently De Vecchi and Estachy (2008), Tellier (2008), Pimentel (2012) and Pettersson (2013). Verbs, like nouns, tend to have particular usages within situational communication between experts of specific fields.

The initial stage of this study focuses on the analysis of certain verbs that can have an unusual significance, or a meaning which is specific to the medical field, such as *affect*, *involve*, *enhance*, etc. Unlike terminological verbs (e.g. *keratinize* or *lyophilize*), they have a lower level of technicalness.

The observation of the behaviour of verbal forms in a corpus of medical texts has been explored by Tellier (2008) and Pettersson (2013) in French. In this work, verbs are examined in French, Italian and

English. The data are extracted from the parallel (translation) EMEA corpus from the EMA (European Medicines Agency). The corpus-based analysis of specialized verb equivalents (lexical units which have the same meaning and usage intra- and cross-linguistically) may be useful for the elaboration of a multilingual terminological resource which covers the subject field of medicine. This could be useful for translators, the teaching of specialized translation and terminological or technical writers.

The description of the verbs in question is based on the theoretical model of Frame Semantics (Fillmore1977a-b, 1982, 1985; Fillmore and Atkins 1992) and on the FrameNet methodology (Ruppenhofer et al. 2010), because verbs are "frame-evoking" or "frame-bearing" words par excellence (Pimentel 2012: 5). Each specialized verb evokes a semantic frame representing a sentence-level scenario which groups together verbal forms that share similar syntactic and semantic valence patterns. The study also tests the hypothesis that semantic frames can function as "interlingual representations" in the organization of a multilingual lexicon (Boas 2005).

This paper is organized as follows. Section 2 provides a brief description of the research methodology: the instruments used for the data collection (2.1., 2.2.) and the theoretical model of Frame Semantics as well as the FrameNet methodology (2.3.). Section 3 illustrates the frame of the specialized verbs examined in this study (The Damaging frame) and their morphological and syntactic patterns. Section 4 follows with some concluding remarks.

2 Methodology

2.1 Corpus

The data are extracted from the multilingual parallel (translation) corpus EMEA from the European Medicines Agency (available in 22 European official languages). The corpus is made up of PDF documents which are representative of a genre of written medical discourse, specifically, package leaflets for medicinal products (Tiedemann 2009). The leaflets are specialized texts that make use of one of the different types of communication between experts and non-experts, such as doctor-patient interactions.

The corpus includes over 311,65 million tokens in all, 14,9 million of which are in French, 14,1 million in Italian, and 12,1 million in English. The corpus is available through the OPUS site (http://opus.lingfil.uu.se/) and can also be accessed through the Sketch Engine interface (Kilgarriff et al. 2004). The verbal items are collected and organized using the Sketch Engine to facilitate their quantitative and qualitative analysis.

Table 1 illustrates the verbal word-types and word-tokens in each language (i.e. French, Italian and English):

Language	Type-frequency	Token-frequency
French	1862	1,836,737
Italian	1855	1,256,154
English	1843	1,328,560

Table 1: Type and token frequency of verbs in EMEA.

The Type- / Token-frequency lists also include verbs which do not have any kind of specific value in medical discourse. Thus, specialized verbs have to be selected from the list of concordances for each language. The contexts which have been examined thus far show a large number of specialized verbs among the three languages. This article simply presents the preliminary results on 8 verbal lexical items (see table 2 below), which actually allow us to observe their special status within medical terminology and lexicology.

2.2 Data

The specialized verbs in question are those which Lorente (2000) calls *verbos fraseológicos* (Eng. 'phraseological verbs'), which are different from the *verbos terminológicos* (Eng. 'terminological verbs'). The former are predicative verb units that appear in specialized texts in order to express states, actions and processes. When isolated, their meaning is similar to the meaning of the verbs in non-specialized contexts, e.g. (Fr.) *administrer*, (It.) *somministrare*, (Eng.) *administer*. However, when they co-occur with arguments (usually terms) in syntagmatic units they acquire a specialized value. For example, we usually say (Fr.) *administrer un médicament*, (It.) *somministrare un farmaco*, (Eng.) *adminster a medicine*, but not (Fr.) *donner un medicament*, (It.) *dare un farmaco*, (Eng.) *give a medicine*, even if their respective meanings in such contexts could justify the alternation of verbal forms. See examples in (1a-c):

(1) a. Lorsqu'il est nécessaire d'administrer des produits radiopharmaceutiques chez la femme en âge de procréer, [...].

b. Quando è necessario somministrare un prodotto radioattivo ad una donna potenzialmente gravida, [...].

c. Where it is necessary to administer radioactive medicinal products to women of childbearing potential, [...].

Phraseological verbs include verbs that appear in collocations (strict lexical selection), in fixed phrases and also in support verb constructions.

The verbs examined in the first stage of this study are listed in Table 2:

French	Italian	English
affecter	coinvolgere	affect
atteindre	interessare	involve
intéresser		
toucher		

Table 2: Verbs examined in French, Italian and English.

Consider the Italian verb *interessare* and its French equivalent *intéresser*. Such verbs are mainly used by experts in the field, and they can be substituted by other words related to the general language (see table 2) without affecting the 'scientific' meaning which is given to them (see forward Section 3). Serianni (2005) labels the verbs *interessare/intéresser* as "tecnicismi collaterali" (subtechnical terms), i.e., words (nouns, adjectives, verbs and phrases) which are used to maintain a high, formal register in specialized languages.

Unlike the phraseological verbs, the terminological verbs correspond to those units whose meanings are specifically related to the specialized field, as in (2a-c):

(2) a. Des études in vitro ont montré que l'irbésartan est oxydé principalement par l'isoenzyme CYP2C9 du cytochrome P450 [...].
b. Studi in vitro indicano che irbesartan viene principalmente ossidato tramite il citocromo P450-enzima CYP2C9 [...].
c. In vitro studies indicate that irbesartan is primarily oxidised by the cytochrome P450 enzyme CYP2C9 [...].

These verbs often have deverbal nouns, which are terms themselves and should be included in terminological resources, e.g. (Fr.) *oxidation*, (It.) *ossidazione*, and (Eng.) *oxidation*.

2.3 Theoretical framework

Over the last few years, some researchers have proposed frame-based organizations of specialized fields, in other languages as well as English, such as environmental science (see Faber et al. 2005, among others), law (see Alves et al. 2005, among others), soccer (see Schmidt 2006 and his following writings), molecular biology (Dolbey et al. 2006 and his following writings), computing and the Internet (see L'Homme 2008).[1]

Frame Semantics (Fillmore 1977, 1982, 1985; Fillmore and Atkins 1992) is a theory of language understanding based on the principle that the meaning of a linguistic item (Lexical Unit, LU) interacts with the scene which it has activated ("meanings are relativized to scenes", Fillmore 1982). Thus, Fra-

[1] For a full bibliography on the application of Frame Semantics within LSPs, see Pimentel (2012).

me Semantics contributes towards understanding the significance of the verbal syntactic patterns, as well as the understanding of the components (Frame Elements, FEs) that form them semantically. For instance, defining the verbal lexical unit *learn* presupposes an educational teaching strategy (i.e., Education_teaching frame). Specialized verbs are often accompanied by other information (non-core Frame Elements, non-core FEs) that may be optionally added to a sentence.

The methodological approach applied to the analysis of specialized verbs is both bottom-up and top-down: the verbs are analyzed and grouped into frames for each language separately, and the use of specialized dictionaries and other reference resources provides helpful background information (Faber et al. 2009: 6). Thus, the analysis of text corpus allows us to observe how the arguments (core FEs and non-core FEs), the organization of syntax and the semantic connection between words put together specialized verbs and their suitable equivalents intra- and cross-linguistically.

The possibility of creating a multilingual specialized lexicon using the FrameNet database of its English-specific lexical descriptions is considered by Boas (2005), since semantic frames are conceptual structures independent of language. In this study, frames are assumed to be "interlingual representations" that can group together not only verbs in one language but also across several languages (French-Italian-English), by transferring semantic annotations from one language to another (Padó 2007; Baker 2009). Thus, frames can group together intra-linguistic and cross-linguistic equivalents (synonyms, near-synonyms, hyponyms, related LUs), as described in the next section.

3 Results

All the verbs examined in this study (see table 2) can be grouped together in the Damaging frame, since they all mean 'to have a strong effect on something or someone', or 'causing physical damage to something or someone', as shown in Table 3 (below). However, the Lexical Unit *to affect* is semantically identified with a general meaning in the FrameNet database, and it is linked to the Objective_influence frame.[2] This frame is the Parent frame of the Transitive_action frame from which the Damaging frame originates. Therefore, the Damaging frame is a Child frame which inherits from more than one Parent frame (multiple inheritance). Unlike *to affect*, the verb *involve* is not listed as a Lexical Unit in the database. Only the adjective *involved* and the noun *involvement* are included, and both belong to the Participation frame.[3] The corpus shows that the verb *involve* is used frequently as a synonym of the

[2] The definition of the Objective_influence frame is as follows: "An Influencing_variable, an Influencing_situation, or an Influencing_entity has an influence on a Dependent_entity, Dependent_variable, or a Dependent_situation".

[3] The definition of the Participation frame is as follows: "An Event with multiple Participants takes place. It can be presented either symmetrically with Participants or asymmetrically, giving Participant_1 greater prominence over Participant_2. If the Event is engaged in intentionally, then there is typically a shared Purpose between the Participants. It is, however, possible that an expressed Purpose only applies to Participant_1."

verb *affect* in medical discourse, and therefore it can be considered as a Lexical Unit of the Damaging frame.

Frame	**Damaging**
Definition	An Agent affects a Patient in such a way that the Patient (or some Subregion of the Patient) ends up in a non-canonical state. Often this non-canonical state is undesirable, and some lexical units (marked with the Negative semantic type) specifically indicate that the Patient is negatively affected.
Core FEs	AGENT [Agt] The conscious entity, generally a person, that performs the intentional action that results in the damage to the Patient. CAUSE [cau] An event which leads to the damage of the Patient. PATIENT [Pat] The entity which is affected by the Agent so that it is damaged.
Non_core FEs	CHARACTER_OF_END_STATE, DEGREE, INSTRUMENT, MANNER, MEANS, PATIENT, PLACE, PURPOSE, REASON, RESULT, SUBREGION, TIME
Contexts	Selon le NCI-CTC, les réactions cutanées de grade 2 sont caractérisées par une éruption **intéressant** jusqu'à 50 % de la surface corporelle, alors que les réactions de grade 3 affectent 50 % ou plus de la surface corporelle. Secondo i criteri NCI-CTC, le reazioni cutanee di grado 2 sono caratterizzate da rash che **interessa** fino al 50 % della superficie corporea, mentre quelle di grado 3 **interessano** il 50 % o più della superficie corporea. According to NCI-CTC, grade 2 skin reactions are characterized by rash up to 50 % of body surface area, while grade 3 reactions **affect** equal or more than 50 % of body surface area. Cette nécrose peut **atteindre** fascias musculaires ainsi que le tissu adipeux et peut par conséquent provoquer la formation d'une cicatrice. Questa può essere estesa e può **interessare** lo strato muscolare così come lo strato adiposo causando quindi la formazione di cicatrici. It can be extensive and may **involve** muscle fascia as well as fat and therefore can result in scar formation. Sintomi che coinvolgono il cervello e i nervi che si sono **manifestati** nell'arco di un mese [...] Réactions **touchant** le cerveau et les nerfs apparues dans le mois suivant la vaccination [...] Symptoms **affecting** the brain and nerves that have occurred within one month after vaccination [...] Les cas les plus graves ont été rapportés chez des patients prenant d'autres médicaments ou atteints de maladies pouvant **toucher** le foie (exemple alcoolisme, infection sévère). I casi più gravi sono stati osservati in pazienti trattati anche con altri medicinali o affetti da disturbi che possono **interessare** il fegato (ad es. abuso di alcolici, infezioni gravi). The most serious were reported in patients taking other drugs or who were suffering from diseases that can **affect** the liver (e.g. alcohol abuse, severe infection). S'ils ne sont pas **atteints**, la main et le pied doivent être protégés par une bande d'Esmarc, un garrot doit être placé au niveau proximal du membre. Mano e piede, se non **interessati**, devono essere protetti da bendaggi Esmarch (espulsione). Hand and foot, if not **affected**, should be protected by Esmarch (expulsion) bandages.

Table 3: The Damaging frame.

The Damaging frame groups together 8 candidate equivalents, i.e. 4 French verbs, 2 Italian verbs and 2 English verbs, more specifically 16 likely combinations of equivalents, as shown in more detail in Table 4:

French			Italian			English		
Cause	target	Patient	Cause	target	Patient	Cause	target	Patient
éruption	intéresser	surface corporelle	rash	interessare	superficie corporea		-	
réaction cutanée	affecter	surface corporelle	reazione cutanea	interessare	superficie corporea	reaction	affect	surface area
nécrose	atteindre	fascia musculaire	questa (necrosi)	interessare	strato muscolare	it (necrosis)	involve	muscle fascia
-	atteint(e)	osseuse	tumore maligno	interessare	osso	malignancie	involve(ing)	bone
maladie	toucher	foie	disturbo	interessare	fegato	disease	affect	liver
affection parodontale	toucher	gencive	disturbo periodontale	interessare	gengiva	periodontal	affect	gum
réaction	toucher	cerveau	sintomo	coinvolgere	cervello	symptom	affect	brain
	affect(ion)	moelle osseuse	patologia	coinvolgere	midollo osseo	conditions	affect(ing)	bone marrow

Table 4: Cross-linguistic comparison of verb (or *noun*) equivalents and FEs.

Most of the FEs in Table 4 are synonyms (or semantic equivalents) because they have similar meanings and distributions (uses). In a few cases, the corpus presents transcategorization phenomena from the verbal to the nominal form, as exemplified in (3a-c):

(3) a. des patients atteints de pathologie maligne à un stade avancé avec atteinte osseuse
b. tumori maligni allo stadio avanzato che interessano l'osso
c. in patients with advanced malignancies involving bone

According to L'Homme (2004), the presence of deverbal nouns, such as (Fr.) *atteinte* (<*atteindre*), (It.) *interessamento* (<*interessare*), (Eng.) *involvement* (<*involve*), establishes the specialized value of these verbs (see Section 2.2. for nouns derived from terminological verbs).

French is the language with the most verbal equivalents, since it distinguishes 4 items, whereas Italian and English contexts show 2, respectively. The English verbs *affect* and *involve* have peculiar features that characterize them as synonyms. In Italian as well as in French, the verbs can also be defined as hyponyms or hyperonyms. For instance, the Italian verb *coinvolgere* is a hyponym of *interessare*. All the verbs in Table 4 are equivalents because no particular differences have been observed: they have the same number of arguments (NP/Subject, NP/Object), the semantic nature of the arguments does not differ (CAUSE and PATIENT) (they refer to the same kind of entities), and their syntactic patterns are similar (see Pimentel 2012 for the criteria identifying equivalents). Only in a few cases, are the verbs not translated because of a syntactic change, e.g. (Fr.) *réaction allergique sévère touchant le corps entier*, (It.) *una grave reazione allergica dell'intero organismo*, and (Engl.) *a severe, whole-body allergic reaction*.

The analysis of the data allows us to identify the typical verbal features (tense, person, number, voice, and mood) which characterize the leaflets, and medical discourse in general. The grammatical persons are the singular and plural third-persons. The realis mood (indicative) is obviously the commonest, whereas the irrealis moods, such as the conditional, imperative and subjunctive forms are less frequent. For instance, the use of the conditional form in Italian and in French are 0,87% and 0,63%, respectively. In relation to the grammatical tense, the present is the most common tense when the indicative is used: (Fr.) 34,56%, (It.) 33,74%, (Engl.) 17,57%. A further note deserves to be made for the grammatical voice: the use of the passive construction placing the thema of a sentence at the beginning of the clause and the rhema at the end is very common in medical discourse, and generally allows one to omit the agent, e.g. *COX-2 is also thought to be involved in ovulation*.

4 Concluding remarks

The exploitation of specialized parallel corpora makes it easy to identify the repertoire of both intra-linguistic and cross-linguistic verb equivalents which acquire specialized value when used in medical texts. The Frame Semantics analysis of each verb pattern as well as the FrameNet methodology allow us to make a description of the interaction of the lexeme, syntax and conceptual background frame. All the verbal items evoke the same frame (Damaging) describing physical damage to something or someone. Thus, the lexicological findings could be useful for the development of a multilingual lexicographical resource specialized in the medical field which could give support with L2 writing. This of course involves a comprehensive and systematic investigation.

5 References

Alves, I., Chishman, R. & Quaresma, P. (2005). Verbos do domínio jurídico: uma proposta de organização ontológica com vistas ao PLN. In *Revista de Estudos Linguísticos da Universidade Federal de Juiz de Fora*, 9(1/2), pp. 123-137.

Baker, C. F. (2009). La sémantique des cadres et le projet FrameNet: une approche différente de la notion de «valence ». In *Langages*, 4, pp. 32-49.

Bertoldi, A., Chishman, R. (2012). Frame Semantics and Legal Corpora Annotation: Theoretical and Applied Challenges. In *Linguistic Issues in Language Technology*, 7(9), pp. 1-15.

Boas, H. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. In *International Journal of Lexicography*, 18(4), pp. 39-65.

Costa, R., Silva, R. (2004). The verb in the terminological collocations. Contribution to the development of a morphological analyser Morphocomp. In *Proceedings of the IV International Conference on Language Resources and Evaluation, May 26-28, 2004*. Lisbon, Portugal, pp. 1531-1534.

De Vecchi, D., Eustachy, L. (2008). Pragmaterminologie: les verbes et les actions dans les métiers. In *Actes des conférences Toth 2008, Annecy 5-7 June 2008*, pp. 35-52.

Dolbey, A., Ellsworth, M. & Scheffczyk, J. (2006). BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. In O. Bodenreider (ed.), Proceedings of KR-MED, pp. 87-94.

Faber, P., Márquez Linares, C. & Vega Expósito, M. (2005). Framing Terminology: A Process-Oriented Approach. In *Meta*, 50(4), CD-ROM.

Faber, P., León, P. & Prieto, J. A. (2009). Semantic relations, dynamicity and terminological knowledge bases. In *Current Issues in Language Studies*, 1(1), pp. 1-23.

Fillmore, C. (1977a). Scenes-and-Frames Semantics, Linguistic Structures Processing. In A. Zampolli (ed.), *Fundamental Studies in Computer Science*, 59, pp. 55-88.

Fillmore, C. (1977b). The Case for Case Reported. In P. Cole, J. Sadock (eds.), Syntax and Semantics, Volume 8: Grammatical Relations. New York: Academic Press.

Fillmore, C. (1982). Frame Semantics. In *Linguistics in the Morning Calm* (ed.), Seoul: Hanshin Publishing Co, pp. 111-137.

Fillmore, C. (1985). Frames and the Semantics of Understanding. In *Quaderni di Semantica*, 6(2), pp. 222-254.

Fillmore, C., Atkins, S. (1992). Towards a Frame-based Lexicon: The semantics of RISK and its Neighbors. In A. Lehrer, E. Kittay (eds.), Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization, Hillsdale: Lawrence Erlbaum Associates, pp. 75-102.

FrameNet: https://framenet.icsi.berkeley.edu/fndrupal/home [September-October 2013; February-March 2014]

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In W. Geoffrey, S. Vessier (eds.), Proceedings of the XI Euralex International Congress, July 6-10, 2004, France: Lorient, pp. 105-111.

L'Homme, M. C. (1995). Définition d'une méthode de recensement et de codage des verbes en langue technique: applications en traduction. In *Traduction, terminologie, rédaction*, 8(2), pp. 67-88.

L'Homme, M. C. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, 73(2), pp. 61-84.

L'Homme, M. C. (2004). *La terminologie: principes et techniques*. Montréal: Les Presses de l'Université de Montréal.

L'Homme M. C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. In *Traduire*, 271, pp. 78-103.

Lorente, M. (2000). Tipología verbal y textos especializados. In M. González Pereira, M. Souto Gómez (eds.), Cuestiones conceptuales y metodológicas de la lingüística. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 143-153.

Lorente, M., Bevilacqua, C. (2000). Los verbos en las aplicaciones terminográficas. In *Actas del VII Simposio Iberoamericano de Terminología RITerm 2000*. Lisboa: ILTEC.

OPUS Corpus : http://opus.lingfil.uu.se/ [September-October 2013]

Padó, S. (2007). Cross-lingual Annotation Projection Models for Role-Semantic Information. Ph. D. thesis. Saarland University.

Pettersson, Å. (2013). Les syntagmes participiaux et les verbes spécialisés dans un texte médical: Une étude contrastive entre le français et le suédois. Ph. D. thesis. Linnaeus University.

Pimentel, J. (2012). Identifying the equivalents of specialized verbs in a bilingual corpus of judgments: A frame based methodology'. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC 2012). Istanbul (Turkey), 21-27 May 2012, pp. 1791-1798.

Picht, H. (1987). Terms and their LSP Environment - LSP Phraseology. In *Meta*, 32(2), pp. 149-155.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R. & Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. ICSI Technical Report.

Schmidt, T. (2006). Interfacing Lexical and Ontological Information in a Multilingual Soccer FrameNet. In Proceedings of OntoLex 2006, Interfacing Ontologies and Lexical Resources for Semantic Web Technologies, Genoa, Italy, May 24-26, 2006.

Serianni, L. (2005). Un treno di sintomi. I medici e le parole: percorsi linguistici nel passato e nel presente. Milano: Garzanti.

Tellier, C. (2008). Verbes spécialisés en corpus médical: une méthode de description pour la rédaction d'articles terminologiques. PhD thesis. Université de Montréal.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.), Recent Advances in Natural Language Processing (vol V). Amsterdam/Philadelphia: John Benjamins, pp. 237-248.

Valente (2000). Peut-on considérer que le verbe est une unité lexicale spécialisée? Description de verbes spécialisés portugais. In *TradTerm*, 6, pp. 171-187.

Neoclassical Formatives in Dictionaries

Pius ten Hacken, Renáta Panocová
Leopold-Franzens-Universität Innsbruck, P.J. Šafárik University Košice
Pius.ten-Hacken@uibk.ac.at, Renata.Panocova@upjs.sk

Abstract

Neoclassical formatives are elements that occur in neoclassical word formations such as *ortho* and *paed(o)* in *orthopaedic*. We can be sure that a separate system of neoclassical word formation is in place when we find words such as *orthopaedic* that use classical elements but cannot have been borrowed from a classical language, because the concept they refer to did not exist yet. However, it must be taken into account that such words can also have been borrowed from a modern language that has such a system. In the lexicographic treatment of neoclassical word formation, a central question is whether neoclassical formatives should be treated in separate entries. We investigated how different English and Russian dictionaries treat them. In order to arrive at an unbiased sample of formatives, we used a Catalan word formation dictionary.

The results of our investigation support the hypothesis that neoclassical word formation constitutes a separate system in English, but not in Russian. This means that in English neoclassical formatives should have entries. In electronic dictionaries they can usefully be connected to the full class of words they appear in. In Russian, neoclassical formations are borrowings from languages such as English or French and their internal structure belongs to the domain of etymology.

Keywords: neoclassical word formation; neoclassical formatives; combining forms; English; Russian

In this paper we investigate the optimal treatment of neoclassical word formation in dictionaries. The main problem of neoclassical word formation in lexicography is to determine how to treat the internal structure and the components that are not words by themselves. First, we give a general introduction to the phenomenon of neoclassical word formation (section 1) and present some considerations as to the treatment of word formation in dictionaries (section 2). In order to investigate the coverage of neoclassical word formation in dictionaries, we present a sampling method that is not directly dependent on the languages and dictionaries under investigation (section 3). Then we compare the realization of neoclassical word formation and its treatment in some standard general dictionaries in two languages, English (section 4) and Russian (section 5). These languages were chosen because, as we will argue, the status of neoclassical word formation in them is different in interesting ways. On the basis of our observations in the preceding sections, we conclude in section 6 that there are good arguments for treating neoclassical formatives in separate entries in English, but not in Russian.

1 Neoclassical Word Formation

The discussion of neoclassical word formation has a long history. In many cases, it is connected to the opposition between learned and non-learned word formation. Bloomfield (1933: 153-54) discusses the opposition between unmarked and learned forms in relation to the way the lexicon of a language is organized and gives examples for various languages. He discusses the phenomenon of learned language mainly as a matter of register. Here, however, we will be interested only in the specific type of word formation involving Greek and Latin stems that is found in a wide range of European languages. An example from English is *orthopaedic*.

An issue that arises immediately in the context of neoclassical word formation is the boundary between word formation and borrowing. This issue arises in two shapes. The first can be illustrated on the basis of the contrast between *orthogonal* and *orthopaedic*. Both of these are based on Ancient Greek, but not in the same way. In Ancient Greek, the word ὀρθόγωνος ([orthógonos] 'rectangular') is attested and can be analysed as formed from ὀρθός ([orthós] 'straight') and γωνία ([gonía] 'angle, corner'). Therefore, it is possible to classify *orthogonal* in English as a borrowing from Ancient Greek. In the case of *orthopaedic*, there is no corresponding word in Ancient Greek. However, in Ancient Greek we have the words ὀρθός ([orthós] 'straight') and παῖς (stem παιδ- [paid-], 'child') which can be analysed as the basis for *orthopaedic*. This is the first realization of the issue of borrowing.

However, there is a second form in which this question arises. In fact, the non-existence of a corresponding Ancient Greek word does not necessarily show that the word was formed in English. In this case, it is possible to trace the origin to French. The French physician Nicolas Andry de Boisregard (1658-1742) created the term *orthopédique* in 1741. From French it was subsequently adopted in other European languages. In this case, it is possible to trace the formation of the word to a single person and a single publication, because the concept was invented by this person. In many cases it is much more difficult to trace the origin so precisely. If a concept is 'in the air' and arises in a context in which different people might think of the same name independently, or spread it very rapidly once someone starts using it, it is almost impossible to reconstruct which language is at the origin of a particular word.

In order to determine the significance in lexicography of the question of borrowing or formation in any particular case, we can choose between two perspectives. One is to aim for a description of the historical development of language use. In this perspective it is important to distinguish the origin of *orthogonal* and *orthopaedic*, stating that the former is borrowed from Ancient Greek and the latter from French where it was constructed from Ancient Greek components. The second possible perspective concentrates rather on the vocabulary as known by the speakers. Here the question is whether these words are structured or not. Most speakers will not be aware of the Ancient Greek origin of *orthogonal* or the French origin of *orthopaedic*. The difference between words formed in Ancient Greek and those formed in modern languages is not one that determines the structure of the mental lexicon of a contemporary speaker of English.

Following ten Hacken (2012), we will assume that in the case of genuine neoclassical word formation, there must be a system of neoclassical formatives in the lexicon. Such a system can emerge as the result of the reanalysis of borrowings. After a sufficient number of words with similar components have been borrowed, speakers may notice the regularity. In the first instance, both a new system and a new set of formatives have to be set up. It is quite likely that *ortho-* belonged to the initial set of formatives, because there are quite a number of old loanwords from Ancient Greek that contain it, e.g. *orthogonal*, *orthodox*, *orthography*. Once the system exists, it is easier to extend it. A single item may be sufficient. Thus, for *paedo-*, the origin may be a word such as *paedagogical*. We even find cases in which a neoclassical formative is borrowed directly from Ancient Greek, e.g. *psepho-* in *psephological*. This was not a formative in English when the word *psephological* was created in the 1950s. The first attestation in the OED is from 1952.

Systems of neoclassical word formation have come into existence in a number of languages, probably in the course of the 18th century. In fact, *orthopédique* is a very early example of a neoclassical formation that cannot have been borrowed directly from Ancient Greek. A much larger set of such words start appearing in the 19th century. Because languages such as French, English, and German all went through the same process, it is often difficult to determine which of these languages is at the origin of a particular word. However, the question of which modern language is at the origin of a particular formation is only relevant if we want to describe the history of a individual words. Speakers of such languages who have the neoclassical subsystem of word formation in their mental lexicon will analyse the relevant words independently of whether the original source is in their own language or in another one. Therefore, it is less relevant which of these languages is the historical origin of a particular word. Once it exists in one language it is quickly adopted in the others.

This is especially the case for words such as *orthopaedic*, because they belong to a very specialized register of language. Neoclassical words are part of the language of science. Not many speakers of English will have known them in the 18th or 19th century, but those who did are more likely to be in contact with and competent in other languages, such as French and German. As languages such as English, French and German only exist as entities when they are rational constructions associated with a conscious aim for codification, the decision to attribute the origin of a particular word to one of them is ultimately arbitrary.

2 Word Formation in Dictionaries

As argued in ten Hacken (2009), dictionaries do not describe the vocabulary of a language, but provide information about words to dictionary users. The aim of a description is not realistic because there is no suitable empirical object to describe. What we have is a number of speakers and a set of texts and utterances. The idea of a language to be described in a dictionary requires the classification of the speakers and texts/utterances and is in this sense a constructed entity, not one empirically found as an entity.

The role of word formation in dictionaries is significantly enhanced by the insight that dictionaries are no more or less than sources of information for their users. Ulsamer (2013) gives an overview of current practice and insights as to the representation of word formation in dictionaries of various types. Booij's (2003: 254) suggestion that the place of word formation in dictionaries is mainly to state "if a possible morphologically complex word actually exists" is unnecessarily restrictive. First of all, Booij's wording suggests that there is an independent sense in which a word should exist. In the same way as for languages, there is no empirical entity corresponding to the word. What the lexicographer has is a number of speakers and a corpus of texts and utterances. In addition, Booij's suggestion reduces the role the representation of word formation can play in supporting the dictionary user.

On the basis of the available corpus and linguistic knowledge, for each morphologically complex word, lexicographers have to take decisions such as the following:

- whether a particular word should be represented in the dictionary,
- whether it should be represented as morphologically complex,
- which information should be given about the word, and
- how this information should be presented.

It is crucial to understand that these are decisions based on judgements. The available data in a corpus can serve to support the judgement, but they do not result in a decision in an algorithmic sense. There is also no objective truth to be discovered as to the existence or the nature of the words considered. Svensén (2009: 131-133) sketches some of the traditional techniques that have been used, including the so-called *run-on* entry that Booij seems to allude to.

In electronic dictionaries, the possibilities for representing and accessing information are greatly enhanced compared to paper dictionaries. However, it is by now well-known that these additional possibilities depend on an appropriate encoding of the information. It is not sufficient to present information in the way it is encoded in a print dictionary and just make it available in electronic form. This argument was made quite systematically by ten Hacken (1998) and has now become commonplace. Ulsamer (2013: 35-50) can refer to a number of dictionaries and lexical resources that were developed with this idea in mind and exploit the specific strengths of the electronic medium.

Ten Hacken (1998) distinguishes three types of lexicographic representation of word formation. One type focuses on the word formation analysis of individual words. If we take as an example *happiness*, this means that the information that this word was formed by adding to the suffix *-ness* to the adjective *happy* is given in the dictionary entry for *happiness*. Giving this information is not as commonplace in paper dictionaries as one might expect. Most general dictionaries only give a run-on entry and in learners' dictionaries we often find a separate entry without this information. In both cases, this leaves it to the user to work out the structure of the word. In electronic dictionaries, there is of course no excuse based on space limitations to motivate such a decision.

A second type of information that dictionaries can give about word formation is the grouping of words that belong to the same word formation class. Here a word formation class is interpreted as a set of words that have some word formation properties in common. Ideally, it would be up to the user to choose which properties are most relevant for their question. In the case of *happiness*, the most obvious class would be that of nouns formed by suffixation of *-ness* to an adjective. This is a very large class and in a paper dictionary it is not normally presented, but in an electronic dictionary it is feasible to do so. Most users will only want to consider some examples, but linguistic researchers (a minority user group, but heavy dictionary users) may well be interested in the entire set. Other word formation classes that *happiness* is a member of include the class of words formed with *happy* as a basis, the class of nouns formed from adjectives, or the class of nouns formed by suffixation. Especially for the larger classes in these examples, any representation in an alphabetically ordered print dictionary will be too unwieldy, both to print it and to use it, but as Ulsamer (2013) shows, in electronic dictionaries there is an advantage in the proper representation of some of the possible classes.

The third type of representation concentrates on the rule as such. In the case of *happiness*, this may mean, for instance, an entry for *-ness* describing the word formation process. This is actually a common type of information also in print dictionaries. COED (2011) gives an entry for *-ness* with the different senses and an example for each sense. Atkins & Rundell (2008: 180) mention such entries as possible lemmata under the category of "partial words" and Svensén (2009: 132-133) presents it as one of the main representations of word formation in the dictionary. Of course, the representation in an electronic dictionary can be much improved by making a hyperlink available between the entries for the individual words (e.g. *happiness*) and the entry explaining the word formation rule, as well as by linking entries such as for *-ness* to the corresponding word formation class (i.e. all nouns formed with *-ness*).

The questions we want to address in relation to neoclassical word formation are then the following:

- which of these possible representations are used in current dictionaries?
- which improvements could be made in these representations, especially in electronic dictionaries? and
- to what extent do the choices depend on properties of different languages?

In order to provide a proper basis for the discussion of the last of these questions, we will consider two languages, English and Russian, in which it can be argued that the mechanism of neoclassical word formation does not work in the same way. First, however, we have to find a way to collect data for the first of these questions.

3 The Lexicographic Representation of Neoclassical Word Formation

In general, the decision how to treat a particular word formation process in a dictionary is determined at least in part by the productivity of the process. As shown by Bauer (2001), *productivity* is a notion that can be interpreted in different ways and has therefore raised a lot of discussion. A very useful analysis of the different notions of productivity is the one by Corbin (1987: 176-178). She distinguishes three aspects of productivity that in obvious cases will point in the same direction, but also give the conceptual vocabulary for a discussion about the choice of relevant properties in more controversial cases. The first is *régularité* ('regularity'). This is the extent to which the form and meaning of a particular formation can be predicted on the basis of the input (base) and the rule. The second is *disponibilité* ('availability'). This is a binary feature, indicating whether or not the rule is available for application to new bases. Finally, there is *rentabilité* ('profitability'). This is the degree to which the rule is applied to many new bases, yielding new formations. In her own work, Corbin (1987) concentrates only on availability. This is understandable because it is the underlying condition that has to be met in the linguistic competence before the other aspects can apply at all. However, in lexicography, the other two concepts are also relevant. Thus, regularity of new formations clearly influences to what extent run-on entries can be used and the profitability of a process will determine how important it is to treat it in the dictionary at all.

Much of the discussion of word formation in the context of lexicography concerns affixation. In the case of affixation it makes sense to consider each particular affix as a rule for which productivity (in its different aspects) can be calculated. In the case of *-ness*, the availability for new formations and the high degree of profitability make it a good choice for a separate entry. The regularity of many of the individual formations, such as *happiness*, makes it attractive to treat these as run-on entries.

If we want to extend this approach to productivity from affixation to neoclassical word formation, we encounter the problem that many neoclassical formations are more like compounding than like affixation. In the example of *orthopaedic*, there is a suffix *-ic*, but the central piece of the formation of this word is the combination of *ortho* with *paed(o)*. That neoclassical formatives such as *ortho* and *paedo* are not affixes is obvious from their distribution as well as from their contribution to the meaning of the resulting word. In lexicography, they are often called *combining forms*, e.g. by Svensén (2009:133). The variation in form is sometimes accounted for by different entries for the initial combining form and the final combining form. Thus, COED (2011) has different entries for *-phone* and *phono-*. This is not optimal for the insight that they represent the same formative, because there is no link from one entry to the other and a user not actively looking for the two variants will not find the connection, as they are separated by nine entries in the alphabetic order.

Returning now to productivity as a criterion to determine how to treat neoclassical word formation in a dictionary, we are faced by the situation that we have to decide for individual neoclassical formatives whether they deserve an entry and in which other ways they might be referred to. In this decision, affixation is not a good parallel, because neoclassical formatives are not affixes, but compounding is not a good model either. In compounding, the lexicographer has to decide whether or not to devote an entry to the compound, but not whether the components of the compound should be treated. In the case of *apple juice*, the only question is whether it deserves an entry of its own, not whether *apple* and *juice* should have an entry. We do not attempt to determine the productivity of *apple* in compounding as the basis for any lexicographic decision. In the case of *ortho* and *paedo*, we have to consider the number of different formations they appear in and the frequency of those formations in some way to determine whether they deserve being treated in a separate entry.

In order to explore the way neoclassical formations are covered in English dictionaries, we selected a sample of neoclassical formations. In many discussions of neoclassical word formation, a very limited set of examples is discussed time and again. However, there is no indication to what extent these examples are representative of the phenomenon. We considered that a sample based on any specific English dictionary would be biased, in particular when we want to compare the coverage in English dictionaries with the coverage in dictionaries of another language. Therefore we used a source from a language that is not in the scope of the study, Bruguera i Talleda's (2006) Catalan dictionary of word formation.

There are several reasons why Bruguera i Talleda (2006) is a good source for our study of neoclassical word formation. First, Catalan has neoclassical word formation in a way similar to other European languages. As a Romance language, it is not biased to English or Russian and it has a more direct link to Latin than either of these languages, so that we can expect that the set of neoclassical formations tends to be larger. Secondly, the dictionary offers a type of access to neoclassical word formation that is convenient for our purposes. The lemmata of the dictionary are affixes and neoclassical formatives. The entries contain basic information about the use of the formatives and a full list of words formed with them. Where appropriate, entries are divided into different senses. In addition, the introduction (2006: 9-50) contains a detailed discussion of the different types of word formation and a classification of the lemmata. Therefore, it was easy to select a randomized sample of relevant formatives. Specialized dictionaries of this type are quite rare, in particular published as paper dictionaries.

Bruguera i Talleda (2006: 31-48) treats neoclassical word formation as *formaciò culta* ('learned formation'), which connects with Bloomfield's (1933) category of learned word formation as treated in section 1. Neoclassical formatives are listed in the section on neoclassical compounding. This is not ideal in principle, because neoclassical word formation is not restricted to compounding, as evidenced by formations such as *ethnic* or *morpheme*. However, compounding is much more prominent in neoclassical word formation than derivation, so it can be expected that in practice any formative that appears in derivation will also appear in compounding. Neoclassical word formation as it is used, for instance, in medicine is based on Greek formatives that in many cases passed through Latin. Bruguera i Talleda (2006: 38-47) gives separate tables for formatives of Greek and of Latin origin. We only considered the former. In addition, the formatives are divided into initial and final combining forms. There is a large degree of overlap between the two, but the former list is significantly longer. An additional practical advantage of initial combining forms is that it is immediately evident when looking them up in an alphabetically ordered dictionary which and how many entries use them. Therefore we based our sample of neoclassical formatives on a random selection from the list of initial combining forms of Greek origin. The only adjustment we had to make is to adapt the spelling from Catalan to English and Russian.

4 Neoclassical Word Formation in English Dictionaries

For English, we took as our dictionaries CED (2000) and COED (2011). In Béjoint's (2000: 42-91) overview of dictionaries of English, it is obvious that British and American dictionaries follow different patterns. Therefore it would not be possible to generalize from one type of dictionary to the other. However, within British dictionaries, CED (2000) and COED (2011) fall into different subtypes. Béjoint (2000: 57-58) classifies the Oxford dictionaries as traditional, where the Collins dictionary belongs to an innovative trend that started in the 1970s. Both belong to the type van Sterkenburg (2003) identifies as 'the' dictionary, i.e. general-purpose dictionaries of a size big enough to give a fairly comprehensive overview of the vocabulary without giving a full scholarly account of its development. Therefore, especially when the findings of the two dictionaries coincide, we can safely draw conclusions for British dictionaries of this type in general.

The first sample we took was a randomized set of items from Bruguera i Talleda's (2006: 40-44) list of initial combining forms of Greek origin. We found that almost all of them had at least two examples of formations that were described in both English dictionaries. This means that the basic condition for identifying the item as a neoclassical formative would be fulfilled. However, only about a third of the formatives is described in separate entries. Thus, CED (2000) gives *thanatology* and *thanatopsis*, but no separate entry for *thanato-*. The structure of the words is only addressed in the section on etymology, where we find the following:

thanatology: [C19: from Greek *thanatos* death + -LOGY]
thanatopsis: [C19: from Greek *thanatos* death + *opsis* a view]

The difference in presentation suggests that *-logy* and *-opsis* are treated differently in the dictionary. In fact, both have an entry as a combining form. The difference between CED (2000) and COED (2011) is small. There are a few cases where COED (2011) gives the etymology and CED (2000) does not, e.g. *mammography*. Conversely, CED (2000) gives slightly more separate entries for combining forms. Thus, only CED (2000) has an entry for *noso-*. However, on the whole the two dictionaries have a remarkably similar treatment of the formatives in the sample.

We were not fully satisfied with our first sample, because for many of the formatives there were so few items that were listed in the dictionary that it was not clear whether lexicographic decisions concerning the inclusion of particular words or the lexicographic approach to neoclassical formatives were responsible for the treatment we found. Thus, for *thanato-*, CED (2000) gives the two entries listed above, but COED (2011) gives only the first. Therefore, we took a second sample, which only included formatives that were relatively profitable in Corbin's (1987) sense.

For this second sample, we took as a selection criterion the length of the entry in Bruguera i Talleda (2006). As this dictionary gives all attested words with the relevant formative in Catalan, the length of an entry gives a measure of the profitability of the formative. From our first sample, we discarded each item for which the entry in Bruguera i Talleda (2006) is shorter than ten lines. Where the same formative occurs as an initial as well as final combining form, we combined the length of both entries. We replaced the discarded items with formatives that fulfilled this length criterion. The results of looking up this second sample in CED (2000) and COED (2011) showed that these formatives are generally more often described in a separate entry. In COED (2011) more than two thirds of the second sample appears as an entry and in CED (2000) we found almost all of them. This difference is in line with the tendency we observed in the first sample that CED (2000) gives more separate entries than COED (2011). As an example of an entry, COED (2011) gives the following for *diplo*:

diplo- ▸**comb. form 1** double: *diplococcus*. **2** diploid: *diplotene*.
- ORIGIN from Gk *diplous* 'double'.

The corresponding entry from CED (2000) is as follows:

diplo- *or before a vowel* **dipl-** *combining form*. double: *diplococcus*. [from Greek, from *diploos*, from DI-[1] + *-ploos* -fold]

The entries in the two dictionaries are very similar, but it is interesting to note the differences. Only CED gives the variant form. Only COED gives the second sense. This second sense is the result of a shortening of one of the complex forms the formative appears in. It is similar to the use of *gastro-* in the sense of 'gastronomic' rather than 'stomach'. In the etymological information, COED gives the classical form as it is likely to have been at the origin of the borrowing, whereas CED gives the oldest attested form in Greek as well as its word formation origin.

On the basis of the samples of neoclassical formatives we considered, we can conclude that British general dictionaries tend to include separate entries for neoclassical formatives when they appear in many different words. In the entries for these formatives, but also in the ones for neoclassical formations, the connection to the Ancient Greek forms is made explicit. There is a slight difference between CED (2000) and COED (2011) in the sense that the former has more separate entries for neoclassical formatives whereas the latter is more systematic in giving etymologies for neoclassical formations.

5 Neoclassical Word Formation in Russian Dictionaries

The position of neoclassical word formation in Russian is not directly comparable to that in English or other Germanic and Romance languages. Neoclassical forms are generally much rarer in Russian and in many cases, a competing form with Slavic roots is preferred. Therefore, our hypothesis is that Russian dictionaries will give fewer neoclassical formations and will not analyse them as readily as English dictionaries.

In exploring this hypothesis, we used the dictionaries by Ušakov (1946-47) and Efremova (2000). They are roughly comparable in size to the dictionaries we used for English. In the Russian lexicographic tradition, the monolingual dictionary by Ušakov (1946-47) is categorized as a normative dictionary of contemporary standard Russian (Šanskij, 1972: 286; Ožegov, 1974: 171; Germanovič, 1979: 264). The dictionary includes nearly 90,000 entries. Germanovič (1979: 265) points out it is the first dictionary which gives separate entries for productive word formation elements such as prefixes or affixoids. The dictionary by Efremova (2000) is more recent and lists around 140,000 entries. The introductory material to the online version of the dictionary gives the information that prefixes, suffixes, initial elements of complex words and final elements of complex words are described in separate entries.

On the basis of this policy description, one might expect that the findings of checking our samples of formatives in Russian dictionaries would be closer to the results in in CED (2000) and COED (2011) for English than originally thought. However, a closer look at the coverage of our first sample of formatives from Bruguera i Talleda (2006) in these two dictionaries confirmed our original hypothesis that only few neoclassical formatives are described in separate entries. The two dictionaries often differ in the treatment of the formatives in the sample. For instance, the dictionary by Ušakov (1946-47) gives этнолог ([etnolog] 'ethnologist'), этнологический ([etnologičeskij] 'ethnological'), этнология ([etnologija] 'ethnology') but does not have a separate entry for этно- ([etno-] 'ethno-'). A large part of the formatives in the first sample follow a similar pattern. For some neoclassical formations, Ušakov (1946-47) gives etymologies explaining the components of words as based on Ancient Greek.

ЭТНОГРАФИЯ, этнографии, мн. нет, ж. (от греч. ethnos - народ и grapho - описываю)

[etnografija, etnografii, mn. net, ž. (ot greč. ethnos – narod i grafo – opisivaju)]

'ethnography, ethnography$_{\text{GEN}}$, no plur., fem. (from Greek ethnos – nation and grapho – writing)'

In principle, such etymological information would enable the user to add a system of neoclassical word formation to their mental lexicon, but there is no indication that many speakers of Russian do this. On the other hand, the dictionary by Efremova (2000) classifies the same formative as an initial part of complex words:

этно-

Начальная часть сложных слов, вносящая значение сл.: народ (этногенез, этнолингвистика, этнопсихология и т.п.).

[etno-]

[načaľnaja časť složnych slov, vnosjaščaja značenie sl.: narod (etnogenenez, etnolingvistika, etnopsichologija i t.d.)]

'ethno-'

'initial part of complex words having the meaning nation (ethnogenesis, ethnolinguistics, ethnopsychology, etc.)'

Efremova (2000) does not give information about the etymology of neoclassical formatives, but the introductory material mentions that entries are included for around 900 combining forms. Of course not all combining forms are neoclassical formatives.

Checking our second sample of formatives from Bruguera i Talleda (2006) in these two dictionaries, we found that our initial expectations were largely confirmed again. Only for just over half of the formatives do the Russian dictionaries give any examples of formations, the number of formations per formative is lower than in English dictionaries, sometimes just one example, and only very rarely is the neoclassical formative described in a separate entry. Despite the generous coverage of combining forms announced in Efremova's (2000) introductory material, we only found very few separate entries for formatives that are part of our second sample. For *phago-*, Efremova (2000) gives the following entry:

фаго

Начальная часть сложных слов, вносящая значения: 1) поедание, пожирание чего-л. (фагоциты); 2) связанный с бактериофагом (фагодиагностика, фагопрофилактика).

[fago]

[Načaľnaja časť složnych slov, vnosjaščaja značenija: 1) pojedanije, požiranije čego-l- (fagocity); 2) svjazannyj s bakteriofagom (fagodiagnostika, fagoprofilaktika).]

'phago'

'initial part of complex words having the meanings: 1) eating of something (phagocytes); 2) related to bacteriophag (phagodiagnosis, phagoprophylaxy)'

The second sense given in this entry is similar in nature to the second sense of *diplo* in section 4. It results from the shortening of a full neoclassical formation. Interestingly, Efremova (2000) lists separately also the formative *-phag* occurring in final position. The entry provides the following information:

фаг

Конечная часть сложных существительных, вносящая значение: поедающий, поглощающий то, что указывается в первой части слова (ихтиофаг, фитофаг и т.п.).

[fag]

[Konečnaja časť složnych suščestviteľnych, vnosjaščaja značenie: pojedajuščij, pogloščajuščij to, čto ukazyvajetsja v pervoj časti slova (ichtiofag, fitofag i t.p.)]

'phag'

'final part of complex nouns having the meaning: eating, eating what is denoted by the initial part of the word (ichtiophag, phytophag, etc.)'

Ušakov (1946-47) does not give a separate entry for the initial neoclassical formative *phago*, but gives фагоцит ([fagocit] 'phagocyte'), and фагоцитоз ([fagocitoz] 'phagocytosis'). Similarly, the final neoclassical formative *-phag* is not given as an independent entry, but can only be traced, at least in principle, in entries such as фитофаг ([fitofag] 'phytophag').

An interesting case is represented by the formative *diplo.* The meaning corresponding to the examples from English in section 4 is found in the formations диплококк ([diplokok] 'diplococcus'), but the formative itself does not appear in a separate entry. In Efremova (2000), an initial part of complex words дип ([dip] 'dip'), is treated as an independent listed item. The meaning is, however, related to дипломатический ([diplomatičeskij] 'diplomatic'). It is a clipping found in formations such as дипкупе ([dipkupe] 'diplomatic compartment'), дипкорпус ([dipkorpus] 'diplomatic corpus') or диппочта ([dippočta] 'diplomatic mail'). This pattern of forming words is very productive in Russian and it is referred to as *stump compounds* or *abbreviated compounds* (cf. Benigni and Masini, 2009, Comrie and Stone, 1978, Molinsky, 1973).

On the basis of our sampling, we may conclude that in Russian monolingual general dictionaries the coverage of neoclassical word formation is much less extensive than in English. First of all there are far fewer entries for neoclassical formations. Moreover the way they are treated reflects to a much smaller extent a system of neoclassical word formation. Ušakov (1946-47) only gives etymologies for neoclassical formations, which is hardly sufficient to retrieve the use of neoclassical formatives. Efremova (2000) gives many entries for combining forms, but from our samples only few neoclassical formatives are actually covered in them.

6 Conclusion

In this paper we investigated the treatment of neoclassical formatives in English and Russian dictionaries. The basis of our research was a sample of initial combining forms. In order to exclude a direct bias towards English or Russian in our sampling, we used a word formation dictionary for a third language, Catalan, as the basis for our samples. The choice of initial combining forms was based on the consideration that where neoclassical formatives appear as the base of a derivation, they also occur as initial combining forms in a neoclassical compound. Almost all final combining forms also appear as initial combining forms. Initial combining forms are a much larger set and they can be easily retrieved also in an alphabetically ordered paper dictionary.

As our dictionaries, we selected CED (2000) and COED (2011) for English and Ušakov (1946-47) and Efremova (2000) for Russian. These dictionaries represent more traditional and more modern trends in British and Russian lexicography in such a way that we can consider our results typical of British and Russian dictionaries in general.

We found that English dictionaries give more neoclassical formations than their Russian counterparts, which suggests that they are more numerous in English. In addition, individual formatives are in many cases described in a separate entry in English dictionaries, but this is very rare in Russian dictionaries. These findings are in line with the hypothesis that for a significant proportion of speakers of English there is a system of neoclassical word formation, whereas this is not the case for speakers of Russian.

To the extent that this hypothesis is correct, the lexicographic policy on the inclusion of entries for neoclassical formatives adopted in English and Russian dictionaries can be justified by properties of the languages they cover. In English, having the information as to what a particular formative means will enable the user to decode new formations that are not in the dictionary, help the user build up a system of neoclassical word formation as part of their mental lexicon, and thus support the acquisition and retention of new neoclassical formations.

In Russian, the situation is different. New neoclassical formations will be borrowings, not the result of applying a neoclassical formation rule. If such a borrowing combines two neoclassical components, an entry for one of these components will often be of little help. In many cases, the other component will not exist yet, so that there is no obvious background structure into which a new formation could be incorporated.

In electronic dictionaries, representing neoclassical formatives is to be recommended for English. In an optimal representation, an entry for a formative will give access to the class of all formations it is part of. This includes both the use as an initial and as a final combining form, because a formative that occurs in both roles (e.g. *paedo* in *orthopaedic* and in *paedagogical*) is basically the same formative. For Russian, there is no similar level of support for setting up such a system.

7 References

7.1 Dictionaries

Bruguera i Talleda, Jordi (2006), *Diccionari de la formació de mots*, Barcelona: Enciclopèdia Catalana.

CED (2000), *Collins Dictionary of the English Language*, 5th edition, Glasgow: Collins.

COED (2011), *Concise Oxford English Dictionary*, 12th edition, Angus Stevenson & Maurice Waite (eds.), Oxford: Oxford University Press.

Efremova, Tatjana F. (2000), Новый словарь русского языка [New dictionary of the Russian language], Moscow: Russkij Jezik, http://www.efremova.info.

OED (2014), *Oxford English Dictionary*, Third edition, edited by John Simpson, www.oed.com.

Ušakov, Dmitrij N. (1946-47), Толковый словарь русского языка [Explanatory dictionary of the Russian language], online edition http://www.dict.t-mm.ru/ushakov.

7.2 Other works

Atkins, B.T. Sue & Rundell, Michael (2008), *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.

Bauer, Laurie (2001), *Morphological Productivity*, Cambridge: Cambridge University Press.

Béjoint, Henri (2000), *Modern Lexicography: An Introduction*, Oxford: Oxford University Press.

Benigni, Valentina and Masini, Francesca, (2009), 'Compounds in Russian', in *Lingue e linguaggio* 2/2009, pp. 171-194.

Bloomfield, Leonard (1933), *Language*, London: Allen & Unwin.

Booij, Geert (2003), 'The codification of phonological, morphological, and syntactic information', in van Sterkenburg, Piet (ed.), *A Practical Guide to Lexicography*, Amsterdam: Benjamins, pp. 251-259.

Comrie, Bernard and Stone, Gerald (1978), *The Russian language since the revolution*, Oxford: Clarendon Press.

Corbin, Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, Tübingen: Niemeyer (2 vol.).

Germanovič, Ivan Klimovič (1979), 'Лексикография' [Lexicography], in Šuba, Pavel Pavlovič, Современный русский язык I [The contemporary Russian language I], Minsk: BGU, pp. 256-312.

ten Hacken, Pius (1998), 'Word Formation in Electronic Dictionaries', *Dictionaries* 19:158-187.

ten Hacken, Pius (2009), 'What is a Dictionary? A View from Chomskyan Linguistics', *International Journal of Lexicography* 22:399-421.

ten Hacken (2012), 'Neoclassical word formation in English and the organization of the lexicon', in Gavriilidou, Zoe; Efthymiou, Angeliki; Thomadaki, Evangelia & Kambakis-Vougiouklis, Penelope (eds.), *Selected papers of the 10th International Conference of Greek Linguistics*, Komotini: Democritus University of Thrace, pp. 78-88.

Molinsky, Steven J. (1973). Patterns of ellipsis in Russian compound noun formations. The Hague/Paris: Mouton.

Ožegov, Sergej Ivanovič (1974), Лексикология, лексикография, культура речи [Lexicology, lexicography, speech culture], Moskva: Vysšaja škola.

van Sterkenburg, Piet (2003), ''The' dictionary: Definition and history', in van Sterkenburg (ed.), *A Practical Guide to Lexicography*, Amsterdam: Benjamins, pp. 3-17.

Svensén, Bo (2009), A Handbook of Lexicography: The Theory and Practice of Dictionary-Making, Cambridge: Cambridge University Press.

Šanskij, Nikolaj Maksimovič (1972), Лексикология современного русского языка [Lexicology of the contemporary Russian language], Moskva: Prosveščenije.

Ulsamer, Sabina (2013), 'Wortbildung in Wörterbüchern – Zwischen Anspruch und Wirklichkeit', in Klosa, Annette (ed.), *Wortbildung im elektronischen Wörterbuch*, Tübingen: Narr, pp. 13-59.

Reports on Lexicographical and Lexicological Projects

Revision and Digitisation of the Early Volumes of *Norsk Ordbok*: Lexicographical Challenges

Sturla Berg-Olsen, Åse Wetås
Norsk Ordbok 2014, University of Oslo
sturla.berg-olsen@iln.uio.no, ase.wetas@iln.uio.no

Abstract

2014 will see the work on the 12th and final volume of the academic dictionary *Norsk Ordbok* (NO) finished. Still, the dictionary will remain heterogeneous due to variation in editorial practice throughout the project and incomplete in the sense that its early volumes are not digitally available. The online version of NO currently only covers the alphabet from the letter *i*. This paper describes the present state of the different parts of NO and argues that the early volumes of the dictionary must be revised and digitised to bring them up to the standards of the rest of the work. The revision and digitisation will not only give the dictionary a unitary profile but also make it possible to use it for a number of other purposes and facilitate the continuous process of keeping the dictionary up to date. The paper discusses some of the lexicographical challenges involved in the planned revision project and displays examples of the changes that must be made to the structure of the early material. It also touches upon questions concerning project organisation and funding.

Keywords: Academic dictionaries; Digitisation; Project planning

1 The history and present status of *Norsk Ordbok*

Norsk Ordbok (NO) is an academic dictionary covering Norwegian Nynorsk and all Norwegian dialects. The dictionary will provide a scholarly and exhaustive account of spoken Norwegian and of texts written in Nynorsk from 1860 up till today, and is to be completed during 2014, the year of the bicentenary of the Norwegian constitution. From 2002 the dictionary work has been organised in the time-limited project organisation Norsk Ordbok 2014 (NO 2014). The project owner is the Department of Linguistics and Scandinavian Studies at the University of Oslo. In 2014, the finished work will include more than 300,000 entries, published in 12 volumes.

When NO was conceived in the late 1920s, Nynorsk was still a written language in the making, and the standard was continuously fed by Norwegian dialect words. The proponents of Nynorsk wanted to make a comprehensive scholarly dictionary building on the works of the famous Norwegian 19th century linguist Ivar Aasen. The immediate goal behind the dictionary was to develop Nynorsk further, and to raise the prestige of the new written standard. The combination of dialects and written standard in one dictionary – somewhat unusual in a wider European context – was considered a natural choice, given the crucial role dialect data had played in the codification of Nynorsk from the outset. Even today the editors of NO regularly write entries entirely based on dialect material. This process often includes codifying the spelling and inflection of these words according to the Nynorsk standard.

The collection of data for a new and comprehensive dictionary of Nynorsk started in 1930. A dictionary board of trained lexicographers instructed and supervised more than 550 volunteers, who during these early years collected dialect data from all over the country and made it possible for the dictionary board to build up a huge slip archive. The learned dictionary board also supervised the extraction of literary excerpts from Nynorsk literature, both fiction and non-fiction. In addition, they compiled a draft manuscript combining Ivar Aasen's dictionaries (Aasen 1850 and 1873) with a range of other canonical dictionaries dating from 1870 to 1910, also adding data from glossaries and local dictionaries dating from 1600 to 1850 (Skard 1932). This draft manuscript for the new, academic dictionary was finished by 1940.

The editing of the dictionary started in 1946, and the first volume of NO was published 20 years later, covering the alphabet from the letter *a* to the adjective *doktrinær*. The original plan was to make a 2–3 volume dictionary, but in 1966 the chief editor estimated that 8–9 volumes would be needed to cover the whole alphabet (Hellevik 1966). During the first 50 years the editing of the dictionary progressed slowly. At the same time the source material grew, and so did the dictionary entries in volumes 2, 3 and 4. All the work was done manually, the slips sorted on the lexicographer's desk and the manuscripts prepared in handwriting.

In 2002 the work was reorganised and moved to a digital platform, making the editing process a lot more efficient. Increased funding allowed the project to employ more editors, and the work gained speed. During the period 2005–2013 7 volumes were published, with the last volume to be finished in 2014. However, the volumes produced before 2002 (volumes 1–4 and roughly half of volume 5) remain only partly digitised and show a number of discrepancies compared to the latter volumes. This has to do with changes in editorial practices that were implemented along the way. The digitisation of the volumes produced before 2002 and the revision of the contents of these volumes to bring them up to date are essential tasks that must be undertaken after the completion of the last volume. This will ensure that NO is a homogeneous dictionary that meets the scholarly standards of the age of electronic corpora and can be updated continuously in the future. Only when the entry database covers the whole alphabet can it be used for other purposes (e.g. the extraction of semantic structures to form the basis of a Nynorsk word net, the extraction of subsets of entries for new, thematic dictionaries etc.). In addition, revisions of the entry database itself can then be organised thematically, and not necessarily alphabetically.

Since 2012 an online version of NO has been available, but this version only contains the material from the letter *i* onwards. A complete online version is dependent on the complete digitisation of the early material and adaptation of this material to the database system used.

The reorganisation into the time-limited project NO 2014 also led to a change in profile for the dictionary. During the whole history of the dictionary, there has been a strict emphasis on constructing a scholarly work that meets scientific demands. However, the editorial profile of the earliest volumes was that of a scientific paradigm still concerned with nation-building. The dictionary was part of the work to document and elaborate on Nynorsk as a cultural object and to further standardise this language, which was still in its formative stage. In the modern project organisation, the emphasis is on editorial practice as descriptive research work. This results in the inclusion of entries that were earlier not considered part of Nynorsk proper, but which have entered Nynorsk during the last 50 years. To take one example, during the work on the entries starting with the Norwegian privative prefix *u-* a number of instances were discovered where the 'positive' counterparts of these words from the earliest parts of the alphabet were not covered (words like *bekvem* 'comfortable'*, bekymra* 'worried', *bemanna* 'manned' etc.). These are loan words from Danish and German that were earlier only used in Bokmål, but they are now also part of modern Nynorsk and should thus be included.

2 The microstructure of *Norsk Ordbok*

The microstructure of NO entries is fairly similar to that found in other comprehensive scholarly monolingual dictionaries, such as the OED, the *Dictionary of the Danish Language* (ODS) and the *Swedish Academy Dictionary* (SAOB). Each headword is followed by a section containing information on early lexicographical sources listing the word and etymology, as well as pronunciation (mainly for borrowed words) and alternative written forms of the word. This section also provides attested dialect forms of the word with geographical indications. Only dialect forms that do not follow automatically from general and well-known rules of sound correspondences in Norwegian dialects are included. The introductory section is the part of the NO entries which has seen the most variation and change during the project period. In the early volumes there was a certain degree of experimenting both with the order of the information given here and the structuring of this information. The digital platform used from 2002 onwards ensures stringency, but the variation found in the introductory section in the early volumes presents big challenges when it comes to digitisation.

The part of the entry following the introductory section is fairly straightforward, with potentially three explicit levels of senses, each sense customarily followed by literary sources and/or geographical indications, as well as examples of usage. In the early volumes, multi-word expressions are treated largely on a par with ordinary examples. Starting from the letter *i*, such expressions have been edited as sublemmas, appearing in boldface.

3 Challenges involved in the digitisation and revision of the early volumes

The goal of the revision project is to bring volumes 1–5 up to the same standards and give the entries in these volumes the same structure as that found in volumes 6–12. The contents of volumes 1–5 must be evaluated in view of the present editorial policies and revised on all levels where necessary in order to reflect these policies. This involves restructuring, adding information and also (particularly in volumes 3–4) removing some information. The result will be a homogeneous product reflecting the Nynorsk of the 21st century as well as the history of this written standard and the diversity of the Norwegian dialects.

There are several possible ways of digitising the oldest volumes of the dictionary. One solution could be OCR-scanning. This process was chosen for the first online version of the *Swedish Academy Dictionary* (SAOB) in 1997, but the result was considered unsatisfying and also turned out to be very expensive (Mattisson 2012). SAOB is currently going through a second re-digitisation process. This time the printed text is punched and stored in digital files in China. When this part of the process is finished, the SAOB editorial staff themselves will process the files by hand into valid XML. A similar process was chosen by the Society for Danish Language and Literature when they digitised their 28-volume *Dictionary of the Danish Language* (ODS) in 2005 (cf. ODS FBTS). The solution chosen for the ODS and for the second digitisation of the SAOB seems to be a good choice for older dictionaries where all the text is produced as typed manuscripts to feed a print version. The situation for NO is not quite similar to these works. Firstly, the dictionary has been produced on a digital platform from the letter *i* onwards. When the work on the 1st edition finishes in 2014, approximately 2/3 of the dictionary entries will be digital entries feeding both the online dictionary and the printed version. Secondly, the punching part of the digitisation process is already done for the oldest volumes of NO. In order to make an online version which covers the whole dictionary, and in order to complete the dictionary database, the only fully satisfactory solution for our dictionary will therefore be to integrate the digitised text from the oldest volumes into the already existing entry structure of the digital dictionary.

The current state for volumes 1–5 of Norsk Ordbok is that the two first volumes were punched and proofread in 2001–02. The manuscripts for volumes 3–4 and the part of volume 5 that covers the letter *h* were produced in simple word processing programmes, and supplied with tags either during the editing process or afterwards. The original text for the oldest volumes of NO thus existed as digital manuscripts as early as in 2002. In 2005, the Norsk Ordbok 2014 project organisation made a pilot study on the integration of this digital text into the modern database system. The adaptation of the texts into the new and stringent database format proved too difficult and too time-consuming for the time-limited project organisation, and was therefore put on hold.

The entries from volumes 1 and 2 are integrated in the database system of NO 2014, but only in an incomplete version. The text is not in line with the current quality when it comes to consistency, and it does not give a complete coverage of older source material. Volumes 3 and 4 are partly integrated in the database, but a lot of the text is not fitted into the correct fields, and the huge amount of dialect data and information on etymology is lacking altogether. The part of volume 5 covering entries starting with the letter *h* is not integrated in the database at all.

3.1 Why digitisation *and* revision?

Why is it so important to do the digitisation and the revision in one integrated operation? As mentioned above, the project organisation made a pilot study in 2005 to see if it would be possible to load the text of the oldest volumes into the modern editorial database. The pilot revealed that a lot of work has to be done to make the old text fit into the strict categories of the new editorial system, and that work inevitably also involves revision. One way of presenting the whole dictionary digitally without performing this integrated process of digitisation *and* revision would be to publish the oldest volumes as searchable PDFs on the Internet. This would be very unsatisfactory for several reasons: low user-friendliness, no possibility to perform searches across the base, lack of access to multi-word expressions in the earliest volumes, lack of possibility to do thematically based revisions and use the dictionary contents for other purposes etc.

Producing a digital dictionary which is identical to the printed version of NO is not the best solution in the view of the project organisation. Instead, we want to fit the entries from volumes 1–5 into the modern editorial database format. Preserving the contents of the oldest volumes in detail would force us to extend the existing database structures in order to adapt it to the structure and the idiosyncrasies of the old entries. Our goal is instead to modernise and standardise these entries and adapt them structurally to the modern online dictionary format.

3.2 Structural changes related to the digitisation

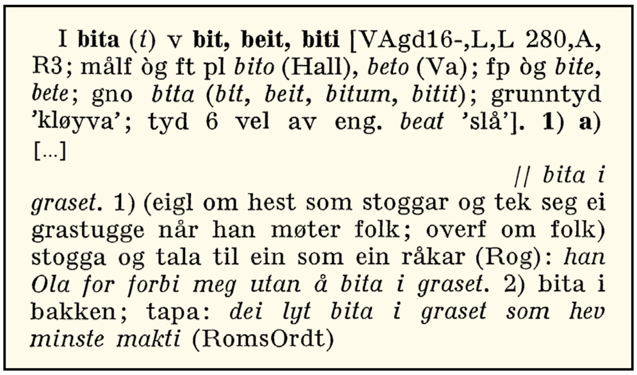
The first four and a half volumes of the dictionary were produced manually. The entries of these volumes are of a high quality for their time, but they often have a very tiered structure (Atkins & Rundell 2008:249) and from time to time include entry-specific structuring of data. This practice is possible and probably inevitable when the manuscripts are produced by hand, but it meets problems with the introduction of a digital production platform.

In 2002, the senior editing staff of the dictionary did a huge job extracting an ideal entry structure from the early volumes. This was used for setting up the electronic editing schema of the modern, digitised dictionary. The entry structure at the macro level (entry status, flat vs tiered structure, content selection etc., cf. Atkins (2008: 36ff)) was created on the basis of what was conceived to be the best practice of the old volumes, but this still leaves a lot of information that will not fit into the categories of the schema, and that will need to be given elsewhere in the entries or, if deemed superfluous, deleted. The planning of the entry structure at the macro level is much in line with the process of dictionary planning described by Atkins (2008), but for a dictionary project that has already published five volumes the options when setting up the macro structure are not open in the way they are when planning new dictionary projects (see also Cantell & Sandström 2012: 166f).

Another task associated with the digitisation of the material is the electronic linking of words in definitions, etymologies and elsewhere, as well as the linking of the first part of compounds to the correct basic word. Such links are an integral part of the structure in the latter volumes, and must be added also in the early material. This linking also requires that the structure of the older volumes is possible to adapt into the new data base system.

3.3 Structural changes related to the revision

Several structural changes must be performed in the old material in order for it to meet the requirements of NO's present editorial practices; a few examples will be mentioned here. As stated above, multi-word expressions were in the early volumes treated more or less on a par with ordinary examples, while starting from the letter *i* they have been edited as sublemmas in boldface. In order to attain a unitary structure throughout the dictionary, multi-word expressions in the early volumes must be identified and changed into sublemmas. A case in point is the phrase *bita i graset* ≈ 'bite the dust' (literally 'bite in the grass'), seen in figure 1. The phrase appears as an example under sense 1a in the entry **bita**, but clearly deserves the status of sublemma in a revised version the entry.



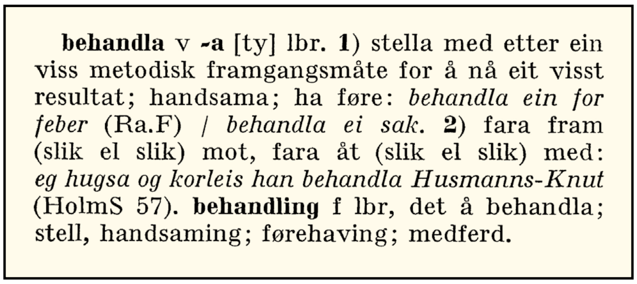
I **bita** (*í*) v **bit, beit, biti** [VAgd16-,L,L 280,A, R3; målf òg ft pl *bito* (Hall), *beto* (Va); fp òg *bite*, *bete*; gno *bíta* (*bít*, *beit*, *bitum*, *bitit*); grunntyd 'kløyva'; tyd 6 vel av eng. *beat* 'slå']. **1**) **a**)
[...]
// *bita i graset*. 1) (eigl om hest som stoggar og tek seg ei grastugge når han møter folk; overf om folk) stogga og tala til ein som ein råkar (Rog): *han Ola for forbi meg utan å bita i graset*. 2) bita i bakken; tapa: *dei lyt bita i graset som hev minste makti* (RomsOrdt)

Figure 1: Part of the entry *bita* with the multi-word expression *bita i graset.*

The structure of senses will – particularly in longer entries – need to be made flatter, more transparent and thus easier to navigate. The fact that the editors have access to a much larger body of linguistic data today (including a ~100 mill. word corpus) than when the early volumes were produced has contributed to less tiered sense structures in the latter volumes, and this will necessarily also be the case for the early material after revision.

There are a lot of structural features where the early volumes differ from today's editorial practice, and where structural revisions along the lines of the present editorial guidelines are required. One example concerns the use of usage labels; certain labels are no longer in use, such as *lbr* (*lite brukande* ≈

'should be used with caution'), which is connected with a certain puristic inclination in the early years of the dictionary. Figure 2 shows two entries with this label from volume 1:



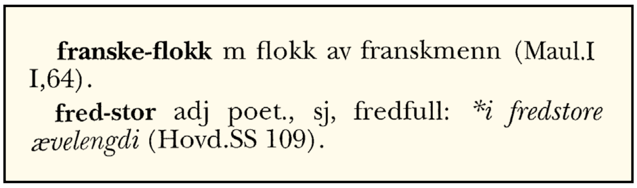
behandla v -**a** [ty] lbr. **1**) stella med etter ein viss metodisk framgangsmåte for å nå eit visst resultat; handsama; ha føre: *behandla ein for feber* (Ra.F) / *behandla ei sak*. **2**) fara fram (slik el slik) mot, fara åt (slik el slik) med: *eg hugsa og korleis han behandla Husmanns-Knut* (HolmS 57). **behandling** f lbr, det å behandla; stell, handsaming; førehaving; medferd.

Figure 2: The entries *behandla* 'treat' and *behandling* 'treatment' are equipped with the label *lbr,* although they are widespread in modern Nynorsk.

Another example concerns the labels *zool* (zoology) and *bot.* (botany), which were earlier used for all definitions covering names of animals and plants respectively, but are today restricted to official terms, while e.g. local names for plants lack the label *bot.*, but are electronically linked to the official term.

3.4 Revision of the lemma list

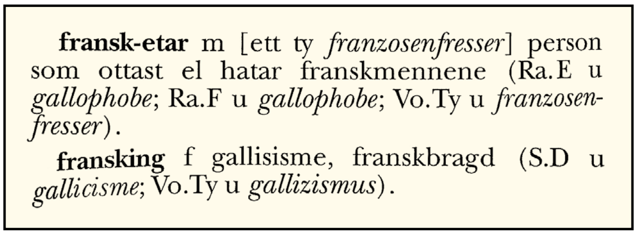
Faced with the task of producing a definite number of volumes on the basis of a certain amount of data, the project NO 2014 has developed effective methods for determining which lemmas should be included and how much space each entry should occupy (Grønvik 2006). The existing lemma list in volumes 1–5 must be revised using the present criteria for inclusion in NO and taking into account the material we have at our disposal today, which is a lot larger than when the first volumes were edited and includes a corpus dominated by 21st century newspaper texts. Neologisms and words that were previously not represented or poorly represented in the material must be included, together with lemmas of German or Danish provenance that were left out for puristic reasons but are used in modern Nynorsk (cf. section 1). In other cases lemmas that were originally included must be excluded – especially in volumes 3–4, where the inclusion criteria were clearly more liberal than today. Thus one can fairly frequently find entries that are based on hapax legomena (figure 3) or exclusively on occurrences in bilingual dictionaries (figure 4). These entries do not qualify for inclusion in the dictionary according to the present editorial guidelines.



franske-flokk m flokk av franskmenn (Maul.II,64).

fred-stor adj poet., sj, fredfull: **i fredstore ævelengdi* (Hovd.SS 109).

Figure 3: Entries in volume 3 based on hapax legomena.



fransk-etar m [ett ty *franzosenfresser*] person som ottast el hatar franskmennene (Ra.E u *gallophobe*; Ra.F u *gallophobe*; Vo.Ty u *franzosenfresser*).

fransking f gallisisme, franskbragd (S.D u *gallicisme*; Vo.Ty u *gallizismus*).

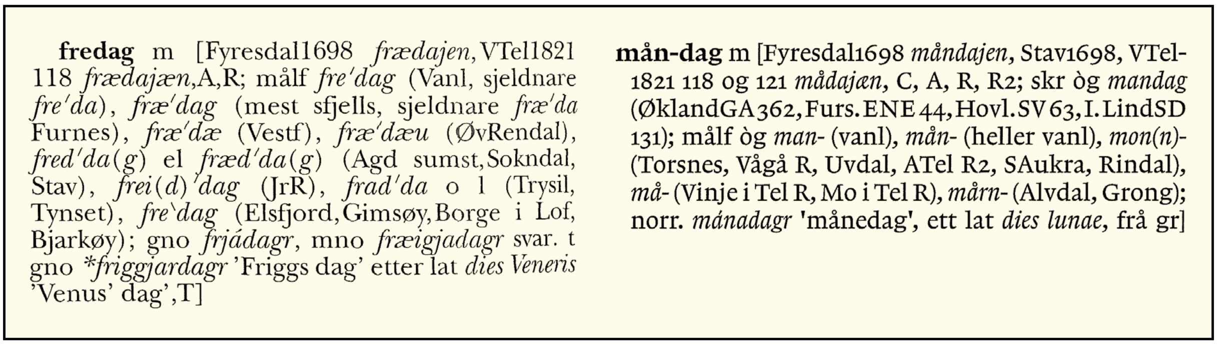
Figure 4: Entries in volume 3 based exclusively on occurrences in bilingual dictionaries.

The oldest entries of the dictionary are not more than some 70 years old. This means that the diachronic dimension of the work itself is less challenging than for dictionaries with a production period that stretches over more than one century. Still, the oldest parts of NO show that some entry revision is needed. New entries have to be added, some old entries should be removed altogether and a lot of existing entries need revision due to broader and sounder empirical evidence, language change or both.

3.5 Revision of the dialect data given in the entries

The dialect material at the editors' disposal is substantially larger today than 80 or even 30 years ago. The geographical indications regarding special dialect forms and dialectal uses of words and word senses can thus be supplemented, in many cases possibly justifying the use of larger areas instead of single counties (the county is the smallest unit used for geographical references in NO). At the same time, the geographical indications in some of the early volumes reflect a more liberal practice than the one followed today, and they must be checked to make sure that the dictionary reflects the actual dialect material at our disposal.

The method of presenting dialect forms has changed somewhat during the history of NO; in particular, volumes 3–4 present such forms in greater phonetic detail and with more parallel forms than both the earlier and the latter volumes (cf. figure 5). Here the revision must imply a certain degree of simplification, following methods established in 2002 and later.



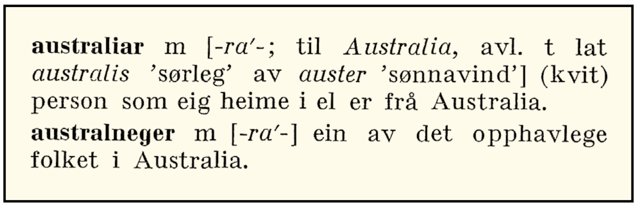
fredag m [Fyresdal1698 *frædajen*, VTel1821 118 *frædajæn*,A,R; målf *fre'dag* (Vanl, sjeldnare *fre'da*), *fræ'dag* (mest sfjells, sjeldnare *fræ'da* Furnes), *fræ'dæ* (Vestf), *fræ'dæu* (ØvRendal), *fred'da*(*g*) el *fræd'da*(*g*) (Agd sumst, Sokndal, Stav), *frei*(*d*)*'dag* (JrR), *frad'da* o l (Trysil, Tynset), *fre`dag* (Elsfjord, Gimsøy, Borge i Lof, Bjarkøy); gno *frjádagr*, mno *fræigjadagr* svar. t gno **friggjardagr* 'Friggs dag' etter lat *dies Veneris* 'Venus' dag',T]

mån-dag m [Fyresdal1698 *måndajen*, Stav1698, VTel-1821 118 og 121 *mådajæn*, C, A, R, R2; skr òg *mandag* (ØklandGA362, Furs.ENE44, Hovl.SV63, I.LindSD 131); målf òg *man-* (vanl), *mån-* (heller vanl), *mon*(*n*)- (Torsnes, Vågå R, Uvdal, ATel R2, SAukra, Rindal), *må-* (Vinje i Tel R, Mo i Tel R), *mårn-* (Alvdal, Grong); norr. *mánadagr* 'månedag', ett lat *dies lunae*, frå gr]

Figure 5: The introductory section of the entry *fredag* 'Friday' from volume 3 and that of *månd-ag* 'Monday' from volume 8. Note the differences in the notation of dialect forms (introduced by 'målf' and 'målf òg' respectively).

3.6 Revision of definitions

In the majority of the entries the actual wording of the definitions will hardly require a lot of revision. Still, since the publication of volume 1 in 1966 the language has certainly undergone quite a few changes on the level of semantics and pragmatics – changes that must necessarily lead to adjustments in a number of definitions. Obvious examples are words that were earlier used neutrally, but later developed derogatory connotations and often have become obsolete altogether. In figure 6, the definition of *australiar* 'Australian' is '(white) person belonging in or coming from Australia'; the first word in the definition should definitely be deleted. The second headword, *australneger*, is no longer in use due to its derogatory character. As NO is a descriptive dictionary which documents actual (historical) usage, the entry should be preserved, but the definition must be updated to reflect the stylistic properties of the words and the fact that it is obsolete. The modern neutral term *aboriginar* 'Aborigine, Native Australian', which is not found in volume 1, must also of course be added.



australiar m [-*ra'*-; til *Australia*, avl. t lat *australis* 'sørleg' av *auster* 'sønnavind'] (kvit) person som eig heime i el er frå Australia.

australneger m [-*ra'*-] ein av det opphavlege folket i Australia.

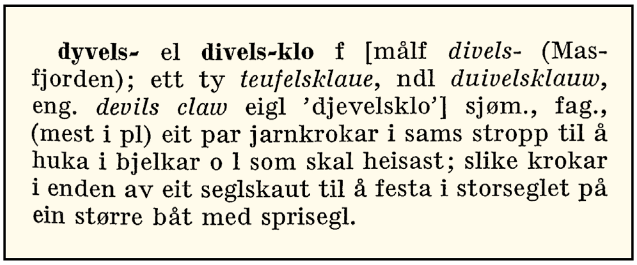
Figure 6: The entries *australiar* 'Australian' and *australneger* 'Australian negro' from volume 1.

On a more general level, there is a tendency in some of the early material towards focusing on the particular rather than the general and to posit separate word senses where the present practice would prefer lumping rather than splitting. Thus especially in longer entries there will be a need for revising or rewriting definitions. Entries that lack a definition altogether but meet the criteria for inclusion in the dictionary must of course be provided with a definition.

Integrating new source material and meeting the requirements of a modern scholarly dictionary

The new source material – including corpus data – must be integrated at all levels of the early volumes of the dictionary. This will be reflected in the addition of new entries (cf. 3.4), the creation of new senses in existing entries, the introduction of new examples, especially the addition of more recent examples (sometimes due to reasons of space and clarity replacing some of the existing examples) as well as in new, updated geographical indications (cf. 3.5).

It is essential to ensure that the early volumes meet the requirements of a modern scholarly dictionary. This implies, firstly, that every entry must be linked to its source material and, secondly, that all entries and all word senses must have a documented source material behind them and contain at least one source reference at the level of definition and/or example. In the electronic version of the dictionary the links between entries and source material will be made explicit, enabling users to verify the information given and potentially falsify it.



dyvels- el **divels-klo** f [målf *divels-* (Masfjorden); ett ty *teufelsklaue*, ndl *duivelsklauw*, eng. *devils claw* eigl 'djevelsklo'] sjøm., fag., (mest i pl) eit par jarnkrokar i sams stropp til å huka i bjelkar o l som skal heisast; slike krokar i enden av eit seglskaut til å festa i storseglet på ein større båt med sprisegl.

Figure 7: The definition in the entry *dyvelsklo* 'devil's claw (a kind of split hook)'from volume 2 lacks source references. One or more references must be added, or the entry must be excluded from the dictionary.

4 Planning and implementing the revision project

As part of the planning it must be decided to what degree the dictionary entries should be rewritten. A plausible strategy is to assume that smaller entries – which constitute the majority – require only revision, while at least a part of the larger entries (especially large verbs and function words) will benefit from being re-edited. This re-editing must be performed with the editor at all times keeping a keen eye on the existing entry and making sure that all essential information that is given there and can be verified is transferred to the new version.

In the modern NO 2014 organisation, all the relevant source material is digitised and stored in a structured relational database system. This makes it possible to quantify relative space for each entry and to estimate the work load for the staff as a group and for each single editor. The experience from the last 12 years of project work shows that this way of working gives a high degree of prediction when it comes to how much time and money are needed to perform the whole operation of revising and digitising the oldest parts of the dictionary.

The whole of the source material behind the earliest volumes is included in the dictionary database system, and this provides a very sound way of estimating the work load for doing the integrated digitisation and revision work. For the whole bulk of 112,500 lemmas it is possible to make fairly accurate estimates that also take into account that some entries will be revised, while others will gain from a full rewriting. Based on experience with producing the last seven volumes of the dictionary over a period of 12 years, feeding both a printed publication (each volume includes 800 pages of entries) and an Internet version, the NO 2014 organisation estimates that the digitisation and revision of the first volumes will be possible with a staff of 16 editors working full time over a period of five years. This is approximately 45 % of the amount of work that was put into volumes 6–12.

5 Funding

The revision project will have a total cost of some 70 million NOK (approx. 8.5 million EUR). The production of NO has so far been funded by the University of Oslo and the Norwegian Ministry of Culture in a joint agreement, but this funding ends in 2014. Language infrastructure, including dictionaries, is cost-intensive and involves huge amounts of manual work. Norway is a relatively small language community, and the commercial potential of the basic language infrastructure resources for Norwegian is quite low. This means that in order to reach the central goals on the field of Norwegian language policy, the building up of basic language resources needs public funding.

A NO dictionary database covering the whole alphabet span will not only offer the public a comprehensive description of spoken Norwegian and written Nynorsk. The full dictionary database will also be an important component in future Norwegian language infrastructure and language technology. In this perspective, public funding of the digital integration of volumes 1–5 of Norsk Ordbok in the dictionary database would hopefully be within reach. In 2013 the Language Council of Norway set up a policy document for dictionaries and other basic lexical resources for the Norwegian languages, including Sami and the official minority languages of Norway. This policy document states the importance of a complete and updated online version of NO, and also states that this needs public funding (LCN 2013-08 and LCN 2014-03).

6 Conclusions

A lot of work is needed to bring the oldest volumes of NO up to the same digital standard as the rest of the dictionary. During the 80 years that have passed since the dictionary work started, the language itself, linguistic theory and preferred publishing platform have all changed. These changes have in turn led to changes in lexicographical practice. For a scholarly dictionary to be scientifically sound and relevant to the dictionary users, it is necessary to revise and upgrade its contents. For the dictionary database to become complete, it is not an option to choose only digitisation, or doing the process in two separate steps.

7 References

Aasen, I. (1850). *Ordbog over det norske Folkesprog*. Christiania: Carl C. Werner & Comp.

Aasen, I. (1873). *Norsk Ordbog med dansk Forklaring*. Christiania: P.T. Mallings Boghandel.

Atkins, B.T.S. (2008). Theoretical Lexicography and its Relation to Dictionary-Making. In: T. Fontenelle (ed.) *Practical Lexicography. A Reader*. Oxford: Oxford University Press, pp. 31-50.

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

Cantell, I. & Sandström, C. (2012), Hur blir en traditionell, tryckt ordbok en webbordbok? In: B. Eaker, L. Larsson, A. Mattisson (eds.) *Nordiska studier i lexikografi 11.* Rapport från Konferensen om lexikografi i Norden Lund 24–27 maj 2011, pp. 157-168.

Grønvik, O. (2006) Verknader av digitalisering på materialvurdering, redaksjonell metode og opplæring. In: *Nordiske Studier i Leksikografi*, 8, pp. 129-142.

Hellevik, A. (1966) Til fyrste bandet. In: NO volume 1, pp. XV-XVI.

LCN 2013-08 = En samlet ordbokpolitikk etter 2014 (letter from the Language Council of Norway to the Ministry of Culture). Accessed at: http://bit.ly/1i06QPN [08/04/2014].

LCN 2014-03 = Norsk ordbokpolitikk (memorandum from the Language Council of Norway to the Ministry of Culture). Accessed at: http://bit.ly/OBoRYU [08/04/2014].

NO = (1966-) *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet.* Oslo: Det Norske Samlaget. Web edition accessed at: http://no2014.uio.no [02/04/2014].

ODS = (1918-2005) *Ordbog over det danske Sprog.* København: Gyldendal. Web edition accessed at: http://ordnet.dk/ods [02/04/2014].

ODS FBTS = Fra bog til skærm. Accessed at: http://ordnet.dk/ods/fakta-om-ods/fra-bog-til-skerm [08/04/2014].

OED = *Oxford English Dictionary.* Accessed at: http://www.oed.com [02/04/2014].

SAOB = (1898-) *Ordbok över svenska språket.* Lund: Svenska Akademien.

Skard, S. (1932) Norsk Ordbok. Historie – plan – arbeidsskipnad. Oslo: Det Norske Samlaget.

A Dictionary Guide for Web Users

Valeria Caruso, Anna De Meo
University of Naples 'L'Orientale'
vcaruso@unior.it, ademeo@unior.it

Abstract

The continuously growing number of specialised lexicographical resources on the Web calls into question the users' ability to solve their information needs autonomously. Neither terminological databanks, nor dictionary aggregators actually represent valuable solutions to these needs (Lew 2011), and in order to guide users towards useful resources, a database was created, which collects evaluation forms of free internet specialised dictionaries and allows users to carry out customised searches on the basis of their subject field expertise (laymen, semi-experts, experts), the desired language, and the kind of support they need (basically with communicative problems or in acquiring new knowledge).

Using a specific evaluation system, the tool displays the best resources available for the desired parameters, assessing dictionaries on the basis of an evaluation scale and explicit guidelines that prevent contradictory responses, such as dictionaries that are simultaneously suited for laymen and experts.

The paper illustrates the current development of the tool, with special reference to its evaluation system, as well as its possible future improvements.

Keywords: Dictionary Guides; Dictionary evaluation; Online Dictionaries; Specialised lexicography

1 Information overload on the Web

Though the Web is the most used source of information, too much data are offered to the Internet surfers, causing what has been called "information death" (Tarp 2010: 41). Search engines are too generic to be of any assistance to users with these tasks, and metalexicographical resources have started to appear. The quickest searches are offered by dictionary aggregators (i. e. *OneLook*) and mesh-ups (i. e. *Your Dictionary*) which show definitions taken from different vocabularies on one page, a system that doesn't seem to be completely effective, because terminology archives - and hence the number of definitions provided - are either too small to cope with users' needs, or too big to solve the problem of an effective and efficient access to information (Heid 2011).

From this point of view, the World Wide Web poses stimulating metalexicographical issues, some of which will be outlined here while presenting a new lexicographical tool for guided searches on the Web, namely a rated inventory of free specialised dictionaries, managed through a relational database which allows users to carry out multiparametric searches.

1.1 Information accessibility

The issue of knowledge accessibility led to the creation of dictionaries, since:

> (t)he truly unique thing about dictionaries is not the various types of data they employ in covering the information needs of users (…). Such data can generally be incorporated into other types of book and text as well. The truly unique thing is the way in which this data is made accessible so users can quickly and easily find the exact data they need. (Tarp 2008: 101)

Nowadays lexicographers focus their work on the customization of dictionaries for their users, and different approaches have been proposed in order to achieve this aim. One in particular seems to be useful not only for writing vocabularies, but also for their critical evaluation, since it offers a synthetic procedure to define the parameters a dictionary must have in order to fulfill its desired functions. Therefore the theory has been named *lexicographical function theory*, and was formulated by Sven Tarp (2008; also Tarp 2009, 2010) as a result of long metalexicographical reflections and debates carried out by the research group of the Aarhus University in Denmark (Nielsen 1994; Geeb 1998; Bergenholz & Tarp 1995). According to this theory, dictionary functions must be identified on the basis of the kind of users, as well as the situations in which the vocabulary is employed, therefore the compilers must think about the specific context in which the need for vocabulary consultation arises (Tarp, 2008: 81). For example, dictionaries may be used in many different situations, such as by students proofreading their homework, or by professional editors working on books to be published, or even in the less common situation of young people reading religious books, in such a case the dictionary "should only explain the meaning of a word or of phrase and noting more" (Bergenholz, 2012: 245). Therefore the more specific the target is, the easier it is to tailor the dictionary to the users' desired functions. As a consequence, the traditional general language dictionaries (or *polyfunctional dictionaries*), offering aid for different kinds of tasks without a specific tailoring of the information they provide are judged as inefficient, since:

> (they) are in many cases so overloaded that this causes information stress and in the worst case may even cause the search to be abandoned if the user cannot find the needle in the haystack (Bergenholz, 2012: 251).

The alternative model proposed is the *monofunctional* electronic vocabulary, extracted from lexical databases using search forms that allow users to tailor the entry to their needs. For example, if the dictionary must supply assistance for text production in an L2, the database will provide a dictionary article displaying grammar information, "synonyms, collocations and examples" (Bergenholz, 2012: 253). Conversely, if the user must understand a text, this information is probably inadequate and certainly not of the outmost important.

Lastly, by fixing explicit parameters that guide good practices of dictionary writing, the theoretical framework of the *lexicographical functions* proves to be suited for the opposite task too, namely dictionary evaluations, which can be undertaken not only in general review terms (see Nielsen, 2009, 2013), but also in a more lexicographical direction, employing the same principles as orienteering parameters among the existing lexicographical resources.

Using these observations as a starting point, a database has been created. The resource, accessible at the *Web Linguistic Resources* (WLR) site, collects free specialized Internet dictionaries which are often more valuable for their unrestricted access than for their overall quality, since the Internet compilers have little or no lexicographical expertise at all. The usability of the majority of these dictionaries is therefore dependent on guides and filters that prevent users from wasting their time and being given inefficient information, in this way they might become quick reference tools for web surfers.

The archived dictionaries were collected during two extensive research sessions in 2010 for the sector of oenology and medicine in different languages: Italian, English and French. A similar intensive exploration was carried out in 2013 for Economics dictionaries of the English language, whilst other sporadic additions gave the database more resources from different specialised sectors on the basis of more occasional findings. A more systematic analysis and upgrade of the inventoried resources is planned to be carried out before the definitive version of the tool is released, since it is currently available only as a pivotal 'beta' version.

2 Dictionaries on the Web: the features to be rated

Instead of providing users with multiple definitions on one page, and leaving them with the task of selecting data, the WLR database offers a rated inventory of dictionaries which help users to find the best resources available for free on the Web.

Moreover, the adaptation of the lexicographical function theory parameters to critical principles of analysis in order to rate and filter dictionaries also fulfills the proposal of Nielsen (2009; 2013) to judge dictionaries on lexicographical principles that are generally applicable in order to make dictionary reviews an integral part of the academic field of lexicography.

The rated inventory of the WLR site is based on an evaluation form (fig. 4 below), managed by a relational database that allows multiparametric searches.

The 53 fields in the form (see table 1) correspond to the possible features of a dictionary, and address all the component parts of vocabularies, i. e. the overall organization and the host site, the mediostructure (Wiegand, 1996; Nielsen, 2003), and microstructure (Hausmann & Wiegand, 1989; Hartmann, 2001). The features were partly set in advance, and partly added - or modified - during the data collection, in order to portray adequately the characteristics of these atypical dictionaries - they are listed in table 1 according to the parts of dictionaries they belong to[1].

[1] See also Caruso [2011] and Caruso & Pellegrino [2012] for a more detailed description of the features considered.

Dictionary parts	Features and sub-features
General Organization and Host Site	Guide, Kind of Site: Amateur/ Blog/ Commercial/ Collective Resource/ Generalist/ Institutional/ Specialised, Learning Resources, Bibliographic Resources, Hyperlinks, User Feedback, Access: Browse / Search Engine / Advanced Search Engine, Entries: 0-49 / 50-100 / Over 100, General Organisation: Concepts / Words, Kind of Dictionary: Monolingual Dictionary/ Monolingual Word List/ Multilingual Dictionary/ Multilingual Word List/ Plurilingual Dictionary, Bidirectionality, Lemmata: Technical And Non-Technical Terms / Only Technical Terms;
Mediostructure	Cross-references, Related terms, Hypernyms & Hyponyms, Hypertexts;
Microstructure: Linguistic fields	Grammatical Category, Morphological Information, Syntactic Pattern, Phonetic Transcription, Pronunciation Notation, Stress Information, Audio Files, Syllabification, Frequency Of Use, Linguistic Variation, Technical Definitions, Translation Equivalences, Example Sentences, Quotations, Idioms, Collocations, Synonyms, Antonyms, Etymology;
Microstructure: Non Linguistic Fields	Definitions, Examples, Domain Field, Video Files, Pictures, Cultural Notes.

Table 1: the listing of the dictionary features and sub-features assessed by the evaluation form of the *Web Linguistic Database.*

The host site may be an important validation criteria of the dictionary quality, since it is to be expected that credited Institutions (universities, ministries, professional associations etc.) publish good lexical resources. In point of fact, 'institution' refers here to authoritative organizations within one field, and it has a more restricted use than in Fuertes-Olivera (2009), where the term refers generically to every dictionary not compiled 'collectively' by non-professional lexicographers (such as *Wiktionary*).

The overall organization, instead, comprises the dictionary type, whether a simple word list, a multilingual dictionary provided or not with bidirectionality (which is a separate field in the form), or a *plurilingual*, a new dictionary added to the list which is typical of the Internet, namely the dictionary within localized sites (Caruso 2011). These sites in fact are optimized for the market of different countries (Pym 2004), and therefore offer many language versions of their pages that are not interlinked with each other. Since one version is completely independent from the others, the many language dictionaries therein also have no direct connection. Therefore, the user must scan the entire word list and check for correspondences in the definitions in order to find any translation equivalences.

Moreover, Internet dictionaries may also offer special access facilities to users, such as advanced search engines (another field in the form). For example, the dictionary of the *Büro für angewandte Mineralogie* allows searches not only in the whole dictionary contents, but also in its classifying ontology: looking for *Elemente*, the listing provided by the engine will include also *Periodensystem der Elemente*, besides all the chemical elements in the dictionary (from *Antimony* to *Sulfur*), and the entries that contain the required word in their definitions.

During the data collection, special attention has also been paid to the mediostructure, or the cross-linking system, which is obviously a key component of electronic vocabularies. Accordingly, the evaluation form registers both *Cross-references* and *Related terms* (see table 1), only the former having direct hyperlinks to other entries, while *Hypernyms and hyponyms* signal semantic hierarchies that also function as internal references.

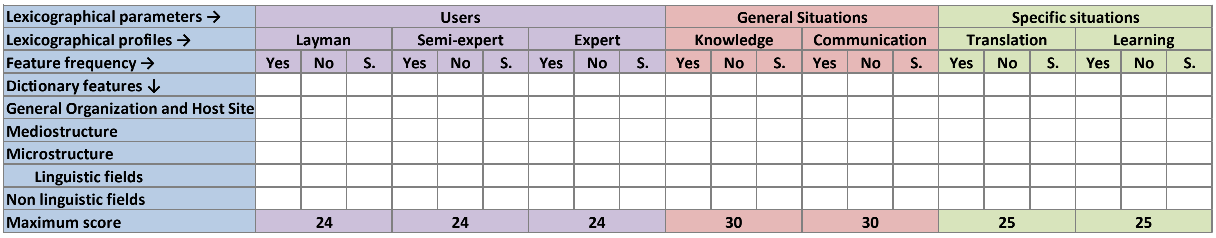
As for the microstructure, or the dictionary entry components, the evaluation form takes note of its linguistic and encyclopedic aspects, and accounts for specific fields that reveal the user-friendly character of these dictionaries, which generally offer non-technical definitions, and pronunciation notations rather than phonetic transcriptions.

3 The rating system

Since this lexicographical project does not aspire to the detailed dictionary reviews of Nielsen (2009, 2013), but to large scale qualitative estimations that filter dictionaries of poor quality or, at least, dictionaries not suited for a specific function, we limit the critical system to a few lexicographically relevant situations and only some types of users.

The most general situations of dictionary use are, according to Tarp, communicative and cognitive contexts in which someone needs to produce texts or know something - in the database we name them *Communication* and *Knowledge*. To these we add two others, which are more specific and are expected to be the most typical for web surfers: contexts in which someone needs to translate (*Translation* in the database) or learn something (*Learning*). Therefore our inventory is made up of three *lexicographical parameters*: three kinds of users, two general and two specific consultation situations (see fig. 1). The kind of user parameter is thus limited to laymen, experts, and semi-experts of one field, e. g. economy journalists who are not economists themselves (Bergenholtz & Kaufmann, 1997; Hartmann, 1989).

To the parameters, feature frequency (see fig. 1) has been added, in order to keep track of the features that are always present and those which occur only sometimes in one dictionary, since the majority of these lexicons lack any strict lexicographical organization, and offer unsystematic assistance to users.



Lexicographical parameters →	Users									General Situations						Specific situations					
Lexicographical profiles →	Layman			Semi-expert			Expert			Knowledge			Communication			Translation			Learning		
Feature frequency →	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.
Dictionary features ↓																					
General Organization and Host Site																					
Mediostructure																					
Microstructure																					
Linguistic fields																					
Non linguistic fields																					
Maximum score	24			24			24			30			30			25			25		

Figure 1: Lexicographical parameters *(Users, General Situations, Specific Situations),* lexicographical profiles (*Layman, Semi-Expert, Expert, Knowledge, Communication, Translation, Learning*), and dictionary features (addressing the *General Organization, Mediostructure, Microstructure*) with their occurrence frequency (Yes, No, S.= Sometimes).

On this basis, the features considered to be more relevant (Bothma & Tarp 2012) for one parameter receive 1 or 2 points score, conversely, negative scores (-1, -2) are given to those judged as contradictory. Thus the evaluation scale is made as follows:

- 2 points to the most relevant features
- 1 point to relevant features
- -1 to contradictory features
- -2 to the most contradictory features

The specifics of each lexicographical parameter determine what we call here a *lexicographical profile*, which is outlined by its characterizing features, as it is displayed in table 2 below.

Lexicographical profile	**Features and scores**
Layman	Institutional Site: Yes (2); Specialised Site: Yes (1); Technical and non-technical terms: Yes (2); Cross-references: Yes (2); Related terms: Yes (1); Hypernyms & Hyponyms: Yes (1); Pronunciation notation: Yes (1); Stress information: Yes (1); Audio files: Yes (2); Technical definitions: Yes (-2), No (2); Example Sentences: Yes (2), No (-2), Sometimes (1); Quotations: Yes (-2), Sometimes (-1); Synonyms: Yes (2), Sometimes (1); Antonyms: Yes (2), Definitions: Yes (2), Sometimes (1); Examples: Yes (2), Sometimes (1), Video files: Yes (2), Sometimes (1), Pictures: Yes (2), Sometimes (1).
Semi-Expert	Institutional Site: Yes (2); Specialised Site: Yes (1); Bibliographic resources: Yes (1) No (-1); Hyperlinks: Yes (1); Access: Advanced search engine: Yes (1); Entries: 0-49: Yes (-2); Entries: 50-100: Yes (-2); Technical and non-technical terms: Yes (1); Cross-references: Yes (1); Related terms: Yes (1); Hypernyms & Hyponyms: Yes (1); Phonetic transcription: Yes (1); Syllabification: Yes (1); Linguistic variation: Yes (2), Sometimes (1); Technical definitions: Yes (1) No (-1); Quotations: Yes (2); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), Sometimes (1); Etymology: Yes (1), No (-1); Definitions: Yes (1); Domain field: Yes (1); Pictures: Yes (1).
Expert	Institutional Site: Yes (2); Specialised Site: Yes (1); Bibliographic resources: Yes (2), No (-2); Hyperlinks: Yes (2); Access: Browse: Yes (-1); Entries: 0-49: Yes (-2); Entries: 50-100: Yes (-2); Hypertexts: Yes (1); Phonetic transcription: Yes (2), Sometimes (1); Syllabification: Yes (2), Sometimes (1); Linguistic variation: Yes (2), Sometimes (1); Technical definitions: Yes (2), No (-2), Sometimes (); Quotations: Yes (2), No (-2), Sometimes (1); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), No (-2), Sometimes (1); Etymology: Yes (2), No (-1), Sometimes (1); Domain field: Yes (1).
Knowledge	Institutional Site: Yes (2); Specialised Site: Yes (1); Bibliographic resources: Yes (2); Hyperlinks: Yes (2); Cross-references: Yes (2); Related terms: Yes (2); Sometimes (1); Hypernyms & Hyponyms: Yes (2), Sometimes (1); Hypertexts: Yes (2); Quotations: Yes (2); Sometimes (1); Etymology: Yes (2), Sometimes (1); Definitions: Yes (2), Sometimes (1); Examples: Yes (2), Sometimes (1); Domain field: Yes (2), Sometimes (1); Video files: Yes (2), Sometimes (1); Pictures: Yes (2), Sometimes (1); Cultural notes: Yes (2).
Communication	Institutional Site: Yes (2); Specialised Site: Yes (1); Technical and non-technical terms: Yes (2); Grammatical category: Yes (2), Sometimes (1); Morphological information: Yes (2), Sometimes (1); Syntactic pattern: Yes (2), Sometimes (1); Phonetic transcription: Yes (2), Sometimes (1); Pronunciation notation: Yes (1); Stress information: Yes (1); Audio files: Yes (2), Sometimes (1); Syllabification: Yes (1); Frequency of use: Yes (1); Linguistic variation: Yes (2), Sometimes (1); Example Sentences: Yes (2), Sometimes (1); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), Sometimes (1); Synonyms: Yes (2), Sometimes (1); Antonyms: Yes (2), Sometimes (1).

Translation	Institutional Site: Yes (2); Specialised Site: Yes (1); Multilingual dictionary: Yes (2); Multilingual word list: Yes (1); Plurilingual dictionary: Yes (2); Bidirectionality: Yes (2); Technical and non-technical terms: Yes (2); Grammatical category: Yes (1); Morphological information: Yes (1); Syntactic pattern: Yes (2), Sometimes (1); Linguistic variation: Yes (2), Sometimes (1); Translation equivalences: Yes (2), No (-2), Sometimes (1); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), Sometimes (1); Cultural notes: Yes (2).
Learning	Institutional Site: Yes (2); Specialised Site: Yes (1); Learning resources: Yes (2), No (-2); Bibliographic resources: Yes (2); Hyperlinks: Yes (2); Monolingual dictionary: Yes (2); Multilingual dictionary: Yes (2); Related terms: Yes (1); Grammatical category: Yes (1); Morphological information: Yes (1); Syntactic pattern: Yes (1); Audio files: Yes (2), Sometimes (1); Example Sentences: Yes (2), Sometimes(1); Definitions: Yes (2), Sometimes(1); Examples: Yes (2), Sometimes(1); Video files: Yes (1);Pictures: Yes (2), Sometimes(1).
Lexicographical profile	Features and scores

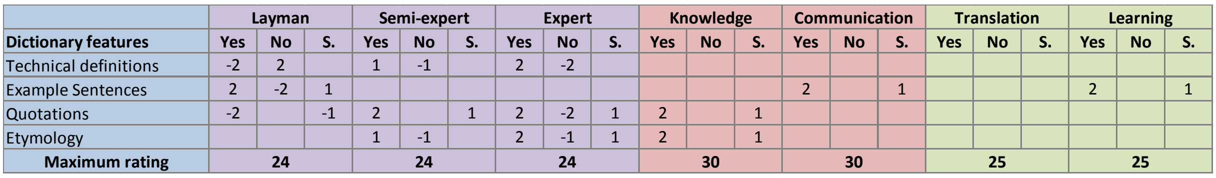
Table 2: Score assignment in the evaluation system of the *Web Linguistic Resources* database. Specific features receive different scores and outline the different *lexicographical profiles* considered.

In addition, the scores were given the following basic guidelines:

1) profiles belonging to the same lexicographic parameter may reach the same maximum score;
2) complementary profiles don't share the same features;
3) similar profiles may share the same features.

According to the first rule, user profiles may reach 24 points maximum each, general situations 30, and more specific consultation situations 25 (see fig. 1 and fig. 2).

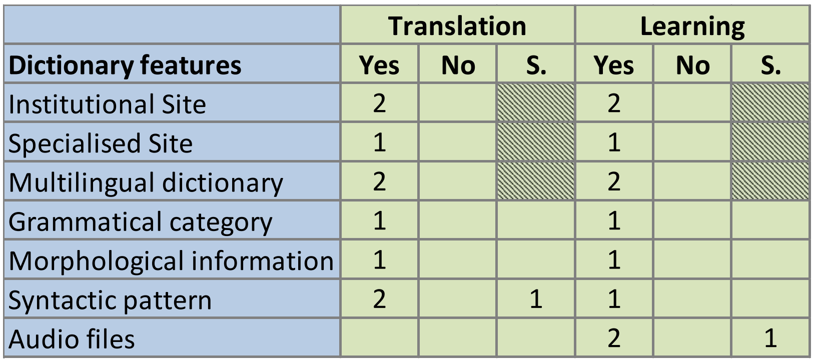
The second principle, however, prevents the database from giving contradictory responses, such as dictionaries suited for laymen and experts at the same time. Therefore, referring to figure 2 below, technical definitions are required in the vocabularies for experts (2 points), but not in those for layman (-2). The opposite is also true: if a dictionary doesn't have technical definitions, it is suited for laymen (2) but not for experts (-2). Similarly, example sentences are expected in dictionaries for laypeople, and quotations in those for experts.



	Layman			Semi-expert			Expert			Knowledge			Communication			Translation			Learning		
Dictionary features	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.
Technical definitions	-2	2		1	-1		2	-2													
Example Sentences	2	-2	1										2		1				2		1
Quotations	-2		-1	2		1	2	-2	1	2		1									
Etymology				1	-1		2	-1	1	2		1									
Maximum rating	24			24			24			30			30			25			25		

Figure 2: Score giving to features according to the different profiles (*Layman*, *Semy-Expert*, *Expert*, *Knowledge*, *Communication*, *Translation* and *Learning*) and their occurrence frequency (Yes, No, S.=Sometimes).

On the contrary, the second guideline states that if the profiles are similar, they can share features and scores, such as a specialized host site and information on syntactic patterns for the translation and learning situations (see fig. 3).



	Translation			Learning		
Dictionary features	**Yes**	**No**	**S.**	**Yes**	**No**	**S.**
Institutional Site	2		▨	2		▨
Specialised Site	1		▨	1		▨
Multilingual dictionary	2		▨	2		▨
Grammatical category	1			1		
Morphological information	1			1		
Syntactic pattern	2		1	1		
Audio files				2		1

Figure 3: Features in common for the *Translation* and *Learning* profiles.

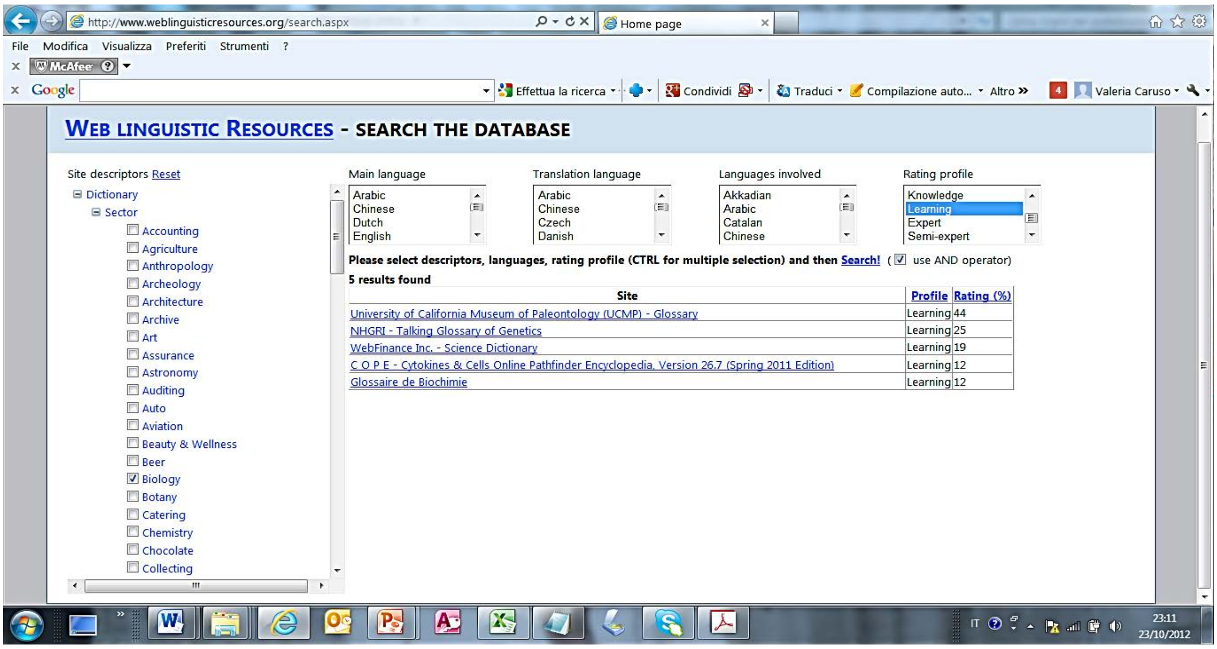
The evaluation procedure adopted is thus purely proscriptive (Andersen & Nielsen 2009), and based on the careful distribution of scores among the profiles inventoried in order to fulfill the requirement of the guidelines stated above. This should guarantee a balanced critical assessment procedure, minimizing the possibility that some profiles are easier to fulfill because they require lower maximum scores. Consequently, even though the comparative methodology used for the distribution of grades among the different profiles is paramount and not dismissible (Caruso forthcoming), at least one test on real users has already been carried out in order to check the overall validity of the proposed evaluation system (Caruso & De Meo 2013). In this study, the higher scoring medicine dictionaries of the WLR database for the *Translation* profile were used by 39 university students in a controlled translation session, and despite the overall low-quality of these vocabularies, students who consulted them to overcome some of the main difficulties in the source text performed better than those who translated freely, without referring to any dictionary whatsoever.

Focusing on the post-consultation phase, this small study on real dictionary use is just a starting point for the examinations that may be carried out in order to validate the assessment procedure of the WLR system, and the features that have been chosen to outline each *lexicographical profile*.

4 How to search the database

The features and the lexicographical (or rating) profiles are the main search options of the Web Linguistic Resources database. Accessing the homonymous site, it is possible to search for the dictionary that is best suited to the user's needs. The available options are listed in the center of the page, where the dictionaries ratings are provided as a percentage, since the score gives evidence of the degree to which the dictionary corresponds to the desired profile. Figure 4, for example, shows the search for a dictionary of biology suited for a learning context. The sector "biology" is a subfield within the dictionary features, which are listed on the left, while in the upper right of the page users can choose the rating profile.

The other available search options are the translation languages, the language in which the dictionary is written (*Main Language*), but also other languages present in the entry list (*Languages Involved*), for example French terms in English wine dictionaries or Latin words in German law lexicons.



WEB LINGUISTIC RESOURCES - SEARCH THE DATABASE

Site descriptors Reset

- Dictionary
  - Sector
    - □ Accounting
    - □ Agriculture
    - □ Anthropology
    - □ Archeology
    - □ Architecture
    - □ Archive
    - □ Art
    - □ Assurance
    - □ Astronomy
    - □ Auditing
    - □ Auto
    - □ Aviation
    - □ Beauty & Wellness
    - □ Beer
    - ☑ Biology
    - □ Botany
    - □ Catering
    - □ Chemistry
    - □ Chocolate
    - □ Collecting

Main language: Arabic, Chinese, Dutch, English

Translation language: Arabic, Chinese, Czech, Danish

Languages involved: Akkadian, Arabic, Catalan, Chinese

Rating profile: Knowledge, Learning, Expert, Semi-expert

Please select descriptors, languages, rating profile (CTRL for multiple selection) and then Search! (☑ use AND operator)

5 results found

Site	Profile	Rating (%)
University of California Museum of Paleontology (UCMP) - Glossary	Learning	44
NHGRI - Talking Glossary of Genetics	Learning	25
WebFinance Inc. - Science Dictionary	Learning	19
C O P E - Cytokines & Cells Online Pathfinder Encyclopedia, Version 26.7 (Spring 2011 Edition)	Learning	12
Glossaire de Biochimie	Learning	12

Figure 4: The search form of the *Web Linguistic Resources* database.

5 What remains to be done

At present the evaluation system filters dictionaries only on the basis of their features, according to explicit lexicographical parameters, but it doesn't provide any assurance about the reliability of contents, which nevertheless is one of the most urgent requirements for anyone browsing the Internet. Obviously it is impossible to vouch for the quality of every single piece of information provided by the web dictionaries or by any other dictionary. What is needed is to avoid resources that create problems for users instead of helping them. This is the case with the following explanations related to the enological term "extra dry":

> Extra-Dry
>
> Don't believe everything you read. What this really denotes is a sweet Champagne.
>
> (Pacific Northwest Wine Company. Terminology and Descriptions)
>
> extra dry
>
> adj. Another step on the sweetness-level scale associated with Champagne. Starting on the low end with brut zéro, the scale ascends to brut nature, extra brut, and brut sauvage (all of which are bone-dry), then brut (dry), extra dry (a hint of sweetness), sec (slightly sweet), demi-sec (moderately

> sweet), and doux (the sweetest of all). Why extra dry is sweeter than brut is a mystery to everyone but Francophiles. The only types of sparkling wine you're likely to see at the store are brut, extra dry, and demi-sec, of which brut is far and away the most popular. FYI, table wine that's slightly sweet is referred to as off-dry (*Wine Lovely – Glossary*).

In these examples, the discrepancy between the ordinary value of the adjective *dry* and its meaning in the *extra dry* specialised compound is particularly highlighted, and in the second definition the difference is also underlined using an indirect question: "Why extra dry is sweeter than brut is a mystery to everyone but Francophiles". However no answer is given.

One useful discriminatory criteria might be that of referring to dictionaries published by leading institutions of one field, but whilst browsing the Internet it is possible to collect examples of the lexicographical inexperience of experts responsible for dictionary writing. Firstly, if definitions are not compiled carefully, they can give bad explanations that eventually turn into information voids, this is the case with the entry *Chromosome* of the *Talking Glossary of Genetics*, published by the highly esteemed National Human Genome Research Institute. The definition says that: "Humans have 23 pairs of chromosomes(...), and one pair of sex chromosomes, X and Y", which is misleading, since XY is the chromosome pair of males, while women have XX, as is clearly explained in the voice for *Sex Chromosome*:

> (...) Humans and most other mammals have two sex chromosomes, the X and the Y. Females have two X chromosomes in their cells, while males have both X and a Y chromosomes in their cells (...).

Secondly, sometimes the lack of any strict lexicographical organization prevents exhaustive meaning explanations. For example, the *University of California Museum of Paleontology* explains *Basement Rock* as follows:

> basement rock -- n. The oldest rocks in a given area; a complex of metamorphic and igneous rocks that underlies the sedimentary deposits. Usually Precambrian or Paleozoic in age.

In fact, since no explicit label nor clear text divisions are provided, it is impossible to decide whether the first part of the definition "The oldest rocks in a given area;" is one possible meaning of "basement rock", or if "The oldest rocks in a given area;" is a synonym of the following part of the definition, particularly that which states: "Usually Precambrian or Paleozoic in age".

These brief examples give an idea of the kind of work that remains to be done, but not of the kind of solutions to be provided. In effect, after having established which features of the definition must be rated, two main evaluation options remain: one is to choose a pair of critical terms for each specialized field and analyze their definitions in every vocabulary, the other is to extract at random a fixed number of terms for each resource and provide statistically relevant assessments. Speaking in general terms, the latter option is preferable, since the 'critical' terms of huge fields (e. g. medicine, economy etc.) are too numerous.

Therefore, the most suitable statistical evaluation model for the matter remains to be chosen, provided that the number of the rated definitions remains the same for every vocabulary, regardless of its entry number. Since the number of definitions considered doesn't change, it is necessary to provide each assessment of the dictionary entries with its variation coefficient, i. e. the precision index of the estimation made for the vocabulary considered. It is therefore unsurprising that small dictionaries will be rated more accurately than the big ones.

6 References

Andersen, B., & Nielsen, S. (2009). Ten Key Issues in Lexicography for the Future. In H. Bergenholtz, S. Nielsen, S. Tarp (eds.) Lexicography at a Crossroads. Dictionaries and Encyclopedias today, Lexicographical Tools tomorrow. Bern etc.: Peter Lang, pp. 355-365.

Bergenholtz, H. & Kaufmann, U. (1997). Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. In *Hermes*, 18, pp. 91-127.

Bergenholtz, H. & Tarp, S., eds. (1995). *Manual of Specialised Lexicography*. Amsterdam, Philadelphia: John Benjamins.

Bergenholtz, H. (2012). Concepts for Monofunctional Accounting Dictionaries. In *Terminology*, 18, 2, pp. 243-263.

Bothma, T. J.D. & Tarp, S. (2012). Lexicography and the Relevance Criterion. In *Lexikos*, 22, pp. 86-108.

Büro für angewandte Mineralogie. Accessed at: http://www.a-m.de/deutsch/inhalt.htm [04/01/2014].

Caruso, V. & De Meo, A. (2013). Comunicare i saperi sul Web: il caso dei dizionari specialistici. In C. Bosisio, C. Cavagnoli (eds.) Comunicare le discipline attraverso le lingue: prospettive traduttiva, didattica, socioculturale. *Proceedings of XII AItLA International Congress, Macerata, 23-24 febbraio 2012*. Perugia: Guerra.

Caruso, V. & Pellegrino, E. (2012). Metadizionari digitali specialistici. In S. Ferreri (ed.) *Lessico e lessicologia. Atti del XLIV Conngresso internazionae di studi della Società Italiana di Linguistica (SLI), Viterbo, 27-29 settembre, 2010*. Roma: Bulzoni, pp. 487-495.

Caruso, V. (2011). Online specialised dictionaries: a critical survey. In I. Kosem, K. Kosem (eds.) Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011, Bled, 10-12 November. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 66-75.

Caruso, V. (Forthcoming). A guide for the quality assessment of dictionaries of economics. In P. Leroyer, S. Tarp (eds.) Dictionaries of Economics in the 21st Century: The Challenges of Online Lexicography. Berlin, Boston: De Gruyter.

Fuertes-Olivera, P. A. (2009). The Function theory of lexicography and electronic dictionaries: Wiktionary as a Prototype of Collective Multiple-Language Internet Dictionary. In H. Bergenholtz, S. Nielsen, S. Tarp (eds.) Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographica Tools Tomorrow. Bern etc.: Peter Lang, 99-134.

Geeb, F. (1998). Semantische und enzyklopädische Informationen in Fachwörterbüchern. Eine Untersuchung zu fachinformativen Informationstypen mit besonderer Berücksichtigung wortgebundener Darstellungsformen. In *Hermes*, 21, pp. 205-216.

Hartmann, R. R. K. (1989). Sociology of the Dictionary User: Hypotheses and Empirical Studies. In F. J. Hausmann, O., Reichmann, E. H. Wiegand, L. Zgusta, (eds.) Wörterbücher/Dictionaries/Dictionnaires.

Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/ Enciclopédie internationale de lexicographie. Berlin-New York: De Gruyter, vol. I, pp. 102-111.

Hartmann, R. R. K., (2001). Teaching and Researching Lexicography. Applied Linguistics in Action. Harlow: Longman-Pearson Education.

Hausmann, F. J. & Wiegand, H. E. (1989). Component Parts and Structures of General Monolingual Dictionaries: A Survey. In F. J. Hausmann, O., Reichmann, E. H. Wiegand, L. Zgusta, (eds.) Wörterbücher/ Dictionaries/Dictionnaires. Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/Enciclopédie internationale de lexicographie. Berlin-New York: De Gruyter, vol. II, pp. 328-360.

Heid, U. (2011). Electronic Dictionaries as Tools: Toward an Assessment of Usability. In P. A. Fuertes-Olivera, H. Bergenholtz (eds.) e-Lexicography. The Internet, Digital Initiatives and Lexicography. London, New York: Continuum, pp. 287-304.

Nielsen, S. (1994) The Bilingual LSP Dictionary. Principles and Practice for Legal Language. Tübingen: Narr Francke Attempto Verlag.

Nielsen, S. (2003). Mediostructures in Bilingual LSP Dictionaries. In R. R. K. Hartmann (ed.) Lexicography. Critical Concepts. Lexicography, Metalexicography and Reference Science, London: Routledge, vol. III, pp. 270-294.

Nielsen, S. (2009) Reviewing Printed and Electronic Dictionaries: a Theoretical and Practical Framework. In S. Nielsen, S. Tarp (eds) Lexicography in the 21st Century. In honour of Henning Bergenholtz. Amsterdam, Philadelphia: John Benjamins, pp. 23-41.

Nielsen, S. (2013). A General Framework for Reviewing Dictionaries. In O. Karpova, F. Kartashkova (eds.): Multi-disciplinary Lexicography: Traditions and Challenges of the XXIst Century. Cambridge: Cambridge Scholars Publishing: 145-157.

OneLook. Accessed at: http://www.onelook.com [04/01/2014].

Pacific Northwest Wine Company. Terminology and Descriptions. Accessed at: http://pcnwwineco.com/terminology-and-descriptions [04/01/2014].

Pym, A. (2004). The moving text: localization, translation, and distribution. Amsterdam, Philadelphia: John Benjamins.

Talking Glossary of Genetics. Accessed at: http://www.genome.gov/glossary/index.cfm [04/01/2014].

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, Berlin-New York: De Gruyter.

Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen, S., Tarp (eds.) Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow. Bern: Peter Lang, pp. 17-32.

Tarp, S. (2010). Functions of Specialized Learners Dictionaries. In P. A. Fuertes-Olivera (ed.) Specialised dictionaries for learners. Lexicographica. Series Maior, Berlin-New York: De Gruyter, pp. 39-53.

University of California Museum of Paleontology – Glossary. Accessed at: http://www.ucmp.berkeley.edu/glossary/glossary.html [04/01/2014].

Web Linguistic Resources. Accessed at: www.weblinguisticresources.org [04/01/2014].

Wiegand, H. E. (1996). Textual Condensation In Printed Dictionaries: a Theoretical Draft. In *Lexikos*, 6, pp. 133-158.

Wiktionary. Accessed at: https://en.wiktionary.org/wiki/Wiktionary:Main_Page [04/01/2014].

Wine Lovely – Glossary. Accessed at: http://www.winelovely.com/index.asp?codice=43 [04/01/2014].

Your Dictionary. Accessed at: http://www.yourdictionary.com [04/01/2014].

Acknowledgements

Special thanks are due to Gianluca Monti for providing the project with all the technical assistance needed.

What a Multilingual Loanword Dictionary can be used for: Searching the *Dizionario di italianismi in francese, inglese, Tedesco* (DIFIT)

Matthias Heinz, Anne-Kathrin Gärtig
Universität Salzburg
matthias.heinz@sbg.ac.ag, anne-kathrin.gaertig@sbg.ac.at

Abstract

The paper outlines the structure and practical use of the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT), a dictionary registering Italian loanwords in three languages published in print and currently being prepared for digital reedition. This specialized lexicographical resource focuses on language contact resulting in synchronic and diachronic lexical transfer from the Italian language to French, English and German. Italian as a culturally extremely rich and diverse source for borrowings has extensively influenced other languages in different historic phases all along the past centuries. The DIFIT reflects this rich European heritage of borrowings in a maximal variety of lexical fields. Major methodological aspects in and prior to editing the dictionary included the delimitation and exploration of the source material and deciding on depth of historical coverage as well as defining the information programme of the microstructure. At the same time, limitations of the general lexicographical documentation on both Italian and the three recipient *languages presented particular challenges. The poster displays results of some first quantitative research on* linguistic aspects of the borrowed elements in DIFIT and gives an outlook on the ongoing project of digitizing the lexicographical data by means of an online database documenting the linguistic impact of Italian as a global donor language (*Osservatorio degli italianismi nel mondo*/OIM, hosted by the Accademia della Crusca).

Keywords: Multilingual Lexicography; Specialized Dictionaries; Online Lexicography; Language contact; History of the Italian/French/English/German Lexicon

"Il faudrait encore, assurément, repérer, cartographier la diffusion de la langue italienne elle-même, cet élément insistant de toute culture européenne." (Braudel 1989: 15)

1 Introduction

This paper intends to give an overview of the structure and possible applications for linguistic research of the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT), published in print in 2008 by the Accademia della Crusca and currently being released online as part of a larger project (*Osservatorio degli italianismi nel mondo* / OIM). As a specialized lexicographic resource it focuses on language contact between Italian and three major European languages, French, English, German, whose lexical outcomes are presented in both synchronic and diachronic perspective.[1] The lexicographic metalanguage of the DIFIT is Italian, but the dictionary is in itself multilingual in so far as it registers lexical transfer from the donor language Italian to the three languages mentioned, the respective loan items following the lemmatization of their Italian source words (etyma).

Literary Italian (based on the Tuscan variety) along with many of its dialectal varieties stands out as a culturally extremely rich and diverse source for borrowings. While its extensive lexical influence on the other three European languages in different historic phases all along the past six centuries has been noted in a host of studies, the somewhat scattered lexicographical documentation is brought together for the first time in a reference work spanning more than one target language. By taking into account a variety of lexical fields, the DIFIT aims at reflecting the rich European heritage of borrowings which becomes evident in parallel loans and lexical interrelations within the French, English and German vocabularies.

A borrowing (It. *prestito*, G. *Entlehnung*) is seen here as the result of the imitation of a foreign linguistic pattern by a speech community (cf. Gusmani 1986, Pinnavaia 2001), meaning elements like the following:

(1) it. *ciao* (→ present in Fr., Eng., G.)
(2) it. *dolce far niente* (→ Eng., G.)
(3) it. –*issimo* suffix (→ e.g. Fr. *Affairissimo*)
(4) it. *eppur si muove!* → (Fr., Eng., in G. loan translation: *Und sie bewegt sich doch!*)
(DIFIT-OIM[2]: s.vv.)

The loan items present in the lexicon of a language can be single words like *ciao* (1), multiword expressions as (2) *dolce far niente*, or sometimes formatives as the Italianizing suffix with superlative meaning –*issimo* (cf. Fr. *affairissimo*, or G. *Transportissimo*, cf. Stammerjohann 2010)[3], as well as phrases (pro-

[1] For details of the macro- and microstructural makeup of the print dictionary cf. Heinz (2008).

[2] Data from the DIFIT corpus available in digitized form via the OIM are labelled here as DIFIT-OIM. These are searchable at www.italianismi.org.

[3] Cf. also examples in Austrian and Swiss German advertisements such as *Vielfaltissimo* (denominal derivation), *Perfektissimo* (deadjectival), *Verwöhnissimo* (deverbal) based on a product name consisting of a hybrid formation (*Caf*- ‘coffee’ + -*issimo*). Such productive formatives based on loan items are described as „induzione“ by Gusmani (1986: 155).

verbial and idiomatic expressions) like „eppur si muove!“ (4) in French and English, rendered in German by the loan translation *Und sie bewegt sich doch!*[4]

At present the DIFIT lists 8951 Italianisms and 4660 Italian etyma (242 of which are without prior attestation elsewhere). Following a concise presentation of the dictionary and its digital counterpart, some examples of how the dictionary data can be used for linguistic analyses will be outlined in this paper.

2 Language contact and lexicographical documentation: DIFIT and OIM

From the metalexicographical point of view informing the classification proposed by Wiegand (2001), the DIFIT can be described as representative of the type of the „aktives polylaterales Sprachkontaktwörterbuch“ (‚active, polylateral dictionary of language contact’); it is active insofar as its lemmatization sets out from the etyma and polylateral in its registering loanwords of one donor language in several recipient languages. Moreover, it is, as mentioned at the beginning, monolingual as to the metalanguage, Italian, while being multilingual as regards its set of recipient languages (French, English, German). The general scope of the dictionary is stated in the Introduction:

[I]l presente *Dizionario* [...] vuole offrire più che una semplice somma dei dati finora raccolti e acclarati. Il suo **scopo** è più specifico, è quello di **mettere a confronto l’incidenza dell’italiano sul francese, l’inglese e il tedesco**, le tre lingue che sono al centro dello spazio europeo e sono a più stretto contatto tra loro, con l’intento di **ricostruire le trafile di penetrazione e la diversa sorte delle parole italiane** in questo circuito. (DIFIT: XI; emphasis added)

In summary, the intention of this work vis-à-vis specialized repertories and studies centred on Italianisms present in single languages, on whose results the DIFIT draws, is to offer more than a simple listing of data yet collected and examined. Its intention is more specifically “mettere a confronto l’incidenza dell’italiano sul francese, l’inglese e il tedesco’, i.e. ‘compare the impact of Italian on French, English, and German’, three languages situated centrally in a European communicative space and in close contact among each other, in order to reconstruct the mechanisms of lexical interpenetration and the ‘variable fate of Italian words in this [historical and cultural] circuit’. By design it presupposes a certain depth of diachronic coverage both in the macrostructural preselection and in the microstructural organization of its entries.

[4] Examples drawn from DIFIT-OIM.

The DIFIT is currently being expanded within the OIM project. The *Osservatorio degli italianismi nel mondo*, an international collaboration hosted by the Accademia della Crusca, has two main purposes:

(a) making available and editing the data collected in the DIFIT in an online database;

(b) extending the contact languages beyond the three major European languages French, English and German documented in the DIFIT, aiming at the lexicographical description of the impact of Italian as a source language e.g. in Spanish, Polish, Japanese etc.

Hence the goal of this collaborative effort is a linguistic 'observatory' for Italian loanwords in the world's languages.

3 The contribution of a loanword dictionary to linguistic analyses

The OIM database[5] offers various search options. Besides a free search option ("ricerca libera") complex searches in the list of etyma ("ricerca negli etimi") are made possible by applying and combining the following criteria:

- Italian etyma (+ variants)
- grammatical categories (etymon)
- year of first attestation (etymon) with optional delimitation of time span
- archaic / obsolete / without lexicographical attestation
- domain of use / lexical field (etymon)
- dialectal origin (etymon)
- register
- recipient language (by year or range of years + one, two or all of the recipient languages)

The sources documenting the etyma can be displayed in a window with the full bibliographical reference by moving the cursor to the respective abbreviation. The "ricerca negli italianismi" (search Italianisms) option allows for complex search paths with the following set of criteria:

- Italianisms (+ variants)
- grammatical categories
- recipient language

[5] The database was designed and implemented as part of the electronic resources of the Accademia della Crusca by Marco Biffi with the help of Giovanni Salucci and Maurizio Rago, while the task of populating the database by digitizing and adapting the DIFIT data for the new online user inferface is the work of Gesine Seymer.

- year of first attestation with optional delimitation of time span
- archaic / obsolete / without lexicographical attestation
- domain of use / semantic field
- dialectal use
- register
- type of borrowing:
  - integral
  - partial (calque: formal / partial / semantic)
  - direct
  - indirect (with mediating language)
  - pseudo-borrowing[6]
  - borrowings whose (Italian) origin is doubtful or unclear (e.g. contradictory etymological information in authoritative sources)

Here too, sources for the Italianisms can be displayed with the full bibliographical reference by moving the cursor to the abbreviation.
In what follows, three exemplary fields of application are sketched out in order to show how the DIFIT data can be used for linguistic research, yielding some first results related to a number of research questions especially in contact linguistics.[7] These are namely the word formation types in the multiword lemmata identified in the DIFIT corpus, the typology of the borrowings and, as a further outlook, the visualization of the results of language contact in the OIMap project.

4 Language contact and word formation

Of the 4660 lemmata listed in the DIFIT, 4161 are single words, while 499 (more than 10%) are classified as *locuzioni*, multiword expressions including complex syntagmatic units or phrases as *con spirito* and lexicalized compounds like *pesce spada*. A semiautomatic count with manual classification of the formation types yields the following percentages:

[6] Also *pseudo-loan* (It. *pseudo-prestito*, G. *Scheinentlehnung* alongside other terms), cf. Winter-Froemel (2011: 44-45), „Bildungen wie dt. *Picobello* [...], die aus einer anderen Sprache (hier dem Italienischen [...]) direkt entlehnt zu sein scheinen, die sich aber in der vermeintlichen AS [Ausgangssprache] in dieser Form als nicht existent erweisen." (45)

[7] The relevance of studying loanwords for the enterprise of general linguistics and linguistic typology becomes evident when using resources like the World Loanword Database (WOLD, Haspelmath/Tadmor 2009), which analyzes lexical borrowability based on a restricted set of core vocabulary meanings (1460) in 41 (mainly non-Indo-European) recipient languages; Italian is present among the 369 donor languages, though only with a small number of loanwords (86 entries in the database, some having an identical source word).

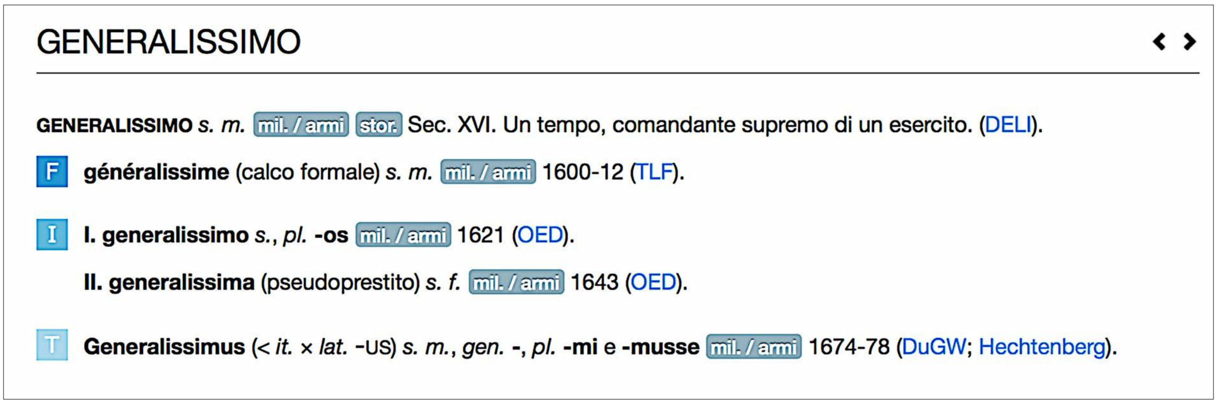
Formation	Examples	%
N + Adj	*salto mortale, sinfonia concertante, opera comica*	33,3
Prep + N	*a battuta, a conto, con spirito*	29,7
Adj + N	*dolce vita, sacra conversazione, onorata società*	15,8
N + Prep + N	*giorno di respiro, lira da gamba, zuppa di pesce*	13,5
N + N	*pesce spada, pensione baby*	4,0
[N + N]	*acquamarina, acquatinta, autostrada*	3,6

Table 1: Distribution and formation type of multiword expressions in DIFIT-OIM.

In a further step these data (as well as the overall database) can be analyzed with regard to the single recipient languages in order to compare how processes of loan integration may display a language specific dynamics.

5 Typology of borrowings

Categorizing and quantifying the diverse types existing alongside classical loanwords (such as formal or semantic calques, rare pseudo-borrowings together with hybrid formations etc.) gives a general picture of the impact of formal and semantic processes on the diachronic evolution and synchronic stratification of the lexicon in the recipient languages.



GENERALISSIMO ‹ ›

GENERALISSIMO *s. m.* mil. / armi stor. Sec. XVI. Un tempo, comandante supremo di un esercito. (DELI).

F **généralissime** (calco formale) *s. m.* mil. / armi 1600-12 (TLF).

I **I. generalissimo** *s.*, *pl.* **-os** mil. / armi 1621 (OED).

II. generalissima (pseudoprestito) *s. f.* mil. / armi 1643 (OED).

T **Generalissimus** (< *it.* × *lat.* -US) *s. m.*, *gen.* **-**, *pl.* **-mi** e **-musse** mil. / armi 1674-78 (DuGW; Hechtenberg).

Figure 1: Different loan outcomes in the recipient languages (DIFIT-OIM, s.v. 'generalissimo').

As a lemma like *generalissimo* (Fig. 1) shows, the loan formations resulting from even a single etymon may be characterized by a variety of outcomes. Here French has a formal calque, English a pseudo-borrowing based on an Italian structural model, whereas the German *Generalissimus* is identified as a hybrid combining an Italian base with a Latin(ate) ending.

Typology of borrowings	**%**
Direct borrowing (loanword)	94,5
Formal calque (loan translation)	3,3
Partial calque	0,8
Semantic calque	1,3
Pseudo-Italianism	0,1

Table 2: Distribution of various types of borrowings registered in DIFIT-OIM.

Table 2 gives the percentages of different types of borrowings. Clearly direct borrowings (loanwords in a strict sense) are by far the most common type among the Italianisms registered in DIFIT, while the most frequent type of calque is the word for word rendering of a foreign model with the means of the recipient language known as loan translation (e.g. It. *monte di pietà* → Fr. *mont-de-piété*), followed by partial calque (It. *fare fiasco* → G. *Fiasko machen*) and semantic calque (It. *futurista* → Eng. *futurist*). Pseudo-Italianisms like Eng. *generalissima* or Fr./Eng./G. *tutti-frutti* (with various meanings) are extremely rare (0,1%) in the DIFIT documentation.

Furthermore, thematic indices based on the different sectors of the lexicon that have contributed Italianisms can be created with the OIM search engine. Thanks to functions permitting combined search criteria it becomes then possible to list the results by chronology and/or target language.[8]

6 Visualizing the areal distribution of Italianisms: OIMap

As a further outlook, we will briefly describe an instrument for visualizing the areal distribution (and dynamics) of Italianisms by means of a mapping tool currently being developed under the name of OIMap.[9] In fact, while lists of borrowed lexical items in a dictionary or database provide multi-faceted information in a very dense manner, new insights into the geographic distribution and the dynamic tendencies of contact relationships between languages can be gained through maps.

Drawing upon a free software tool for digital cartography (http://batchgeo.com), OIMap intends to show the geographic direction of language contact. Moreover, the dialectal origin of borrowed elements can be obtained (other than manually) by using refined search options. Thereby dialect varieties having contributed a relatively high number of loanwords to the lexicon of one, two or all three of the extant recipient languages can be singled out, and as in the case of Venetian (or, with a comparatively lower number of lexical items, Genoese) a clearer picture of the areal dynamics emerges. In Fig.

[8] The ‚ups and downs' of Italian lexical influence in the three languages taken into account by DIFIT are illustrated by Stammerjohann/Seymer (2007: 51), with the availability of the OIM search engine the raw data for even more fine-grained analyses can be elicited as described in 3.3.

[9] OIMap is designed for the cartographic representation of language contact dynamics based on the DIFIT data (project situated at the University of Salzburg under the direction of M. Heinz).

3 an arrow indicates the main direction of borrowings where an absolute (Genoese: 7 out of 9) or relative (Venetian: 15 out of 45) majority of loan elements 'migrates' exclusively towards one language (French and German respectively).

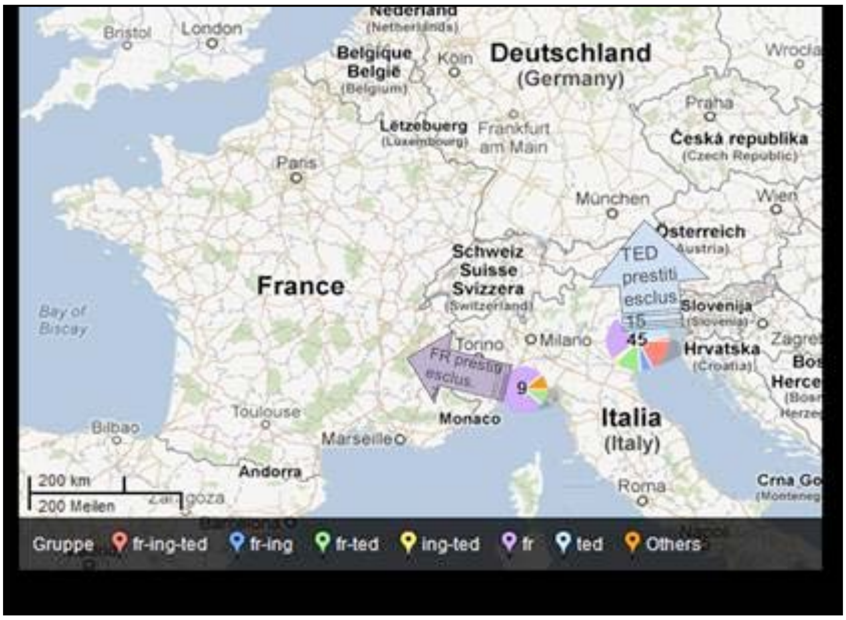


Figure 3: OIMap – Areal dynamics of Genoese and Venetian borrowings (FR = French, TED = German).

As the OIM database is developed further and the number of recipient languages grows,[10] an interface with the cartographic tool OIMap becomes possible, resulting in a world map[11] of lexical 'migration' setting out from Italian and its varieties.

7 References

Biffi, M. et al. (2014). *Osservatorio degli italianismi nel mondo*. Firenze: Accademia della Crusca (www.italianismi.org [10/04/2014]).

Braudel, F. (1989). *Le modèle italien*. Paris: Arthaud.

DIFIT = Stammerjohann, H., Arcaini, E. , Cartago, C., Galetto, P., Heinz, M. , Mayer, M., Rovere, G., Seymer, G. (2008). *Dizionario di italianismi in francese, inglese, tedesco*. Firenze: Accademia della Crusca.

Gusmani, R. (1986). *Saggi sull'interferenza linguistica. Seconda edizione accresciuta*. Firenze: Le Lettere.

Haspelmath, M., Tadmor, U. (eds.) 2009. *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology (http://wold.livingsources.org [10/04/2014]).

[10] For quite a number of languages electronic lists of Italianisms are at hand, though unpublished and in different formats, so integrating the existing material into the OIM database may soon result in substantial progress as for the documentation of Italian lexical influence.

[11] Such a map could provide more detailed cartographic and lexical information than the useful but rather plain one figuring on the WOLD website (Haspelmath/Tadmor 2009).

Heinz, M. (2008). L'expérience du Dizionario di italianismi in francese, inglese, tedesco (DIFIT): objectifs, structure et aspects méthodologiques. In F. Pierno (ed.) *Aspects lexicographiques du contact entre les langues dans l'espace roman*. Strasbourg: Université Marc Bloch, pp. 165-180.

Pinnavaia, L. (2001). *The Italian Borrowings in the* Oxford English Dictionary: *A lexicographical, linguistic and cultural analysis*. Roma: Bulzoni.

Stammerjohann, H., Seymer, G. (2007). L'italiano in Europa: italianismi in francese, inglese e tedesco. In N. Maraschio (ed.) *Firenze e la lingua italiana fra Nazione e Europa*. Firenze: Accademia della Crusca, pp. 41-55.

Stammerjohann, H. (2010). Italianismi. In *Enciclopedia dell'italiano*. Roma: Treccani (http://www.treccani.it/enciclopedia/italianismi_(Enciclopedia-dell'Italiano)).

Wiegand, H.-E. (2001) Sprachkontaktwörterbücher: Typen, Funktionen, Strukturen. In *Germanistische Linguistik* 161/162, pp. 115-224.

Winter-Froemel, E. (2011). *Entlehnung in der Kommunikation und im Sprachwandel*. Berlin/New York: de Gruyter.

http://batchgeo.com [10.04.2014]

The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian

Jelena Kallas, Maria Tuulik, Margit Langemets
Institute of the Estonian Language
jelena.kallas@eki.ee, maria.tuulik@eki.ee, margit.langemets@eki.ee

Abstract

This paper is a report on a lexicographical project completed by the Institute of the Estonian Language. The Basic Estonian Dictionary was published in print in March 2014, and the online version will be available by September 2014. The BED is aimed at learners of Estonian as a foreign language or as a second language at the elementary and intermediate levels (A2 to B1) according to the Common European Framework of Reference for Languages.

The dictionary contains about 5,000 headwords, including single items and multi-word lexical items. The BED provides lexicographical information on pronunciation, morphological information, definitions, word formation, government and collocation patterns, multi-word phrases, semantically related words and usage notes. In the online version, sound recordings (mp3 audio files) are provided. Morphological information was generated automatically. The most frequent government and collocation patterns were analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004).

The BED contains approx. 400 illustrations, study pages and picture pages (e.g. related to animals). In the appendix, geographical names and grammar tables are included.

In addition, the Dictionary of the Estonian Sign Language (containing approx. 6,700 video recordings), based on the BED database, was published online in March 2014.

Keywords: L2 monolingual lexicography; active dictionary; Estonian

1 Purpose and structure of the dictionary

The Basic Estonian Dictionary (henceforth BED) is a monolingual active dictionary aimed at learners of Estonian as a foreign language or as a second language at the elementary and intermediate levels. The dictionary contains about 5,000 headwords, which were chosen on the basis of their frequency in the Estonian Reference Corpus[1], with 250 million tokens as input. In addition, headwords that are necessary in everyday life, but might not be as frequent in corpora, were added, for example *pott* 'pot‘,

[1] http://www.cl.ut.ee/korpused/segakorpus/ [01/04/2014]

pann 'pan', *jahu* 'flour', and *köhima* 'cough'. To get systemic content, some semantic classes (e.g. animals, plants and professions) were specially analysed.

Headword list includes not only single items, but also multi-word lexical items. Multi-word lexical items presented independently are multi-word verbs – particle verbs (verb + adverb particle, e.g. *alla kukkuma* 'fall down') and expression verbs (verb + noun/adjective phrase, e.g. *aru saama* 'understand') – and multi-word interjections (e.g. *tere õhtust* 'good afternoon'). The headword list of the BED was considered to be the controlled vocabulary list of the whole dictionary, other words were used in the entries. This was intentional so that users can look up unknown words in the dictionary.

Morphological information was generated automatically by using a morphological synthesizer for Estonian[2]. The BED as a learner's dictionary uses a comprehensive form-based presentation of data. For declinable words, grammatical cases in singular and plural, as well as the short form of the Illative, are presented explicitly. For verbs, the *-ma* and *-da* infinitives, and *he/she* forms, the past participle forms are given. However, after automatic generation there was a need for manual control of generated forms. Mostly this was necessary for the identification of homonymy. But forms were also deleted and added according to their frequency in corpora.

Information on pronunciation (palatalization, stress and syllabic quantity (in Estonian, a tripartite correlation of three syllabic quantities of stressed syllables exists)) is presented on the level of basic morphological forms of headwords. This is done by means of special palatalization, quantity and stress marks. Stress is shown only in cases where it is not on the first syllable, the normal stress pattern in Estonian.

Information on word formation is built into the micro-structure. Compounds with the headword as a second element (base word) are presented as references/links to their own entries, without additional information in the entry of the base word. Only transparent compounds, where the meaning of the base word has been preserved, were selected. All referenced compounds are presented as independent headwords as well.

Semantically related words (synonyms, antonyms and paronyms) of headwords are shown using the simplest possible metalanguage, e.g. *sama mis* 'same as' for synonyms and *vastand* 'opposite' for antonyms.

At the end of some entries, there are usage notes. Usage notes show differences between words and help to build vocabulary, e.g. polite phrases related to particular headwords are given and usages prone to error are pointed out.

The XML database of the Basic Estonian Dictionary is organized into several fields: lemma, pronunciation, inflectional information, definition, word formation, government, collocation, multi-word patterns, semantically related words and usage notes.

[2] http://www.eki.ee/keeletehnoloogia/projektid/morfana/ [01/04/2014]

1.1 Government and collocation patterns in the BED

As the BED is an active dictionary, the explicit presentation of syntagmatic relations (government and collocational patterns, also multi-word phrases) are of the utmost importance.

The most frequent government and collocation patterns were analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004).

Estonian Sketch Grammar (Kallas 2013) is geared towards the specification of the Estonian Reference Corpus and it contains 85 rules (14 UNARY, four SYMMETRIC, 62 DUAL and five TRINARY grammatical relations). As a result, the system searches for 32 types of lexicogrammatical constructions.

For nouns, the system searches for modifying adjectives, participles, oblique-case substantives, adverbs, pronouns, prepositional phrases, non-finite verbs and (by identifying conjunctive words) subordinate clauses.

For adjectives, the system searches for modifying adjectives, adverbs, oblique-case substantives, prepositional phrases, non-finite verbs and (by identifying conjunctive words) subordinate clauses.

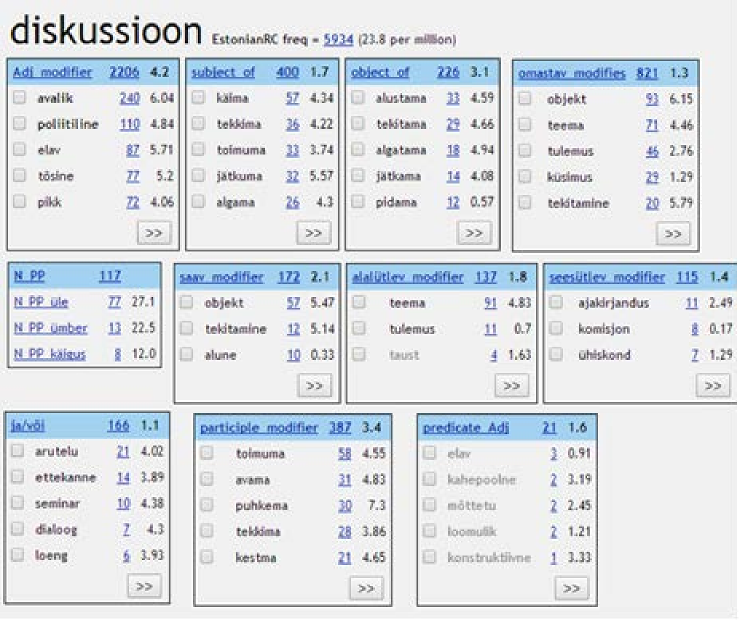
For adverbs, the system searches for modifying adverbs, oblique-case substantives, prepositional phrases and (by identifying conjunctive words) subordinate clauses.

For verbs, the system searches for substantives that function as subjects, objects and adverbials, and also for modifying adjectives, adverbs, prepositional phrases, non-finite verbs, gerundives and (by identifying conjunctive words) subordinate clauses.

Multi-word verbs, i.e. particle verbs (verb + adverb particle, e.g. *alla kukkuma* 'fall down‘), expression verbs (verb + noun/adjective phrase, e.g. *aru saama* 'understand‘), catenative verbs (verb + non-finite verb, e.g. *käima panema* 'start‘, lit. 'make [the engine] work‘), and support verb constructions (e.g. *läbirääkimisi pidama* 'negotiate‘) are considered separately. Since adverbial particles are tagged in the corpus as regular adverbs, a list of adverbial particles was compiled. The system identifies the most frequent adverbial particles used with particular verbs. This feature has great value when lexicographers need to choose what kind of particle verbs should be presented in the dictionary. Secondly, it is possible to see components of expression verbs if the component concerned has the part-of-speech tag X. Other components of multi-word verbs are identified as objects, adverbials or modifying non-finite verbs.

In addition, constructions with the conjunctions *ja/või* ‘and/or’, and *kui/nagu* ‘as’ can be found for all content words. For nouns, the system also searches for predicatives (complements of the copula-like verb *olema* ‘be’).

Figure 1 shows the word sketch for the noun diskussioon 'discussion'.



diskussioon EstonianRC freq = 5934 (23.8 per million)

Adj_modifier	2206	4.2
□ avalik	240	6.04
□ poliitiline	110	4.84
□ elav	87	5.71
□ tõsine	77	5.2
□ pikk	72	4.06

subject_of	400	1.7
□ käima	57	4.34
□ tekkima	36	4.22
□ toimuma	33	3.74
□ jätkuma	32	5.57
□ algama	26	4.3

object_of	226	3.1
□ alustama	33	4.59
□ tekitama	29	4.66
□ algatama	18	4.94
□ jätkama	14	4.08
□ pidama	12	0.57

omastav_modifies	821	1.3
□ objekt	93	6.15
□ teema	71	4.46
□ tulemus	46	2.76
□ küsimus	29	1.29
□ tekitamine	20	5.79

N_PP	117	
N_PP_üle	77	27.1
N_PP_ümber	13	22.5
N_PP_käigus	8	12.0

saav_modifier	172	2.1
□ objekt	57	5.47
□ tekitamine	12	5.14
□ alune	10	0.33

alaütlev_modifier	137	1.8
□ teema	91	4.83
□ tulemus	11	0.7
□ taust	4	1.63

seesütlev_modifier	115	1.4
□ ajakirjandus	11	2.49
□ komisjon	8	0.17
□ ühiskond	7	1.29

ja/või	166	1.1
□ arutelu	21	4.02
□ ettekanne	14	3.89
□ seminar	10	4.38
□ dialoog	7	4.3
□ loeng	6	3.93

participle_modifier	387	3.4
□ toimuma	58	4.55
□ avama	31	4.83
□ puhkema	30	7.3
□ tekkima	28	3.86
□ kestma	21	4.65

predicate_Adj	21	1.6
□ elav	3	0.91
□ kahepoolne	2	3.19
□ mõttetu	2	2.45
□ loomulik	2	1.21
□ konstruktiivne	1	3.33

Figure 1: Word sketch of the noun *diskussioon* 'discussion'.

The word sketch offers the lexicographer the most frequent collocates that occur as adjectival modifiers (e.g. *avalik* 'public', *poliitiline* 'political', *elav* 'lively', *tõsine* 'serious', *pikk* 'long' and *avatud* 'open'), various oblique-case substantive modifiers (e.g. *diskussiooni objekt/teema/tulemus* 'object/topic/result of discussion') and in the 'and/or' relation to the node word (e.g. *diskussioon ja arutelu* 'discussion and debate').

Also identified are relations where the node word functions as subject (e.g. *diskussioon käib/tekib/jätkub* 'discussion takes place/starts/continues') and object (e.g. *diskussiooni alustama/algatama/jätkama/avama* 'start/initiate/continue/open a discussion').

The most frequent extracted patterns are mostly included in the entry of particular words and registered in the dictionary database.

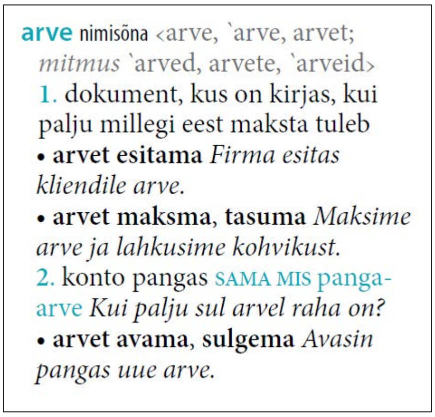
In the BED database, the government pattern field contains data about the government pattern, together with attributes for the type of government (object, case, adposition, infinitive and conjunctive word government), as well as the position of the complements and complementation variability.

The collocation pattern field contains data about the collocation pattern, together with attributes for the type of collocation. Collocation patterns are described by means of categorical and functional-relational labels. There are 13 types of collocation in the BED database:

- N(S)+V noun (as grammatical subject) + verb: *päike paistab/tõuseb/loojub* 'sun shines/rises/sets';
- N(O)+V noun (as grammatical object) + verb : *arvutit sisse lülitama* 'switch on the computer';
- N(A)+V noun (as adverbial modifier) + verb: *kinnisvarasse investeerima* 'invest in property';
- Adj+V adjective + verb: *määravaks osutuma* 'prove decisive';

- Adv+V adverb + verb: *kiiresti jooksma* ‘run fast’;
- N+N noun + noun: *ekspertide hinnang/arvamus* ‘assessment/opinion of experts’;
- Adj+N adjective + noun: *hea/halb eeskuju* ‘good/bad example’;
- Num+N numeral + noun: *sada meetrit/kilo* ‘hundred meters/kilograms’;
- Adv+N adverb + noun: *kergesti süttiv* ‘easily flammable’;
- Adv+Adv adverb + adverb: *väga aeglaselt* ‘very slowly’;
- Prep+N preposition + noun: *enne/pärast jõule* ‘before/after Christmas’;
- N+Post noun + postposition: *interneti/raadio kaudu* ‘on television/ radio’.

Collocations of the same type are divided into semantic sets and presented explicitly as separate bundles. Figure 2 shows the entry for *arve* ‘invoice, account’ in the printed version of the BED.



arve nimisõna ‹arve, `arve, arvet; *mitmus* `arved, arvete, `arveid›
1. dokument, kus on kirjas, kui palju millegi eest maksta tuleb
• **arvet esitama** *Firma esitas kliendile arve.*
• **arvet maksma, tasuma** *Maksime arve ja lahkusime kohvikust.*
2. konto pangas SAMA MIS panga-arve *Kui palju sul arvel raha on?*
• **arvet avama, sulgema** *Avasin pangas uue arve.*

Figure 2: The BED entry for the noun *arve* ‘invoice, account’ in the printed version.

1.2 Extra materials

The BED also contains approx. 400 illustrations. These are single illustrations with legends, structural illustrations (particular objects are highlighted by means of arrows), functional illustrations (mostly for adpositions), scenic illustrations (mostly for phrasal verbs) and nomenclatory illustrations (see figure 3).

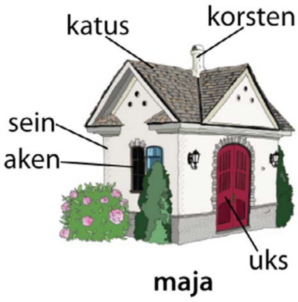
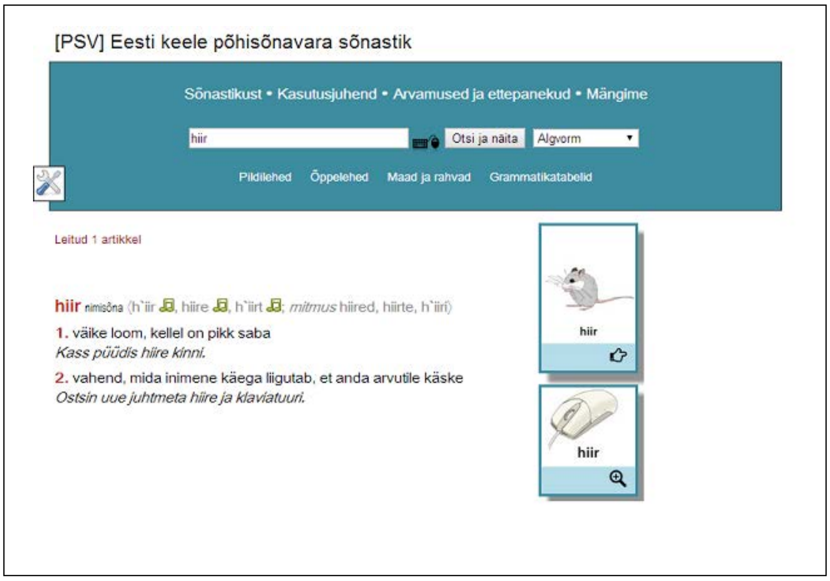


Figure 3: Nomenclatory illustration for the entry *maja* ‘house’.

Besides illustrations, the dictionary has a centre section of 16 study pages (including instructions for producing numbers, time and dates, writing letters and emails, punctuation marks, common abbreviations, useful phrases) and 17 picture pages (e.g. on insects, animals, flowers and transportation). In the appendix, a list of countries, people and languages is given, as well as grammatical tables. Grammatical tables show how to decline and conjugate words, also they give guidance for producing all other word forms when moving on from the basic forms given in the dictionary.

2 The BED as an online-dictionary

The online version of the BED has some innovative features, which are implemented in the Estonian lexicography for the first time. Figure 4 illustrates the interface of the online version of BED.


[PSV] Eesti keele põhisõnavara sõnastik
Sõnastikust • Kasutusjuhend • Arvamused ja ettepanekud • Mängime
hiir
Otsi ja näita
Algvorm
Pildilehed
Õppelehed
Maad ja rahvad
Grammatikatabelid
hiir
hiir

Leitud 1 artikkel

hiir nimisõna (h`iir, hiire, h`iirt; *mitmus* hiired, hiirte, h`iiri)

1. väike loom, kellel on pikk saba
Kass püüdis hiire kinni.

2. vahend, mida inimene käega liigutab, et anda arvutile käske
Ostsin uue juhtmeta hiire ja klaviatuuri.

Figure 4: Online BED entry for the noun *hiir* 'mouse'.

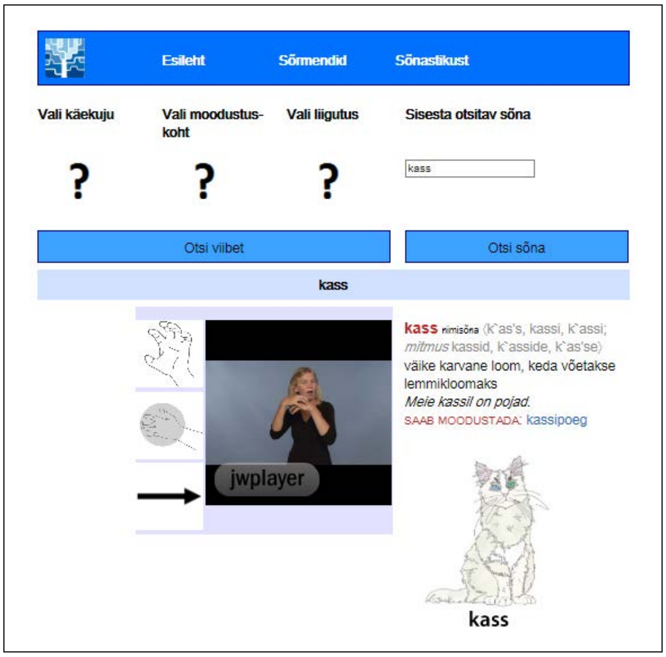
Pictures are aligned with particular word meanings. If the picture is topically related to one of the special picture pages provided in the dictionary as extra material (e.g. *hiir* 1. 'mouse' as related to *animals*), then these pictures are linked together. Otherwise it is possible to enlarge the picture.

Green musical note symbols indicate that there are sound recordings (mp3 audio files) linked to particular morphological forms. The audio files were pre-recorded.

The contents of the entire dictionary have been morphologically analysed. As a result, users can click on any word in a definition or example to find the entry for that word. And, vice versa, it is possible to type into the search box a word in any form (previous dictionaries allowed for searching only on the basis of lemma), and the lemma entry will be provided.

3 The BED as a basis for the Dictionary of the Estonian Sign Language

The BED database was used to compile the online Estonian Sign Language – Estonian Language dictionary[3]. Figure 5 illustrates the interface of the dictionary. There are approx. 6,700 video files. For every sign, the BED database contains information on the initial hand form, the location where the sign is articulated (face, lips, cheek, chest, neutral space etc.) and the movement with which the sign is formed. Based on these three parameters, it is possible – for the first time in Estonian lexicography – to search for a certain sign by choosing the hand form, the location, or the movement of the sign. This enables the deaf dictionary user to find the Estonian equivalent for a sign. The interface allows for searching in the opposite direction as well, making it possible for the non-deaf to learn Estonian Sign Language.



kass nimisõna (k`as's, kassi, k`assi; *mitmus* kassid, k`asside, k`as'se)
väike karvane loom, keda võetakse lemmikloomaks
Meie kassil on pojad.
SAAB MOODUSTADA: kassipoeg

Figure 5: Online entry for the noun *kass* ‘cat’ in the Estonian Sign Language – Estonian Language online-dictionary.

4 The BED as a lexical resource in the Dictionary Writing System EELex

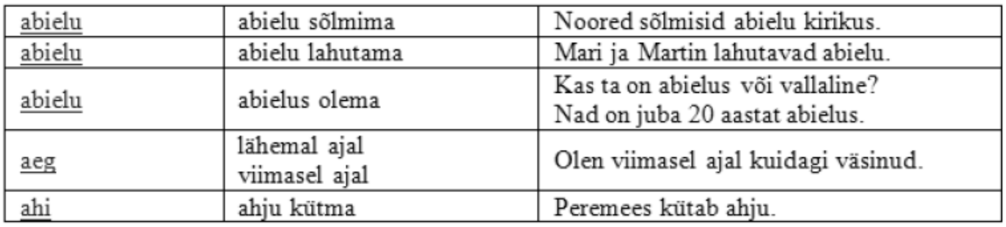
The BED was compiled in the web-based dictionary writing system EELex[4] (Jürviste et al. 2011). Nearly 50 dictionaries of different types (monolingual and bilingual, general and learner’s dictionaries, etc.)

[3] http://www.eki.ee/dict/viipekeel/ [01/04/2014]
[4] http://eelex.eki.ee/ [01/04/2014]

with a standard XML mark-up make EELex a multi-purpose lexicographical database. XML-based compilation allows for the generation of different outputs: for example, specialised dictionaries based on partial database output (Kallas, Langemets 2012). There are two options for the automatic generation of specialised dictionaries: reorganising the preview (and layout) of the existing dictionary articles, or generating a new dictionary database (i.e. to clone only a part of the source database).

The function of the article preview generator makes it possible to modify the preview, i.e. to set a character, text or line break between, in front of or after a specific element or group of elements, to show or hide specific elements in the article editing preview, to assign a condition for displaying a specific element (according to the value of the attribute or neighbouring elements) or to assign a hyperlink to an element. So, by specifying elements in the print preview, it is possible to get output consisting of only those elements that are specified by the user.

The same result may be achieved by the customization of the regular XML query. It is possible to select particular elements to be displayed instead of the whole content of the dictionary article. Table 1 shows a dictionary-like extract from the BED database consisting only of the following elements: lemma, collocational patterns and usage example.



abielu	abielu sõlmima	Noored sõlmisid abielu kirikus.
abielu	abielu lahutama	Mari ja Martin lahutavad abielu.
abielu	abielus olema	Kas ta on abielus või vallaline? Nad on juba 20 aastat abielus.
aeg	lähemal ajal viimasel ajal	Olen viimasel ajal kuidagi väsinud.
ahi	ahju kütma	Peremees kütab ahju.

Table 1: An example of the collocations extracted from the BED database.

In this way, it is possible to reuse the BED database in order to generate specialised dictionaries (e.g. a dictionary of government and collocations).

5 Conclusion

The XML-based compilation makes the BED database a useful lexical resource, which can be used in different ways for development materials meant for the teaching and learning of the Estonian language as a second or a foreign language. The dictionary is special in many ways. It is the first monolingual dictionary meant for learners of Estonian at the elementary and intermediate levels (previously there were bilingual dictionaries). Government and collocation patterns were analysed and selected using the Sketch Engine corpus query system. The online version allows learners to listen to the pronunciation of words. In addition, the morphological analysis implemented in the BED make it a very innovative and user-friendly dictionary.

The first online Estonian Sign Language – Estonian Language dictionary has also been compiled. This dictionary is unique in that it enables the deaf dictionary user to find the Estonian equivalent for a sign, and not only for a word.

In future, it may be possible to convert the dictionary web page into a language-learning portal by combining the dictionary with other resources (corpora, different specialised dictionaries etc.).

6 References

Jürviste, M., Kallas, J., Langemets, M., Tuulik, M., Viks, Ü. (2011). Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In I. Kozem, K. Kozem (eds.) eLexicography in the 21st Century: New Applications for New Users, Proceedings of eLex 2011, Bled, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovenian Studies, pp. 106-112.

Kallas, J., Langemets, M. (2012). Automatic Generation of Specialized Dictionaries Using the Dictionary Writing System EELex. In A. Tavast, K. Muischnek, M. Koit (eds.) Human Language Technologies – The Baltic Perspective, Proceedings of the Fifth International Conference Baltic HLT 2012. IOS Press, (Frontiers in Artificial Intelligence and Applications), pp. 103-110.

Kallas, J. (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. PhD thesis. Tallinn: Tallinna Ülikool.

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.

Acknowledgements

The project was supported by the National Programme for Estonian Language and Cultural Memory (2009–2013), as well as the National Programme for Estonian Language Technology (2011–2017).

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

Meike Meliss
Universidad de Santiago de Compostela-Spanien
meike.meliss@usc.es

Abstract

Der Beitrag beschäftgt sich mit den verschiedenen Such-, Aufffindungs- und Auswahlsprozessen, die für die fremdsprachige Produktion nowendig sind und von DICONALE-*online*, einem onomasiologisch-konzeptuell ausgerichteten, zweisprachig-bilateral konzipierten Verbwörterbuch der spanischen und deutschen Gegenwartsspache, besonders berüksichtigt werden. Der Ausgangspunkt von DICONALE ist ein unbefriedigendes Informationsangebot in den bestehenden ein- und zweisprachigen Lernerwörterbüchern für den L2-*output* und bestätigt das Projektteam in der Notwendigkeit, ein neuartiges benutzer- und situationsdefiniertes *onlin*e-Nachschlagewerk zu erstellen. Zwei Bezugsrahmen bilden die Grundlage für einen komplexen, konzeptuell und framegeleiteten Zugriffspfad, der dem Benutzer bei der Suche und Auswahl von Ausdrucksmöglichkeiten und der adäquaten Anwendung behilflich sein soll. Das Novum dieses Wörterbuchprojekts besteht hauptsächlich darin, eine onomasiologisch-konzeptuelle Perspektive für den fremdsprachigen Produktionsprozess nutzbar zu machen und mit einem semasiologischen Zugriff zu verbinden, durch den es möglich ist, die inter- und intralingualen Unterschiede zwischen den Lexemen eines lexikalisch-semantischen (Sub)Paradigmas hervorzuheben.

Ziel des Beitrages ist es daher, den Ausgangspunkt, sowie die theoretischen und methodologischen Grundlagen von DICONALE-*online* unter der speziellen Perspektive der Benutzer- und Situationsorientiertheit zur Diskussion zu stellen, die einzelnen Zugriffspfade für den Such- und Auffindungsprozess vorzustellen und das Angebot zur Auswahl und zum adäquaten Gebrauch aus inter- und intralingualer Perspektive zu präsentieren.

Keywords: Lernerlexikographie; Argumentstrukturgrammatik; kontrastive Linguistik

1 Einleitung

DICONALE-*online* ist ein Forschungsprojekt[1] zur Erstellung eines onomasiologisch-konzeptuell orientierten, zweisprachig-bilateralen Wörterbuches deutscher und spanischer Verben[2] im Kontrast, welches sich an fortgeschrittene Lernende des Deutschen bzw. Spanischen als Fremdsprache (DaF und Ele) richtet und besonders die fremdsprachige freie Produktion anvisiert. Es handelt sich somit um ein Wörterbuchprojekt, welches sowohl den Benutzerkreis als auch die Benutzersituation klar vordefiniert und damit sogleich deutlich einschränkt. Die Struktur des zukünftigen Wörterbuches, sowie die Auswahl des Informationsangebotes und der Zugang stehen in direkter Verbindung mit der vordefinierten Benutzersituation.

Trotz zahlreicher zweisprachiger Wörterbücher für das besagte Sprachenpaar in *prin*t- und *online*-Format besteht nach unserer Auffassung die Notwendigkeit, ein neuartiges, konzeptuell orientiertes *online*-Verbwörterbuch zu konzipieren, um damit gerade die Aspekte in den Mittelpunkt zu stellen, die in der herkömmlichen ein- und zweisprachigen Lernerlexikographie für das Sprachenpaar Deutsch-Spanisch bis jetzt zu kurz gekommen sind. Diese sind u.a. die konsequente Berücksichtigung des situationsbedingten Such- und Auffindungsprozesses eines geeigneten Lexems für den fremdsprachigen *output*, sowie eine benutzerorientierte Darbietung der Information für den Auswahlprozess und den anschließenden situationsgerechten Gebrauch in der jeweiligen Fremdsprache**.** Die sich daraus ergebenen notwendigen Suchsequenzen rechtfertigen eine primär onomasiologisch-konzeptuell orientierte Zugriffsstruktur. Diese Prämissen stellen uns vor die Herausforderung, einen Benutzerkreis, der prinzipiell an alphabetisch-semasiologisch konzipierte Wörterbücher gewöhnt ist, durch ein geeignetes Suchleitsystem zu den gewünschten Resultaten zu führen. Das Novum dieses Wörterbuchprojekts besteht daher hauptsächlich darin, eine onomasiologisch-konzeptuelle Perspektive für den fremdsprachigen Produktionsprozess nutzbar zu machen und mit einem semasiologischen Zugriff zu verbinden, durch den es möglich ist, die inter- und intralingualen Unterschiede zwischen den Lexemen eines lexikalisch-semantischen (Sub)Paradigmas hervorzuheben.

Ziel des Beitrages ist es daher, den Ausgangspunkt, sowie die theoretischen und methodologischen Grundlagen von DICONALE-*online* unter der speziellen Perspektive der Benutzer- und Situationsorientiertheit zu präsentieren (Kapitel 2). In Kapitel 3 sollen im Einzelnen die Zugriffspfade für den Such- und Auffindungsprozess vorgestellt, in Kapitel 4 das Angebot zur Auswahl und zum Gebrauch präsentiert und abschließend in Kapitel 5 ein kurzer Ausblick angeboten werden. Zur Veranschauli-

[1] DICONALE (= Diccionario conceptual del alemán y del español): Es handelt sich um ein von dem spanischen Ministerium gefördertes Forschungsprojekt (MINECO – FEDER: FFI2012-32658: 2013-2015), das außerdem in Verbindung mit dem lexikographischen Netzwerk RELEX (Xunta de Galicia: CN2012/290) entwickelt wird. Das Forschungsteam besteht aus Mitgliedern verschiedener spanischer, deutscher und portugiesischer Universitäten und Forschungseinrichtungen und wird von der Autorin dieses Beitrages an der Universidad de Santiago de Compostela (Spanien) geleitet.

[2] Neben der Hauptlemmaliste werden in einer sekundären Lemmaliste auch deverbale Nomen, Adjektive und Adverbien und komplexe, mehrteilige Lexeme aufgenommen.

chung werden einige ausgewählte Beispiele aus dem Bereich der Verben der AUDITION herangezogen.

2 Ausgangspunkt und theoretisch-methodologische Grundlagen

Als Ausgangspunkt unserer Überlegungen stützen wir uns auf Untersuchungen, die aufzeigen konnten, dass die lexikographischen Ressourcen, die normalerweise im DaF- und Ele-Bereich zur Verfügung stehen, bis jetzt zu wenig Wert auf fremdsprachige Produktionssituationen für die fortgeschrittene Lernerebene (ab B2) gelegt haben. Dies gilt gleichermaßen für die zweisprachigen, als auch für die einsprachigen Lernerwörterbücher des besagten Sprachenpaars (Meliss 2013a, 2013b, 2014a, 2014b, 2014c).

Nach der Untersuchung der gängigsten zweisprachigen Großwörterbücher Spanisch-Deutsch (GWB-sp/dt)[3] in *print* und *online*-Format lässt sich zusammenfassen, dass in Verbindung mit der hier im Fokus stehenden fremdsprachigen Produktionssituation dem Benutzer zu wenig Information zu dem syntagmatischen Kombinationspotenzial der möglichen Entsprechungen in der fremdsprachigen Zielsprache angeboten und die semantisch orientierte Disambiguierung bei Entsprechungsvielfalt durch Angabe von paradigmatischen Sinnrelationen zu wenig für einen angebrachten Gebrauch genutzt werden (Fuentes Morán 1997, Haensch&Omeñaca ^{2}2004, Hausmann 1991, Meliss 2013a, 2013b, 2014a, 2014b, Model 2010).[4] Der Such- und Auswahlprozess zur geeigneten Bennenung wird außerdem von der Muttersprache geleitet und führt daher in vielen Fällen nicht zu der zielsprachigen Ausdrucksvarietät (Meliss 2014c).

Einsprachige Lernerwörterbücher (LWB) für DaF[5] und Ele[6] weisen zwar in den meisten Fällen ein relativ hohes Informationsangebot bezüglich des Kombinationspotenzials der einzelnen Lexeme durch Angabe von Strukturformeln auf (Dentschewa 2006, Engelberg 2010; Meliss 2013b, 2014b) und bieten dem Benutzer somit nützliche Information zum korrekten Gebrauch.[7] Der klassische, alphabetisch orientierte Zugang und die semasiologische Zugriffsperspektive favorisieren hingegen nicht die Auffindung eines unbekannten Lexems zur Benennung eines bestimmten Konzepts im fremdsprachigen

[3] LHWB und LHWBe, LEO, Pons: Das Sprachenportal, SGIWBe (Slaby/Grossmann/Illig);

[4] Dies steht im Einklang mit Untersuchungen zu zweisprachigen WB anderer Sprachen (Engelberg/Lemnitzer 42009, Herbst/Klotz 2003). Siehe dazu auch: Abel 2008.

[5] LGWB-DaF (Götz et al.) & online-Version, GWB-DaF (Kemcke), PGWB-DaF und Pons-DaF-online, Duden-DaF, Wahrig-DaF und online-Version;

[6] DS (Diccionario Salamanca), DA (Diccionario Alcalá) und online-Version, SM-Clave-online;

[7] Dieses Informationsangebot ist besonders ausgeprägt in den DaF-Lernerwörterbüchern von Langenscheidt (LGWB-DaF: Götz et al. 32010) und Kempcke (1999). So ist z.B. die mikrostrukturelle Information zu den Verballemmata von Langenscheidt DaF von einer syntagmatisch orientierten Grundstrukturierung geprägt (Engelberg 2010: 116).

output-Prozess und zieht auch keine spezifische Hilfestellung zur Auswahl aus einer Vielfalt von bedeutungsähnlichen Ausdrucksmöglichkeiten in Betracht. [8]

Verschiedene andere lexikographische Ressourcen wie z.B. syntagmatische und paradigmatische Spezialwörterbücher[9] können zwar die genannten Informationslücken und Zugriffsblokaden für die fremdsprachige *output*-Situation teilweise beheben, stehen aber im DaF- und Ele-Bereich entweder nicht zur Verfügung, oder genießen im Falle von frei verfügbaren *online*-Ressourcen, zu denen der Benutzer durch externe Links der gängigen ein- und zweisprachigen Wörterbuchportale zwar fast automatisch gelangt[10], nicht den erwünschten Bekanntheitsgrad (Meliss 2013b, 2014b). Die begrenzten Sprachkenntnisse und die mangelhafte lexikographische Vorbildung des hier anvisierten prototyischen Benutzers führt außerdem zu einer wenig optimierten Nutzung des inzwischen sehr breiten Informationsangebots bei gleichzeitiger Gefahr des „Sich Verirrens" („lost in hyperspace": Storrer 2010).

Das Forschungsprojekt DICONALE hat sich daher zum Ziel gesetzt, die unterschiedlichen Such-, Auffindungs- und Auswahlprozesse, die in der freien L2-Sprachproduktion durchlaufen werden müssen, konsequent zu berücksichtigen. Im Mittelpunkt der Überlegungen stehen daher folgende Problemkomplexe: (i) die Ausdrucks- bzw. Benennungssuche, (ii) die Ausdrucksauswahl aus der Vielfalt und (iii) der Gebrauch unter Berücksichtigung kontrastiv relevanter Divergenzen. Eine sich daraus ableitende onomasiologisch-konzeptuelle Zugriffsstruktur und ein modulares Informationsangebot bilden die Grundlagen für die Makro- und Mikrostruktur von DICONALE.

Dementsprechend setzt sich die MAKROSTRUKTUR in einer ersten Arbeitsphase aus 10 konzeptuellen Feldern[11] zusammen, die in weitere konzeptuelle Subfelder „zweiten und dritten Grades" mittels einer immer feiner differenzierenden Konzeptualisierung gegliedert werden. Diese Subfelder bilden die Grundlage für die lexikalisch-semantischen (Sub)Paradigmen (SPls), denen durch die auf dieser Stufe erfolgte Lesartdifferenzierung einzelne Lexeme in beiden Sprachen zugeordnet werden können. Das mehrstufige Beschreibungsmodell basiert auf unterschiedlichen lexikologischen Parametern die in 4 Modulen erfasst werden (Meliss & Sánchez Hernández 2014, González Ribao & Meliss 2014). Zur Lesartdisambiguierung wird besonderer Wert auf die Beschreibung der Bedeutung und bestimmter kombinatorischer Merkmale gelegt. Die Argumentstrukturbeschreibung zusammen mit entsprechender Information zu den morpho-syntaktischen, funktionalen und semantisch-kategoriellen Füllungen und Kollokatoren[12] stehen hier – neben den paradigmatischen Sinnrelationen - im Mittel-

[8] Ausnahmen sind für das Deutsche der Teil 2 des Wörterbuches von Kempcke (1999) und für das Spanische der onomasiologisch-konzeptuell angelegte Teil 2 des zukünftigen „Diccionario de Coruña" (Porto Dapena et al. 2008).

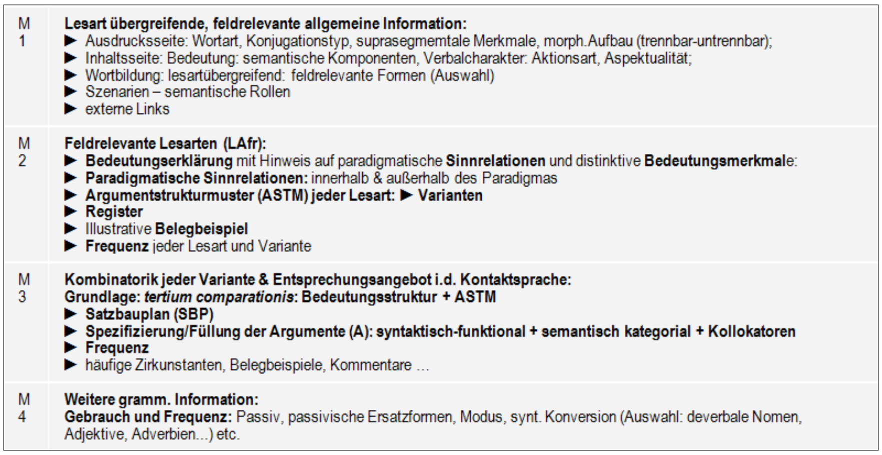
[9] Insbesondere sind hier die Konstruktionswörterbücher (ValenzWB, KollokationsWB etc.) und die SynonymWB zu nennen.

[10] So gelangt der Benutzer über das Pons-Sprachenportal zu einsprachigen Wörterbüchern der deutschen Sprache, wie z.B. DWDS und zu einigen Spezialwörterbüchern. Über CanooNet gelangt der etwas geschulte Benutzer über TheFreeDictionary zu der online-Version von Langenscheidts DaF-WB (LGWB-DaF-online). Allerdings muss festgehalten werden, dass kaum mit syntagmatisch-orientierten Ressourcen verlinkt wird.

[11] Es werden z.Z. Felder der Wahrnehmung, Kommunikation, Zwischenmenschlichen Beziehung, Kognition, Transfer, Konsum, Fortbewegung und der Existenz untersucht.

[12] Siehe dazu u.a.: Engelberg 2014a, Engelberg 2014b, Engelberg et al. 2012.

punkt der Module 2 und 3 (*Abbildung 1).*[13] Zu den empirischen Grundlagen in Verbindung mit der problematischen Erstellung von vergleichbaren Korpora für beide Sprachen soll auf die Studie von González Ribao (2014) verwiesen werden. Bezüglich des Formats haben uns jüngste Studien zur Benutzerforschung im ein- und zweisprachigen Kontext (Domínguez et al. 2013, Klosa et al. 2011)[14] in der Notwendigkeit bestätigt, ein lexikographisches Werk zu schaffen, welches über einen freien Internetzugang, ein schnell zugängliches, modular organisiertes, benutzerfreundliches und -adaptives, intern und extern verlinktes Informationsangebot offeriert[15], welches unserem DaF- und Ele-Benutzerkreis auch sprachlich und metasprachlich entgegen kommt.



M 1	**Lesart übergreifende, feldrelevante allgemeine Information:** ▶ Ausdrucksseite: Wortart, Konjugationstyp, suprasegmemtale Merkmale, morph.Aufbau (trennbar-untrennbar); ▶ Inhaltsseite: Bedeutung: semantische Komponenten, Verbalcharakter: Aktionsart, Aspektualität; ▶ Wortbildung: lesartübergreifend: feldrelevante Formen (Auswahl) ▶ Szenarien – semantische Rollen ▶ externe Links
M 2	**Feldrelevante Lesarten (LAfr):** ▶ **Bedeutungserklärung** mit Hinweis auf paradigmatische **Sinnrelationen** und distinktive **Bedeutungsmerkmal**e: ▶ **Paradigmatische Sinnrelationen:** innerhalb & außerhalb des Paradigmas ▶ **Argumentstrukturmuster (ASTM) jeder Lesart: ▶ Varianten** ▶ **Register** ▶ Illustrative **Belegbeispiel** ▶ **Frequenz** jeder Lesart und Variante
M 3	**Kombinatorik jeder Variante & Entsprechungsangebot i.d. Kontaktsprache:** **Grundlage: *tertium comparationis*: Bedeutungsstruktur + ASTM** ▶ **Satzbauplan (SBP)** ▶ **Spezifizierung/Füllung der Argumente (A): syntaktisch-funktional + semantisch kategorial + Kollokatoren** ▶ **Frequenz** ▶ häufige Zirkunstanten, Belegbeispiele, Kommentare …
M 4	**Weitere gramm. Information:** **Gebrauch und Frequenz:** Passiv, passivische Ersatzformen, Modus, synt. Konversion (Auswahl: deverbale Nomen, Adjektive, Adverbien...) etc.

Abbildung 1: Die 4 Beschreibungsmodule von DICONALE.

3 Suchen und Finden

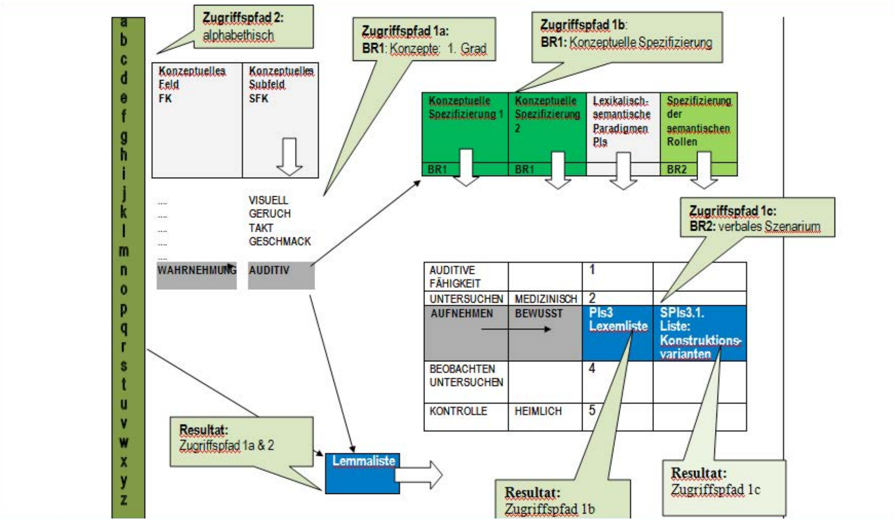
Der übliche und bekannteste Zugriff auf ein Wörterbuch erfolgt zwar aus einer semasiologisch-alphabetisch geleiteten Perspektive, für freie fremdsprachige Produktionszwecke ist diese Perspektive jedoch nur geeignet, wenn man das sprachliche Ausdrucksmittel schon kennt, bzw. schon ausgewählt hat und das WB nur noch zwecks Überprüfung bestimmter lexikologischer Parameter zur korrekten Anwendung konsultieren möchte. Wenn man aber noch kein sprachliches Ausdrucksmittel ausgewählt hat, weil man das Ausdrucksmittel für die Ausdrucksbedürfnisse gar nicht kennt, dann kann

[13] Siehe dazu auch verschiedene sprachvergleichende Studien in Verbindung mit lexikalisch-semantischen Paradigmen in Engelberg et al. (eds.) (2014).

[14] Im Rahmen von DICONALE ist eine breit angelegte Wörterbuchbenutzerumfrage entwickelt worden, die unter https://www.usc.es/gl/proxectos/diconale/aleman/enquisa.html [11.04.2014] abgerufen werden kann. Sie erfragt die Benutzergewohnheiten und Erwartungen im Bereich DaF und Ele an universitären und nicht universitären (Gymnasien, Sprachenschulen, Volkshochschule etc.) Lehrinstitutionen in Spanien, Deutschland und Portugal. Die Umfrageauswertungen liegen Ende 2014 vor.

[15] Siehe dazu auch Haß & Schmitz (2010), Klosa (ed.) (2008), Mann (2010), Müller Spitzer & Engelberg (2013), Storrer (2010) etc. und spezifisch zu dem Mehrwert der Internetlexikographie Engelberg & Lemnitzer 42009: 220 und Tarp 2012: 253).

man in einem einsprachig semasiologisch konzipierten WB nicht wirklich fündig werden (González Ribao & Proost 2014; Proost 2007). Daher soll die onomasiologisch-konzeptuelle Zugriffsstruktur[16] besonders für den fremdsprachigen Produktionskontext genutzt werden, wobei die konzeptuelle Referenz zusammen mit den jeweiligen verbalen Szenarien die zwei Hauptbezugsrahmen bilden und gleichzeitug das *tertium comparationis* für den Sprachvergleich stellen. Der erste Bezugsrahmen (BR1) bezieht sich auf die konzeptuellen Referenzen mit unterschiedlichem Spezifizierungsgrad, während der zweite Bezugsrahmen (BR2) sich an die Beschreibung der verbalen Szenarien annähert und die am verbalen Geschehen beteiligten Rollen beschreibt[17]. Beide Bezugsrahmen bilden die Grundlage des Zugriffspfades 1 mit den entsprechenden Spezifizierungen (1a-1c). Zusätzlich wird auch ein klassisch alphabetisch geleiteter Zugriffspfad 2 angeboten, der aber nicht zu der erwarteten semasiologischen Informationsperspektive führt, sondern die Benutzenden über eine Lemmaliste zu dem Zugriffspfad 1 zurückleitet. Die verschiedenen Zugriffspfade werden in *Abbildung 2* visualisiert und sollen im Folgenden genauer beschrieben werden. Als Beispiel dient uns das konzeptuelle Subfeld ersten Grades: AUDITIVE WAHRNEHMUNG (SFK[1]), welches dem Wahrnehmungsfeld untergeordnet ist (FK WAHRNEHMUNG). Der Benutzer greift auf das Wörterbuch aus einer konzeptuellen Suchperspektive zu, die durch anfängliche Benutzereinstellung in spanischer oder deutscher Sprache realisiert werden kann. Ein detailliertes Optionsleitsystem ermöglicht die Suche, Auffindung, Auswahl und den anschließenden Gebrauch der Ausdrucksform, die am besten dem kommunikativen Ausdrucksbedürfnis der jeweiligen Situation entspricht.


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
Zugriffspfad 2: alphabethisch
Zugriffspfad 1a: BR1: Konzepte: 1. Grad
Zugriffspfad 1b: BR1: Konzeptuelle Spezifizierung
Konzeptuelles Feld FK
Konzeptuelles Subfeld SFK
Konzeptuelle Spezifizierung 1
Konzeptuelle Spezifizierung 2
Lexikalisch-semantische Paradigmen Pls
Spezifizierung der semantischen Rollen
BR1
BR1
BR2
....
VISUELL
GERUCH
TAKT
GESCHMACK
WAHRNEHMUNG
AUDITIV
Zugriffspfad 1c: BR2: verbales Szenarium
Resultat: Zugriffspfad 1a & 2
Lemmaliste
Resultat: Zugriffspfad 1b
Resultat: Zugriffspfad 1c

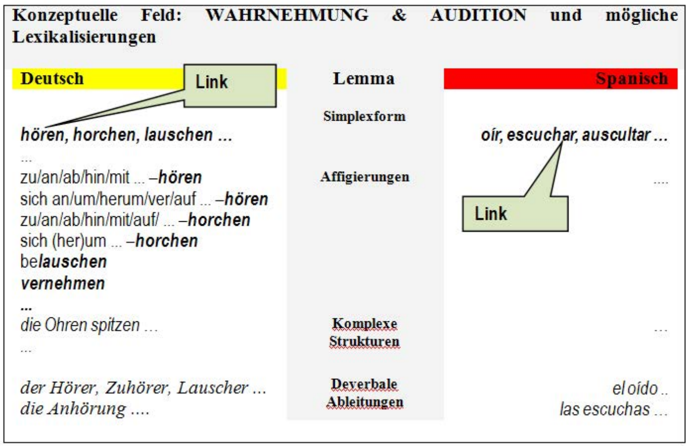
AUDITIVE FÄHIGKEIT		1	
UNTERSUCHEN	MEDIZINISCH	2	
AUFNEHMEN	BEWUSST	Pls3 Lexemliste	SPls3.1. Liste: Konstruktionsvarianten
BEOBACHTEN UNTERSUCHEN		4	
KONTROLLE	HEIMLICH	5	

Abbildung 2: Unterschiedliche Zugriffspfade für DICONALE

[16] Die zur Verfügung stehenden klassischen konzeptuell-onomasiologisch geordneten Wörterbücher und Nachschlagewerke wie Wehrle & Eggers und Dornseiff oder Casares für das Spanische sind wegen ihrer Komplexität für den hier anvisierten Benutzerkreis ungeeignet. Neuere Studien, wie z.B. zu den Kommunikationsverben (Harras et al.) mit online-Zugang (Proost) verfolgen einen sehr komplexen Bezugsrahmen, der für unsere Benutzersituation ebenfalls nicht geeignet ist.

[17] Siehe dazu Studien zu FrameNet: Boas 2013, Boas & Dux 2013, FrameNet Spanisch: Subirats 2009;

Der Zugriffspfad 1 führt den Benutzer mithilfe des entsprechenden Bezugsrahmens BR1 (WAHRNEHMUNG & AUDITION) von einer allgemeinen konzeptuellen Referenz zu einer Lemmaliste, in der neben den verbalen Lemmata in Simplexform und den entsprechenden affigierten Formen auch einige mehrteilige Lemmata und deverbale Formen aufgeführt werden (*Abbildung 3*).

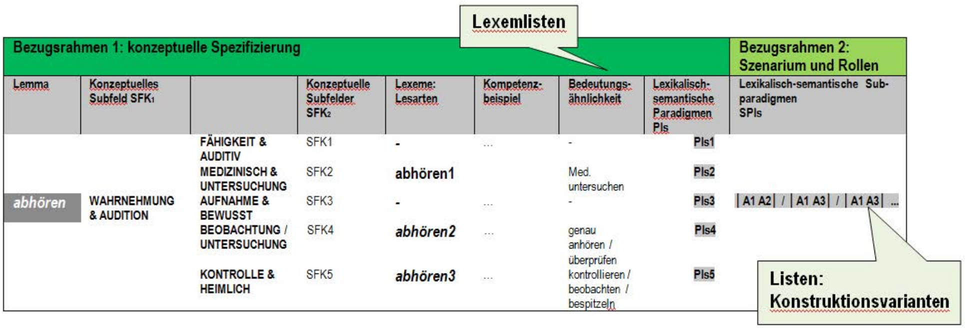


Konzeptuelle Feld: WAHRNEHMUNG & AUDITION und mögliche Lexikalisierungen

Deutsch	**Lemma**	**Spanisch**
hören, horchen, lauschen … …	**Simplexform**	***oír, escuchar, auscultar …***
zu/an/ab/hin/mit … –***hören*** sich an/um/herum/ver/auf … –***hören*** zu/an/ab/hin/mit/auf/ … –***horchen*** sich (her)um … –***horchen*** be***lauschen*** ***vernehmen*** …	**Affigierungen**	….
die Ohren spitzen … …	**Komplexe Strukturen**	…
der Hörer, Zuhörer, Lauscher … *die Anhörung ….*	**Deverbale Ableitungen**	el oído .. las escuchas …

Abbildung 3: Lemmalisten in beiden Sprachen zu dem BR1: WAHRNEHMUNG & AUDITION.

Diese Lemmalisten sind allerdings nur von Nutzen, wenn die angeführten internen Links den Benutzer zu den verschiedenen Lesarten führen und die Vernetzung innerhalb des entsprechenden konzeptuellen Feldes und den dazu gehörigen lexikalisch-semantischen Paradigmen abgerufen werden kann (*Abbildung 4*: Beispiel *abhöre*n).


Lexemlisten
Listen: Konstruktionsvarianten

Bezugsrahmen 1: konzeptuelle Spezifizierung								Bezugsrahmen 2: Szenarium und Rollen
Lemma	Konzeptuelles Subfeld SFK₁		Konzeptuelle Subfelder SFK₂	Lexeme: Lesarten	Kompetenz-beispiel	Bedeutungs-ähnlichkeit	Lexikalisch-semantische Paradigmen Pls	Lexikalisch-semantische Sub-paradigmen SPls
abhören	WAHRNEHMUNG & AUDITION	FÄHIGKEIT & AUDITIV	SFK1	-	…	-	Pls1	
		MEDIZINISCH & UNTERSUCHUNG	SFK2	abhören1		Med. untersuchen	Pls2	
		AUFNAHME & BEWUSST	SFK3	-	…	-	Pls3	| A1 A2 | / | A1 A3 | / | A1 A3 | …
		BEOBACHTUNG / UNTERSUCHUNG	SFK4	*abhören2*	…	genau anhören / überprüfen	Pls4	
		KONTROLLE & HEIMLICH	SFK5	*abhören3*	…	kontrollieren / beobachten / bespitzeln	Pls5	

Abbildung 4: Lemma und Linkangebot zu Lesarten in Verbindung mit dem BR1 und BR2.

Auf dem Zugriffspfad 1 kann der Benutzer auch direkt – geleitet durch weitere konzeptuelle Spezifizierungen des BR1(1a-1c) - zu möglichen Benennungseinheiten gelangen, die mit Bezug auf das entsprechende lexikalisch-semantische Paradigma in Lexemlisten mit Kompetenzbeispielen angeordnet werden. So wird z.B. der Benutzer durch den BR1 und eine Spezifizieung durch AUFNEHMEN & BEWUSST zu dem lexikalisch-semantischen Paradigma Pls3: „zuhören" geführt. Die Ausdruckmittel, die dem BR1 entsprechen, sind für das Deutsche: (*sich) anhören1, zuhören1, hören2, horchen1, lauschen1* und für das Spanische*: escuchar 2 y oír2*. Eine Auswahl von Kompetenzbeispielen illustrieren die Bedeutung durch den Kontext (*Abbildung 5*: Suchergebnisse bezüglich BR1). Verben wie Dt.: *abhören, abhorchen, belauschen* und Sp.: *auscultar* etc., werden in diesem lexikalisch-semantischen Paradigma konsequenterweise nicht aufgeführt, da sie in keiner ihrer Lesarten zu dem besagten Paradigma gehören. Der Bezugsrahmen 2 ermöglicht die Spezifizierung des verbalen Szenariums durch die beteiligten Rollen und Muster. Die zum lexikalisch-semantischen Paradigma Pls3 gehörigen Rollen sind u.a. der Hörer (R1), die Äußerung, die wahrgenommen wird (R2), ein Lebewesen, das als Geräuschquelle wahrgenommen wird (R3), eine nicht belebte Entität, die als Geräuschquelle wahrgenommen wird (R4) und das wahrgenommene Geräusch selbst (R5). Der BR2 stellt einen wichtigen Auswahlfaktor für den Benutzer dar, denn nicht alle Lexeme in einem lexikalisch-semantischen Pardigma deuten auf dasselbe Szenarium hin. Aus der Kookkurrenzanalyse zu *lauschen* und *zuhören* (CCDB: Cyril) lässt sich z.B. schließen, dass *lauschen* häufiger in Verbindung mit dem Szenarium: "jemand (R1: Hörer) nimmt ein Geräusch wahr (R5: Klang, Gesang, Geräusch...)" auftritt, während *zuhören* andere Szenen vorzieht: (z.B.: R1: Hörer & R3: belebte Geräuschquelle: Leute ...). Je nach Beteiligung der Rollen liegt das eine oder andere Argumentstrukturmuster vor, welches in einem Menu mit den entsprechenden Vorgaben selegiert werden kann. Wenn der Benutzer Ausdrucksformen für "jemand hört aufmerksam zu, was jemand sagt" sucht, bezieht es sich auf ein ganz bestimmtes Szenarium, in dem die Rollen Hörer (R1) und Äußerung (R2) verbalisiert werden sollen. Die Argumentstruktur |jemand A1 *V* etwas A2| bildet daher den BR2, durch den nur die Konstruktionsvarianten einer Lesart selegiert werden, die dieses Szenarium realisieren können. In unserer L2-*output*-Benutzersituation kann davon ausgegangen werden, dass der Benutzer genau und nur das Szenarium sucht, welches zu seinem Ausdruckswunsch passt. Nach der Auswahl des entsprechenden Musters wird dem Benutzer eine Liste von Konstruktionsvarianten eines lexikalisch-semantischen Paradigmas zusammen mit Kompetenzbeispielen für die ausgewählte Sprache (Beispiele 1-5$_{dt}$) angeboten, die beide Bezugsrahmen miteinander teilen (*Abbildung 5*).[18] Nach der Auswahl der einen oder anderen Konstruktionsvariante kann der Benutzer in weiteren Schritten detaillierte Information zu den 4 Beschreibungsmodulen erhalten (*Abbildung 1*).

(1.1$_{dt}$) Sie ***hören sich*** die Probleme ***an,*** die den Kindern auf den Nägeln brennen [...]. (R97/SEP.72386 Frankf. Rundschau, 15.09.1997, S. 4).

(1.2$_{dt}$) Vier Jahre hätten die Vorbereitungen gedauert, man habe sich in anderen Städten vergleichbare Ansagen ***angehört*** und sich jetzt für diese Lösung entschieden. (M13/JUL.01277 Mannh. Morgen, 04.07.2013, S. 20).

(1.3$_{dt}$) Auch die Mitarbeiterinnen im Bürgercenter werden aufatmen, denn sie mussten sich in den vergangenen drei Wochen so manche nicht immer freundlich vorgetragene Beschwerde ***anhören.*** (BRZ10/NOV.09775 Braunschw. Z., 19.11.2010).

(2.1$_{dt}$) Sie ***hörte*** die Aussagen der verängstigten Kinder. (RHZ05/JUN.16629 RZ, 15.06.2005).

(2.2$_{dt}$) Frei nach dem Grundsatz „Erst mal *hören,* **was** die Zeugen wissen", verlegte sich das Muskelpaket und Vater von drei Kindern aufs Schweigen. (RHZ03/JUL.12417 RZ, 16.07.2003).

[18] Zur Behandlung von Argumentstrukturen und Lesartdisambiguierung siehe u.a. Engelberg (2010) für das Deutsche und Porto Dapena et al. (2008) für das Spanische.

(3.1$_{dt}$) Man setzt sich, ***horcht*** der Feldpredigt und geniesst die Sonne. (A00/JUN.38103 St. Galler Tagblatt, 02.06.2000)

(3.2$_{dt}$) Er ***horchte*** auf die Worte der Reisenden, die von Bord gingen, und wenn sie deutsch sprachen, redete er sie an. (P97/MAR.11400 Die Presse, 22.03.1997).

(3.3$_{dt}$) Vor kurzem lud der Tischtennisverein Züllig zur 15. Generalversammlung. Zahlreiche Mitglieder folgten der Einladung und ***horchten*** den erfreulichen Neuigkeiten. (A99/SEP.68281 St. Galler Tagblatt, 30.09.1999).

4.1$_{dt}$) Jago, das Aas, ***lauscht*** hinter Säulenfluchten den honorigen Erklärungen Othellos vor dem Rat von Venedig. (UN93/JUN.01879 NN, 26.06.1993, S. 22).

(4.2.$_{dt}$) Rund ein Drittel der Steinacher Ortsbürger ***lauschten*** den Worten des Präsidenten, als dieser auf das vergangene Jahr zurückblickte. (A13/APR.06663 St. Galler Tagblatt, 17.04.2013, S. 34).

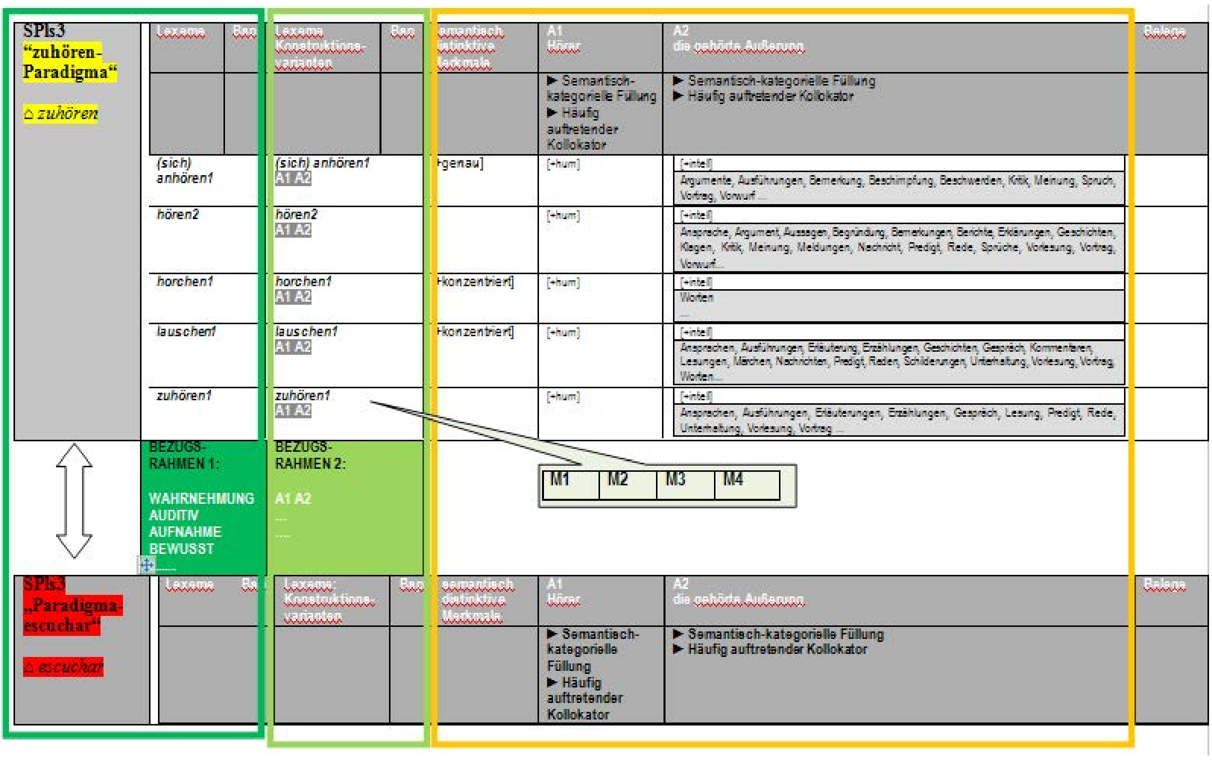
(4.3$_{dt}$) Plätzchenduft schwebte durch die Schule in Wattenheim, denn dort entpuppten sich die Jungen und Mädchen nicht nur als gute Zuhörer, sondern auch als exzellente Plätzchenbäcker. Gemeinsam mit Lehrerin Ilka Peter hatten sich zehn Kinder aus den Klassen Zwei, Drei und Vier versammelt und ***lauschten*** auf die Erzählung „Weihnachten im Möwenweg" von Kirstin Boie. (M08/DEZ.95513 Mannh. Morgen, 08.12.2008, S. 19).

(5.1$_{dt}$) Mit großem Interesse hatten die Pflegekräfte [...], dem Vortrag ***zugehört*** [...]. (M03/FEB.11701 Mannh. Morgen, 22.02.2003).

(5.2$_{dt}$) Helmut Mägdefrau, der stellvertretende Tiergartendirektor, hat der Debatte lange schweigend ***zugehört*** [...].(NUZ13/JUL.01777 Nürnberger Zeitung, 20.07.2013, S. 11).

(5.3$_{dt}$) Er marschiert von Haustür zu Haustür und ***hört*** *zu,* was ihm die Leute erzählen. (RHZ96/AUG.03591 RZ, 07.08.1996).

Der Zugriffspfad 2 kann genutzt werden, wenn der Benutzer schon ein mögliches Lexem für seine Ausdrucksbedürfnisse kennt. Über eine alphabetisch angeordnete Leiste erhält er Zugriff zu dem gesuchten Lemma, und wird dann über die Zuordnung zu dem entsprechenden konzeptuellen Feld (*Abbildung 2*) und den feldrelevanten Lesarten zu dem einen oder anderen lexikalisch-semantischen Paradigma – entsprechend dem Zugriffspfad 1, geleitet (siehe *Abbildung 4*: *abhören*).



SPls3 “zuhören-Paradigma“ ⌂ zuhören	Lexeme	Bsp	Lexeme Konstruktions-varianten	Bsp	emantisch istinktive Merkmale	A1 Hörer	A2 die gehörte Äußerung	Belege
						▶ Semantisch-kategorielle Füllung ▶ Häufig auftretender Kollokator	▶ Semantisch-kategorielle Füllung ▶ Häufig auftretender Kollokator	
	(sich) anhören1		(sich) anhören1 A1 A2		+genau]	[+hum]	[+intel] Argumente, Ausführungen, Bemerkung, Beschimpfung, Beschwerden, Kritik, Meinung, Spruch, Vortrag, Vorwurf ...	
	hören2		hören2 A1 A2			[+hum]	[+intel] Ansprache, Argument, Aussagen, Begründung, Bemerkungen, Berichte, Erklärungen, Geschichten, Klagen, Kritik, Meinung, Meldungen, Nachricht, Predigt, Rede, Sprüche, Vorlesung, Vortrag, Vorwurf...	
	horchen1		horchen1 A1 A2		+konzentriert]	[+hum]	[+intel] Worten ...	
	lauschen1		lauschen1 A1 A2		+konzentriert]	[+hum]	[+intel] Ansprachen, Ausführungen, Erläuterung, Erzählungen, Geschichten, Gespräch, Kommentaren, Lesungen, Märchen, Nachrichten, Predigt, Reden, Schilderungen, Unterhaltung, Vorlesung, Vortrag, Worten...	
	zuhören1		zuhören1 A1 A2			[+hum]	[+intel] Ansprachen, Ausführungen, Erläuterungen, Erzählungen, Gespräch, Lesung, Predigt, Rede, Unterhaltung, Vorlesung, Vortrag ...	
⇕	BEZUGS-RAHMEN 1: WAHRNEHMUNG AUDITIV AUFNAHME BEWUSST		BEZUGS-RAHMEN 2: A1 A2		M1 M2 M3 M4			
SPls3 „Paradigma-escuchar“ ⌂ escuchar	Lexeme	Bs	Lexeme: Konstruktions-varianten	Bsp	semantisch distinktive Merkmale	A1 Hörer	A2 die gehörte Äußerung	Belege
						▶ Semantisch-kategorielle Füllung ▶ Häufig auftretender Kollokator	▶ Semantisch-kategorielle Füllung ▶ Häufig auftretender Kollokator	

Abbildung 5: Suchergebnisse bezüglich BR1 und BR2 und Selektion durch distinktive Bedeutungsmerkmale und Kollokatoren (Ausschnitt).

4 Auswählen und Anwenden

Nachdem der Benutzer über die zwei Bezugsrahmen zu einer Lexemliste mit den möglichen Konstruktionsvarianten, die dem Szenarium entsprechen, gelangt ist (*Abbildung 5*), spielen einige distinktive Merkmale bei der weiteren Auswahl aus der Ausdrucksvielfalt eine Rolle. Eine Gesamtdarstellung des szenenspezifischen Teilparadigmas soll dem Benutzer dabei helfen, für seine Kommunikationssituation das adäquate Ausdrucksmittel zu finden. Dabei sind die semantisch distinktiven Merkmale ebenso relevant, wie die semantisch-kategorielle Information[19] und Hinweise zu häufig auftretenden Kollokatoren[20]. Aus der Liste der Ausdrucksmöglichkeiten für die Lexikalisierung von WAHRNEHMUNG & AUDITION & AUFNAHME & BEWUSST (BR1) und dem Argumentstrukturmuster │A1 HÖRER A2 ÄUßERUNG│ sollen z.B. die Lexeme ausgewählt werden, die außerdem das Merkmal [konzentriert] lexikalisieren und die mit „Wort“ als möglichem Kollokator auftreten können. Als Ergebnis erhält der Benutzer die Lexeme *horche*n und *lauschen*(Beispiel 3.2_{dt} & 4.2_{dt}). Im Fall einer Verbalisierung

[19] Siehe dazu: Engel 2004.

[20] Die Information zu den in Tabelle 5 aufgeführten häufigen Kollokatoren ist anhand der entsprechenden CCDB-Kookkurrenzanalysen (Cyril Belica) und den Wortprofilen aus DWDS erstellt worden. Die empirische Analyse für DICONALE sieht bei der Kodierung der Belegsamples auch eine Häufigkeitsanalyse der Rollenfüller vor und soll mit korpusgenerierten Daten, die auf umfangreichem Korpusmaterial beruhen, abgeglichen werden.

in Verbindung mit „Beschwerde / Beschimpfung“ scheint *anhöre*n (Beispiel 1.3$_{dt}$) am adäquatesten zu sein (*Abbild 5*). Nach Auswahl einer Konstruktionsvariante hat der Benutzer dann die Möglichkeit detallierte einzellexematische Information zu den vier Beschreibungsmodulen (*Abbildung 1*) zu erhalten. An dieser Stelle des Such- und Auswahlprozesses angelangt, wird die onomasiologisch orientierte Ausgangsperspektive, für die zwei Zugangspfade angeboten wurden, mit einer semasiologischen Zugriffsstruktur verbunden (Blank & Koch 2003, Mingorance 1994, Proost 2007, Reichmann 1989) und bietet auf der Mikrostrukturebene die relevante Gebrauchsinformation an.

Daneben ist es aber auch möglich, die einzellexematische Information weiterhin im Kontrast zu den anderen Elementen des Teilparadigmas zu erhalten (*Abbildung 6*). In diesem Fall wird dem Benutzer z.B. deutlich, dass nicht alle Elemente des Paradigmas denselben Satzbauplan (SBP) aufweisen. Das zweite Argument erfährt unterschiedliche morphosyntaktische Realisierungsformen. Während *sich anhören1* eine Akkusativergänzung (Eakk) (Bsp. 1$_{dt}$) und *zuhören1* (Bsp. 5.1$_{dt}$, 5.2$_{dt}$) eine Dativergänzung (Edat) regiert, ist bei *lauschen1* (Bsp. 4$_{dt}$) und *horchen1* (Bsp. 3$_{dt}$) die Alternanz zwischen Edat/Eprp auffällig. Die Information zu den einzelnen SBP können zwar auch in Valenzwörterbüchern konsultiert werden, aber erst der Überblick der Vielfalt in einem Teilparadigma durch ein strukturiertes Informationsangebot ermöglicht dem L2-Benutzer eine bewusste Auswahl und Anwendung.

Ebenso ist der strukturierte Überblick der Information in Teilparadigmen für den Sprachenkontrast von höchstem Interesse. Bei der Selektion der möglichen Entsprechung *escuchar* 2 zu allen Elementen des besagten Paradigmas erhält der Benutzer die Information zu dem spanischen Verb bezüglich der vier Beschreibungsmodule. Besonders auffällig sind Unterschiede in der morphosyntaktischen Realisierungsform. Dem dativisch realisierten A2 in *zuhören* entspricht z.B. ein direktes Objekt in *escuchar*2 und *oír2* (Bsp. 1-2$_{sp}$), während die Realisierungsmöglichkeit durch eine Präpositionalphrase in mehreren deutschen Lexemen des Subparadigmas in *escuchar*2 nicht möglich ist. Eine weitere kontrastiv relevante Auffälligkeit in diesem Subparadigma ist die Beobachtung, dass sich die deutschen Verben teilweise durch distinktive Bedeutungsmerkmale unterscheiden lassen können, während das spanische Verb eine viel allgemeinere Bedeutung besitzt, und daher die kommunikative Notwendigkeit zu Spezifizierungen über adverbiale Zusätze erfolgen muss (1.4-1.5$_{sp}$). Weitere aufschlussreiche Divergenzen zwischen beiden Sprachen und bezüglich aller konzeptueller Felder, die im Rahmen von DICONALE behandelt werden, sind in den unterschiedlichen satzförmigen Komplementrealisierung zu erwarten (2.3$_{dt}$, 5.3$_{dt,}$ 1.3$_{sp}$). Für den korrekten Gebrauch in der fremdsprachigen Produktionssituation sind derartige kontrastive Informationen von enormer Relevanz und sollten dem Benutzer klar vor Augen geführt werden.

(1.1$_{esp}$) En la tribuna de invitados ***escucharon*** el debate el secretario general de UGT [...] (El Mundo, 20/11/2002)

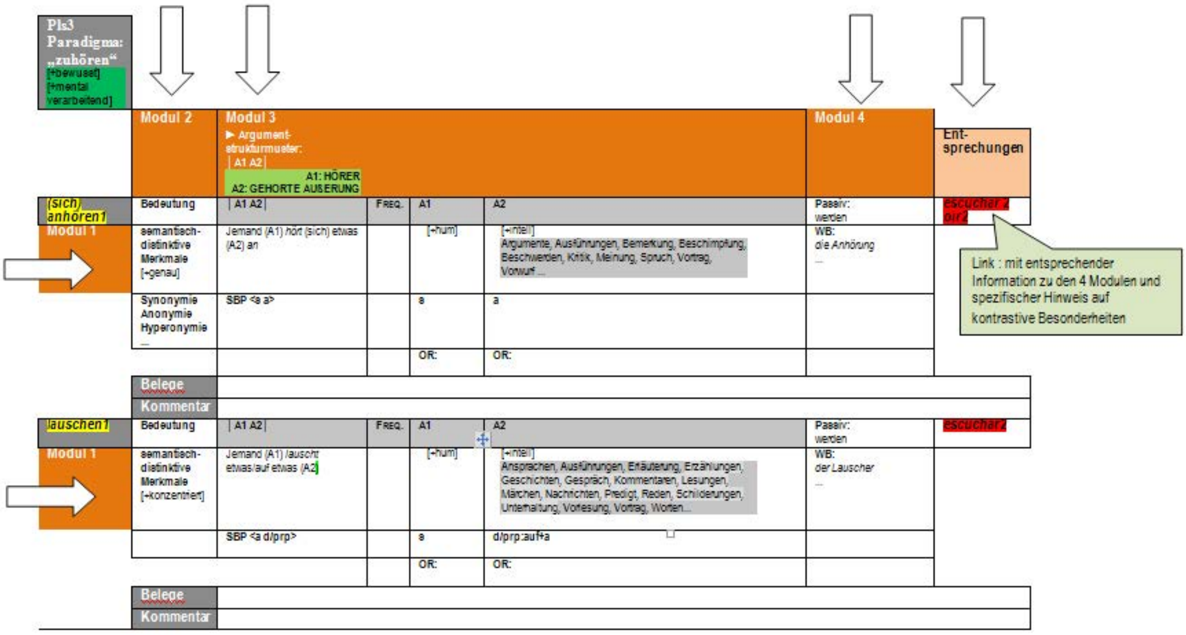
(1.2$_{esp}$) Allí, mientras ***escuchaba*** las noticias por televisión, se quedó impresionada cuando una locutora narraba con frialdad el siguiente suceso [...]: (El Diario Vasco, 31/01/2001)

(1.3$_{sp}$) Fue en tono de broma, pero también hay que ***escuchar*** lo que dicen los demás“ (El País, 02/06/1989)

(1.4_{sp}) La comitiva ***escuchó*** atentamente las explicaciones de Fernando Checa, director del museo, y de José Antonio Fernández Ordoñez, presidente del patronato. (El País, 18/11/1997)

(1.5_{sp}) Horacio ***escucha*** con atención mi relato. (La Razón, 20/12/2001)

(2.1_{esp}) Prácticamente a la misma hora, el juez de instrucción de París, Gilles Boulouque, que se encarga de los atentados de 1986, y un tribunal islámico, ***oían*** las declaraciones de los dos personajes. (El País, 01/12/1987)



Pls3 Paradigma: „zuhören“ [+bewusst] [+mental verarbeitend]	Modul 2	Modul 3 ► Argumentstrukturmuster: | A1 A2 | A1: HÖRER A2: GEHORTE AUßERUNG				Modul 4	Ent-sprechungen
(sich) anhören1	Bedeutung	| A1 A2 |	FREQ.	A1	A2	Passiv: werden	escuchar 2 oír2
Modul 1	semantisch-distinktive Merkmale [+genau]	Jemand (A1) *hört* (sich) etwas (A2) an		[+hum]	[+intell] Argumente, Ausführungen, Bemerkung, Beschimpfung, Beschwerden, Kritik, Meinung, Spruch, Vortrag, Vorwurf ...	WB: *die Anhörung* ...	
	Synonymie Anonymie Hyperonymie ...	SBP <a a>		a	a		
				OR:	OR:		
	Belege						
	Kommentar						
lauschen1	Bedeutung	| A1 A2 |	FREQ.	A1	A2	Passiv: werden	escuchar2
Modul 1	semantisch-distinktive Merkmale [+konzentriert]	Jemand (A1) *lauscht* etwas/auf etwas (A2)		[+hum]	[+intell] Ansprachen, Ausführungen, Erläuterung, Erzählungen, Geschichten, Gespräch, Kommentaren, Lesungen, Märchen, Nachrichten, Predigt, Reden, Schilderungen, Unterhaltung, Vorlesung, Vortrag, Worten...	WB: *der Lauscher* ...	
		SBP <a d/prp>		a	d/prp:auf+a		
				OR:	OR:		
	Belege						
	Kommentar						

Link : mit entsprechender Information zu den 4 Modulen und spezifischer Hinweis auf kontrastive Besonderheiten

Abbildung 6: Tabellarische Übersicht eines SPIs: Informationsangebot für Module 1-4 zur Auswahl (Teilinformation).

5 Ausblick

In dem Beitrag wurden die verschiedenen Zugriffspfade zu der für fremdsprachige Produktionszwecke relevanten Information vorgestellt und der Versuch unternommen, die Perspektive eines Wörterbuchbenutzers im fremdsprachigen Produktionsprozess zu verfolgen, um gemäß seiner Bedürfnisse den Such-, Auswahl- und Anwendungsprozess zu gestalten. Im Laufe der Ausführungen ist deutlich geworden, dass das Verfolgen einer onomasiologisch-konzeptuell angeordneten Informationsdarbietung im zweisprachigen Kontext ein komplexes Unterfangen darstellt. Das hier vorgestellte Projekt soll als Prototyp eines neuartigen Konsultationswerkes verstanden werden, welches durch einen primär konzeptuell und szenenorientierten Bezugsrahmen den Zugang anbietet und durch ein komplexes Such- und Selektionsverfahren den Benutzer zu einer Reihe von bedeutungsähnlichen Lexemen und Konstruktionsvarianten führt, aus denen nach verschiedenen Kriterien für den kontextadäquaten Gebrauch ausgewählt werden muss. Gebrauchsrelevante einzelsprachige und kontrastive Information sollen eine korrekte Anwendung in der jeweiligen L2 ermöglichen. Übersichtliche, tabellarische inter- und intralinguale Gegenüberstellungen der Elemente eines Paradigmas sollen dem Benutzer bei der bewussten Auswahl im Falle der Ausdrucksvarietät behilflich sein. Obwohl gerade in den letzten Jahren immer mehr lexikographische Ressourcen, vor allem mit *online*-Zugang, entwickelt wurden, steht bis jetzt noch ein umfassendes, verlagsunabhängiges und benutzergerechtes lexikographisches Informationsangebot für den fremdsprachigen Produktionsprozess in den Bereichen DaF und Ele aus. DICONALE hat sich zum Ziel gesetzt, diesem Desideratum einen Schritt näher zu kommen.

6 Literatur

6.1 Wörterbücher und andere Ressourcen

CanooNet: Portal: Deutsche Wörterbücher und Grammatik. http://www.canoo.net/ [11.04.2014].

Casares, J. (1942/^{3}2001): Diccionario ideológico de la lengua española. Barcelona.

CCDB: Cyril Belica: Kookkurrenzdatenbank V3.3. Eine korpuslinguistische Denk- und Experimentierplattform © 2001 ff., Institut für Deutsche Sprache, Mannheim. http://corpora.ids-mannheim.de/ccdb/ [11.04.2014]

DA=Diccionario de Alcalá: Alvar Ezquerra, M. (dir.): Diccionario para la enseñanza de la lengua española. Español para extranjeros, Barcelona, Vox y Universidad de Alcalá, (1995/^{2}2000). Online-Version über http://www.diccionarios.com/ [11.04.2014]

Dornseiff, Franz/Wiegand, Herbert E./Quasthoff, Uwe (^{8}2004). Der deutsche Wortschatz nach Sachgruppen. 8. Neubearbeitete Fassung. Berlin. (print+elektronisch).

DS=Diccionario Salamanca: Gutiérrez Cuadrado, J. (dir.): Diccionario Salamanca de la lengua española, Madrid, Santillana y Universidad de Salamanca,1996/2007.

Duden-Portal: online http://www.duden.de/ [11.04.2014]

DWB-DaF: Duden (^{2}2010): Deutsch als Fremdsprache – Standardwörterbuch. Mannheim.

DWDS: Digitales Wörterbuch der deutschen Sprache. http://www.dwds.de/ [11.04.2014]

GWB-DaF: Kempcke, G.: Wörterbuch Deutsch als Fremdsprache. Berlin/NewYork: de Gruyter, 1999.

LEO: zweisprachiges Wörterbuchportal. http://www.leo.org/ [11-04.2014]

LGWB-DaF: Götz, D.,Haensch, G. & Wellmann, H.: Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Berlin, München: Langenscheidt, ^{3}2010. Online-Zugang über: TheFreeDictionary http://de.thefreedictionary.com [11-04.2014]

LHWB: Langenscheidts Handwörterbuch Spanisch ^{12}1982: Teil 1: Spanisch - Deutsch (LHWB-SD), Teil 2: Deutsch – Spanisch (LHWB-DS), Berlin, München: Langenscheidt.

LHWBe: Langenscheidts Handwörterbuch Spanisch 2006. Spanisch – Deutsch (LHWBe-SD) / Deutsch – Spanisch (LHWBe-DS). Berlin, München: Langenscheidt. Elektronische Fassung.

PGWB-DaF: Pons Großwörterbuch DaF (2004): Stuttgart: Pons.

Pons: Das Sprachenportal. http://de.pons.eu/ [11.04.2014]

SGIWBe: Slaby, R. J. & Grossmann, R. & Illig, C. (^{5}2003): Wörterbuch der spanischen und deutschen Sprache. Spanisch - Deutsch (SGIe-SD), Deutsch – Spanisch (SGIe-DS). Wiesbaden: Brandstetter Verlag. + Elektronische Fassung.

SM-Clave: SM-portal: Diccionario de uso del español actual. http://clave.smdiccionarios.com/app.php [11-04.2014]

SM-Diccionario: Maldonado, C. (dir.): Diccionario de español para extranjeros. Madrid: SM, 2002.

The free Dictionary: http://de.thefreedictionary.com/ [11.04.2014]

Wehrle, H.&Eggers, H. (1961/17 1993): Deutscher Wortschatz: ein Wegweiser zum treffenden Ausdruck. Stuttgart.

WGWB-DaF: Wahrig (2008): Großwörterbuch Deutsch als Fremdsprache. Berlin. Online-Zugang über Wissens-Portal: http://www.wissen.de [11.04.2014]

Wortschatz Universität Leipzig: Portal http://wortschatz.uni-leipzig.de/ [11.04.2014]

6.2 Fachliteratur

Abel, A. (2008). *ELDIT* (Elektronisches Lernerwörterbuch Deutsch-Italienisch) und *elexiko* im Vergleich. In Klosa, A. (ed.) 1/2008. 175-189. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2008-1.pdf. [11.04.2014]

Blank, A. & Koch, P. (eds.) (2003). *Kognitive romanische Onomasiologie und Semasiologie.* Tübingen: Niemeyer.

Boas, H. (2013). Wieviel Wissen steckt in Wörterbüchern? Eine Frame-semantische Perspektive. In *Zeitschrift für Angewandte Linguistik* 57, 75-97.

Boas, H. C. & Ryan Dux (2013). Semantic frames for foreign-language education: Towards a German frame-based dictionary. In *Veridas On-Line* 1/2013, 81-100.

Dentschewa, E. (2006). DaF-Wörterbücher im Vergleich: Ein Plädoyer für „Strukturformeln". In Dimova, Ana et al. (eds.): *Zweisprachige Lexikographie und Deutsch als Fremdsprache*. Hildesheim: Olms, 113-128.

Domínguez Vázquez, Mª J. et al. (2013). Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen. In Domínguez Vázquez, Mª J. (ed.), 135-172.

Domínguez Vázquez, M. (ed.) (2013). *Trends in der deutsch-spanischen Lexikographie.* Frankfurt: P. Lang Edition.

Domínguez Vázquez, Mª J., Gómez Guinovart, X. & Valcárcel Riveiro, C. (eds.). *Lexicografía románica. Aproximaciones a la lexicografía moderna y contrastiva.* (Coord.: Sánchez Palomino, Mª. D. & Domínguez Vázquez, Mª J. (vol. 2). Berlin: de Gruyter (im Druck).

Engel, U. (2004). Deutsche Grammatik. – Neubearbeitung. München: iudicium.

Engelberg, S. & Lemnitzer, L. (2001), (4 2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Engelberg, St. (2010). Die lexikographische Behandlung von Argumentstrukturvarianten in Valenz- und Lernerwörterbüchern. In Fischer, K., Fobbe, E. & Schierholz, St. (eds.), *Valenz und Deutsch als Fremdsprache,* Frankfurt a. M.: P. Lang, 113-141.

Engelberg, St. (2014a). The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment. In Boas, H. & Ziem, A. (eds.): *Constructional approaches to argument structure in German.* Boston, Berlin: De Gruyter Mouton. (im Druck).

Engelberg, St. (2014b). Gespaltene Stimulus-Argumente bei Psych-Verben. Quantitative Verteilungsdaten als Indikator für die Dynamik sprachlichen Wissens über Argumentstrukturen. In: Engelberg, St. et al. (eds.): *Argumentstruktur – Valenz – Konstruktionen.* Tübingen: Narr. (im Druck).

Engelberg, St., Koplenig, A., Proost, K., Winkler, E. (2012). Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations. In *Lexicographica 28,* 13-48.

Engelberg, St., Meliss, M., Prosst, K. &Winkler, E. (eds.) (2014). *Argumentstruktur – Valenz – Konstruktionen.* Tübingen: Narr. (im Druck)

Fuentes Morán, Mª T. (1997). Gramática en la lexicografía bilingüe. Morfología y sintaxis en diccionarios español-alemán desde el punto de vista del germanohablante. Tübingen: Niemeyer.

González Ribao, V. & Meliss, M. (2014). Vorschläge zur Ausarbeitung eines onomasiologisch-konzeptuell orientierten Produktionswörterbuches im zweisprachigen Lernerkontext: Deutsch-Spanisch. In Calañas Continente, J.A. et al. (eds.). *Wörterbücher des Deutschen.* Frankfurt a. M.: P. Lang (Reihe: Studien zur Linguistik des Deutschen – Spanische Akzente) (im Druck).

González Ribao, V. & Proost, K. (2014). El campo léxico al servicio de la lexicografía: Un análisis contrastivo en torno a algunos subcampos de los verbos de comunicación en alemán y español. In Domínguez Vázquez, Mª J. et al. (eds.): *Lexicografía*. (Coord.: Sánchez Palomino, Mª. D. & Domínguez Vázquez, Mª J.). (vol. 2). (im Druck).

Haensch, G. & Omeñaca, C. (1997, ^{2}2004). (coords.): *Los diccionarios del español en el siglo XXI*. Salamanca: Ediciones Universidad.

Haß, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin: de Gruyter.

Haβ, U. & Schmitz, U. (2010). Lexikographie im Internet 2010 – Einleitung. In Gouws, R. H. et al. (eds.). *Lexicographica. Internationales Jahrbuch für Lexikographie. 26*. Berlin: de Gruyter, 1-18.

Hausmann, F. J. (1991). Die zweisprachige Lexikographie Spanisch-Deutsch, Deutsch-Spanisch. In Steger H. & Wiegand, H.E. (eds.). *Wörterbücher: Ein Internationales Handbuch zur Lexikograph*ie. Berlin/NewYork: de Gruyter. 2987-2991.

Herbst, Th. & Klotz, M. (2003). *Lexikografi*e. Paderborn: Schöningh.

Harras, G., Winkler, E. et al. (2004): *Handwörterbuch deutscher Kommunikationsverben*. Teil 1: Wörterbuch. Berlin.

Harras, Gisela, Proost, K. & Winkler, E. (2007). *Handbuch deutscher Kommunikationsverben*. Teil 2: Lexikalische Strukturen. Berlin. Und: Proost, K.: Online-Nachschlagewerk Kommunikationsverben:

Kemmer, K. (2010). Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationsraster mit 39 Beurteilungskriterien. In *Online publizierte Arbeiten zur Linguistik*. OPAL2/2010, 1-33. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2010-2.pdf. [11.04.2014]

Klosa, A. (ed.) (2008). *Lexikographische Portale im Internet*. *OPAL-Sonderheft* 1/2008, . [http://pub.ids-mannheim.de/laufend/opal/pdf/opal2008-1.pdf]. [11.04.2014]

Klosa, A., Koplenig, A. & Töpel, A. (2011). Benutzerwünsche und Meinungen zu einer optimierten Wörterbuchpräsentation – Ergebnisse einer Onlinebefragung zu „elexiko". In *Online publizierte Arbeiten zur Linguisti*k: OPAL 3/2011. Mannheim: Institut für Deutsche Sprache. [11-04.2014]

Mann, M. (2010). Internet-Wörterbücher am Ende der „Nulljahre". In Hass, U. & Schmitz, U. (eds.): Thematic Part: Lexikographie im Internet 2010. *Lexicographica. Internationales Jahrbuch für Lexikographie*, 26/2010. Berlin, 19-45.

Martín Mingorance, L. (1994). La lexicografía onomasiológica. In: Hernández, H. & Mederos, H. (Coord.). *Aspectos de lexicografía contemporánea*. Barcelona: Biblograf, 15-27.

Meliss, M. (2013a). Das zweisprachige Wörterbuch im bilateralen deutsch-spanischen Kontext. Alte und neue Wege. In Domínguez Vázquez, Mª J. (ed.), 61-87.

Meliss, M. (2013b). Online-Lexikographie im DaF-Bereich: Eine erste kritische Annäherung: Bestandsaufnahme – Nutzen – Perspektiven. In *Real Revista de Estudos Alemães, 4*, 176-199. http://real.fl.ul.pt/textos.page/pag/2. [11.04.2014]

Meliss, M. (2014a). (Vor)überlegungen zu einem zweisprachigen Produktionslernerwörterbuch für das Sprachenpaar DaF und ELE. In Reimann, D. (ed.). Kontrastive Linguistik und Fremdsprachendidaktik Iberoromanisch – Deutsch. Studien zu Morphosyntax, nonverbaler Kommunikation, Mediensprache, Lexikographie und Mehrsprachigkeitsdidaktik (Spanisch/Portugiesisch/Deutsch) (Reihe: Romanistische Fremdsprachenforschung und Unterrichtsentwicklung). Tübingen: Narr.

Meliss, M. (2014b). Das verbale Kombinationspotenzial in einsprachigen DaF-Lernerwörterbüchern: Kritische Bestandsaufnahme – Neue Anforderungen. In *ZDaF* (im Druck).

Meliss, M. (2014c). Propuestas para un diccionario conceptual bilingüe para ELE y DaF. In: Domínguez Vázquez, Mª J. et al. (eds.): (Coord.: Sánchez Palomino, Mª. D. & Domínguez Vázquez, Mª J. (vol. 2). (im Druck).

Meliss, M. & Sánchez Hernández, P. (2014). Theoretical and methodological foundations of the DICONALE project: a conceptual dictionary of German and Spanish. In Silvestre João Paulo et al. (eds.): *Dicionários que não existem*. Lissabon.

Model, B. (2010). Syntagmatik im zweisprachigen Wörterbuch. Berlin: de Gruyter.

Müller-Spitzer, C. & Engelberg, St. (2013). Dictionary Portals. In Rufus H. Gouws et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography.* Supplementary volume: Recent developments with special focus on computational lexicography. Berlin, New York: de Gruyter (im Druck).

Porto Dapena, J. A., Conde Noguerol, E., Córdoba Rodríguez, F. & Muriano Rodríguez, Mª M. (2008): Presentación del diccionario "Coruña" de la lengua española actual. En: Bernal, E. & DeCesaris (eds.): *Proceedings of the Xiii Euralex International Congress.* Barcelona: Documenta Universitaria, Série Activitats, 20.753-762.

Proost, K. (2007). *Conceptual structure in lexical items: The lexicalisation of communication concepts in English,* German and Dutch. Pragmatics & Beyond New Series; 168. Amsterdam/Philadelphia: Benjamins.

Reichmann, O. (1989). Das onomasiologische Wörterbuch: Ein Überblick. In Sterger H. & Wiegand, H.E. (eds.). *Wörterbücher: Ein Internationales Handbuch zur Lexikographie.* Berlin/New York: de Gruyter, 1057-1067.

Storrer, A. (2010). Deutsche Internet-Wörterbücher: Ein Überblick. In Gouws, R. H. et al. (ed.). *Lexicographica. Internationales Jahrbuch für Lexikographie 26.* Berlin: de Gruyter, 154-164.

Subirats, C. (2009). Spanish Framenet: A frame-semantic analysis of the Spanish lexicon. In Boas, H. (ed.). *Multilingual FrameNets in Computational Lexicography. Methods and Applications.* Berlin/New York: Mouton de Gruyter, 135-162.

Tarp, S. (2012): Online dictionaries: today and tomorrow. In Heid, U. (ed.): Thematic Part: Corpora and Lexicography. *Lexicographica (International Annual for Lexicography)* 28/2012. Berlin, 253-267.

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

Peter Meyer
Institut für Deutsche Sprache, Mannheim
meyer@ids-mannheim.de

Abstract

This paper reports on an ongoing lexicographical project that investigates Polish loanwords from German that were further borrowed into the East Slavic languages Russian, Ukrainian, and Belorussian. The results will be published as three separate dictionaries in the *Lehnwortportal Deutsch*, a freely available web portal for loanword dictionaries having German as their common source language. On the database level, the portal models lexicographical data as a cross-resource directed acyclic graph of relations between individual words, including German 'metalemmata' as normalized representations of diasystemic variants of German etyma. Amongst other things, this technology makes it possible to use the web portal as an 'inverted loanword dictionary' to find loanwords in different languages borrowed from the same German etymon. The different possible pathways of German loanwords that went through Polish into the East Slavic languages can be represented directly as paths in the graph. A dedicated in-house dictionary editing software system assists lexicographers in producing and keeping track of these paths even in complex cases where, e.g, only a derivative of a German loanword in Polish has been borrowed into Russian. The paper concludes with some remarks on the particularities of the dictionary/portal access structure needed for presenting and searching borrowing chains.

Keywords: online dictionary; graph databases; loanwords

1 Introduction

1.1 The Lehnwortportal Deutsch

The *Lehnwortportal Deutsch* (lwp.ids-mannheim.de) is a freely accessible online lexical information system developed at the Institute for German Language (IDS) that has been designed to provide unified access to a large number of both existing and newly produced XML-based dictionaries of German loanwords in other languages.[1] The modular architecture of the portal allows for easy integration of new resources of possibly very heterogeneous structure; each portal dictionary may have its own XML schema, as long as the underlying lexicographical information of the different constituent parts of an

[1] The web portal in its present form has been developed in a project funded by the Federal Government Commissioner for Culture and the Media upon a Decision of the German Bundestag.

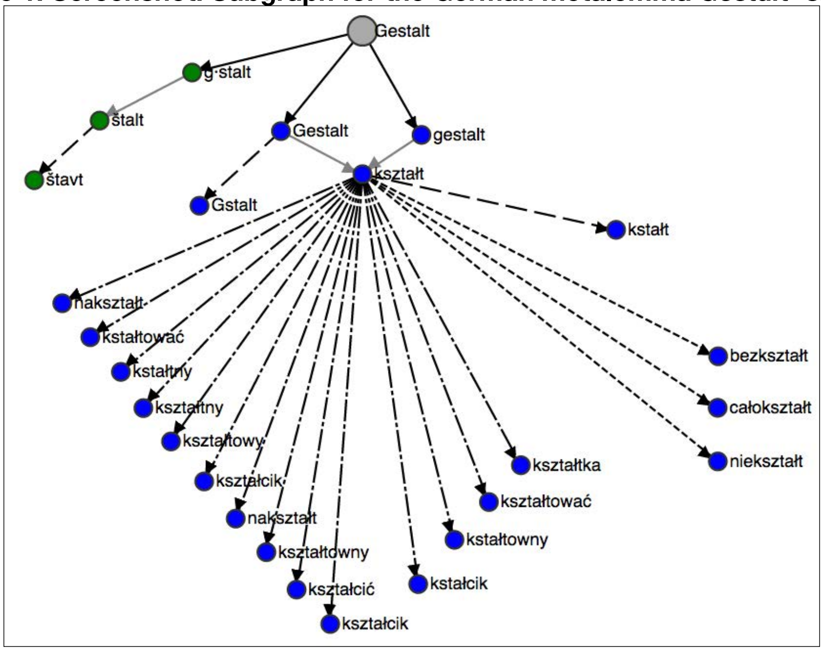
entry are unambiguously and explicitly encoded and separated in the markup, analogous to what is called the 'lexical view' in the TEI.dictionaries module, cf. Burnard, Bauman 2007, section 9.5 (online at http://en.guidelines.tei-c.org/html/DI.html#DIMVLV [04/11/2014]).

Apart from conventional access to the individual dictionaries, the portal offers complex cross-dictionary search functionality; in particular, it can be used as an 'inverted loanword dictionary' (Engelberg 2010) to trace the way of German words into different recipient languages, comparable to the manually compiled dictionary of Dutch loanwords in the world's languages by van der Sijs (2010). As any German etymon may appear in a variety of orthographical, phonetic/phonological and diasystemic variants in different entries within and across loanword dictionaries, these different forms are mapped in manual lexicographical work to etymologically corresponding 'normalized' word forms, wherever possible contemporary Standard German words. This is accomplished at the IDS with the help of an in-house software tool during the integration of a loanword dictionary into the web portal. These normalized entries, henceforth *metalemmata*, are used as headwords of the inverted loanword dictionary.

1.2 Graph-based Data Modeling

The XML-based representation of entries in the individual component dictionaries mainly serves as input for XSLT transformations that produce a fairly conventional, dictionary-specific HTML-based online presentation of the entries. For advanced search functionalities, however, a relational database is used that represents lexicographical information as a cross-resource network (a *directed acyclic graph*) of relations between words that are, as we say throughout this paper, 'recorded' in the individual dictionaries. These recorded words include metalemmata, etyma and loanwords alongside their variant forms, derivatives etc. Interactive visualizations of parts of this graph are available online; figure 1 below shows the subgraph for the German metalemma *Gestalt* 'shape'. Differently colored discs correspond to words recorded in different dictionaries (vertices/nodes in the graph); different kinds of relations (arcs/directed edges) are symbolized by different types of arrays between two discs. In the example we see that the 'normalized' contemporary German metalemma *Gestalt* 'corresponds to' [=dark solid arrow] a New High German etymon *Gestalt* and a Middle High German etymon *gestalt* as recorded in the portal's Polish loanword dictionary (color tag: dark blue) and to a Middle High German / Bavarian etymon *g·stalt* as recorded in the Slovene dictionary (color tag: green). We further see that, e.g., the etymon *gestalt* 'has been borrowed into' Polish [=grey solid arrow] as *kształt* which 'has a variant phonetic' form *kstalt* [=black long-dashed arrow] and from which (amongst other words) the verb *kształcić* 'has been derived' [=dashed-dotted arrow]. The relationships between words recorded in the same loanword dictionary entry are programmatically extracted from the XML source of the entry on a per-dictionary basis, making use of the fact that different kinds of relations correspond to different structural configurations in the entry document that depend on the XML schema of the dictionary and can be described using XPath expressions. Every word in the graph has a set of attributes (diasystemic, grammatical, semantic information) obtained by encoding the appropriate pieces of information as contained in the entries in a portal-wide unified data format. For more details, cf. Meyer (2013b; to appear).

Figure 1: Screenshot: Subgraph for the German metalemma *Gestalt* 'shape'



(http://lwp.ids-mannheim.de/art/meta/lemma/Gestalt).

1.3 Tracing the Way of Polish Loanwords from German into East Slavic

In a joint project funded by the German Research Foundation the Institute of Slavic Studies at the University of Oldenburg and the Institute of German Language (IDS, Mannheim) are currently developing dictionaries of German loanwords in the three East Slavic languages Russian, Ukrainian, and Belorussian that were mediated through Polish words recorded in the portal's dictionary on German loans in Standard Polish (previously published as a standalone resource: de Vincenz, Hentschel 2010). This endeavor draws on a rich Slavic tradition of historical lexicography; a wealth of (partially unpublished) dictionary material (starting with a basis of 15 historical and contemporary monolingual dictionaries of Russian, Ukrainian, and Belorussian) will be excerpted and analyzed both in Oldenburg and at the editorial offices of those dictionaries that are still work in progress, while the portal integration of the resulting dictionaries with an estimated total of 1900 new entries will be carried out in Mannheim.

The present paper focuses on data modeling issues and the technical and procedural specifics of dealing with (arbitrarily long) borrowing chains in the context of this project.

2 Modeling and Editing Borrowing Chains in the Portal's Graph Database

2.1 Data Modeling Aspects: Borrowing Chains as Paths in the Graph

Borrowing chains are the premier *raison d'être* for the graph-based data modeling in the *Lehnwortportal*. Figure 2 (below) shows, if only in a highly schematic fashion, the sample case of German *Drucker* 'printer' that has entered the East Slavic languages mostly through Polish. The different pathways obviously form a small directed graph that can be added more or less directly as a new subgraph to the portal data graph. The dashed arrows indicate less likely borrowing pathways; correspondingly, edges in the portal graph may be assigned weights to indicate likelihood of a borrowing relation and to calculate rankings of search results. Note that there are three different paths in this subgraph all leading from the German etymon *Drucker* to the Russian loanword *drukar'*.

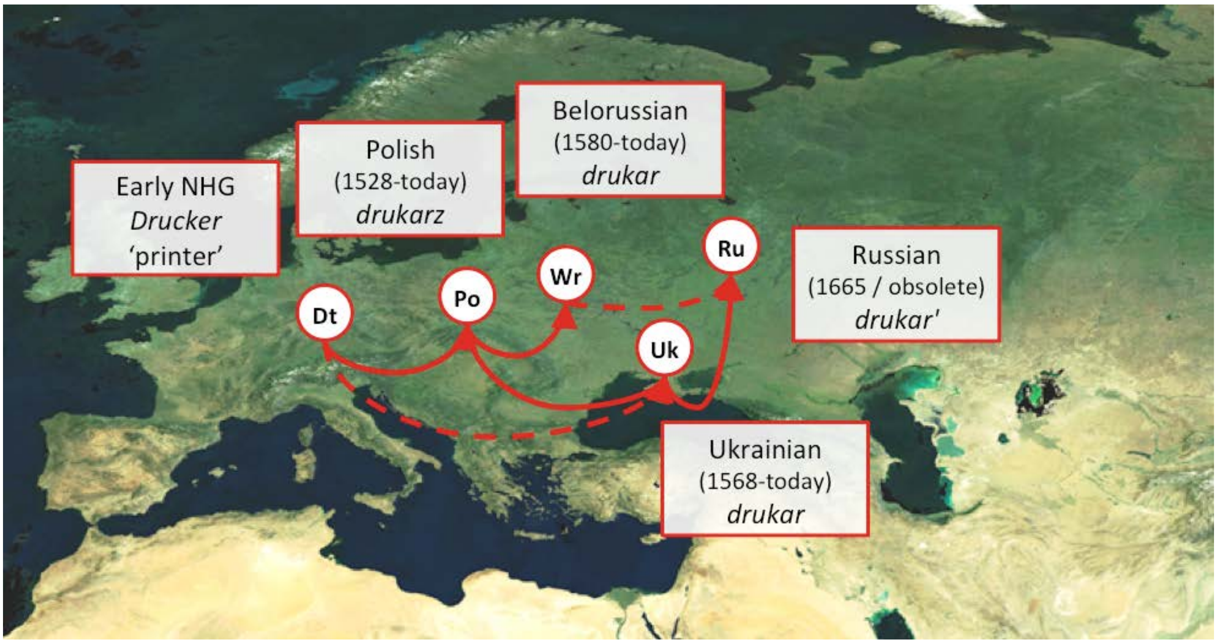


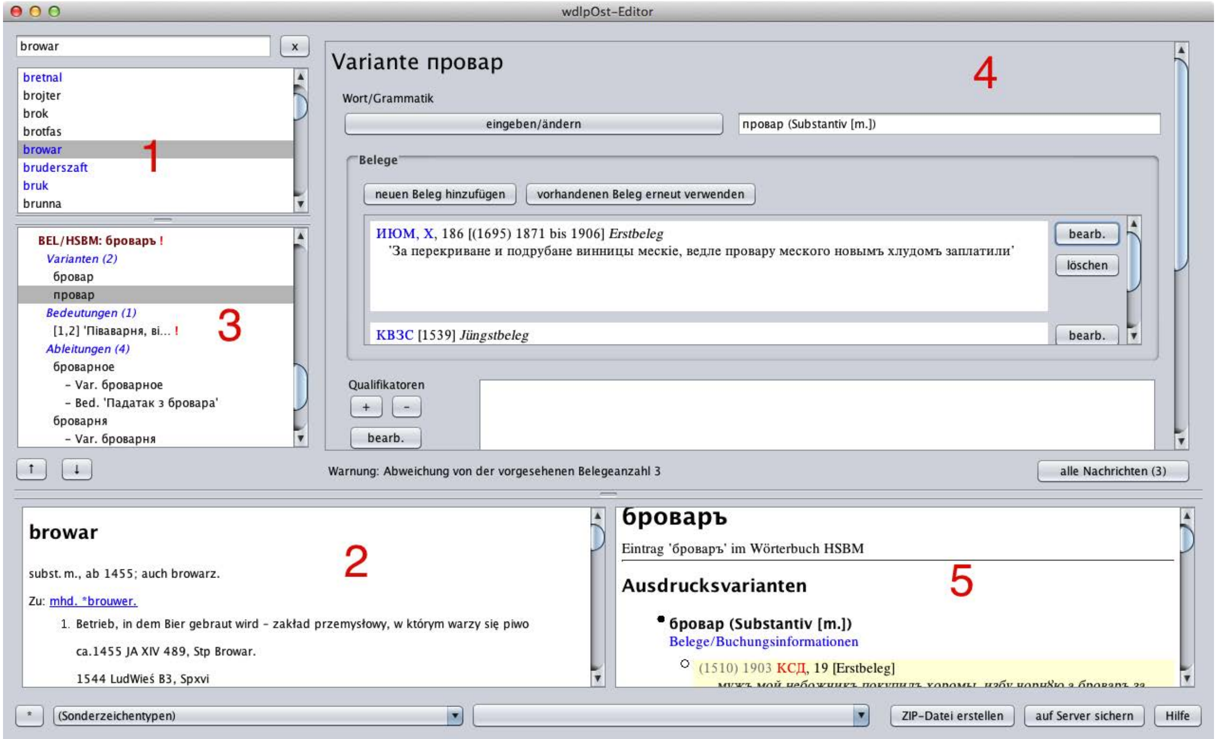
Figure 2: Possible borrowing paths of German Drucker 'printer' into the East Slavic languages.

2.2 Some Technical Aspects of the Lexicographical Process

The graph data layer atop the XML resources of individual loanword dictionaries requires a highly specialized lexicographical process of its own (cf. Meyer, to appear). The graph is not a self-contained resource; instead, it must be constructed anew from the individual dictionaries and portal-specific cross-reference information as soon as any of the portal's data sources changes. Tracing borrowing chains complicates the picture considerably. For the project presented here, a complex in-house desktop dictionary editing software is being developed which allows lexicographers to collaboratively compile and edit excerpts *using the lemmata (and other recorded words and meaning definitions) of the portal's Polish loanword dictionary* (de Vincenz, Hentschel 2010) *as a common frame of reference*. Figure 3 (below) shows a screenshot of a preliminary version of the editor used for excerpting. The working lexicographer selects a Polish loanword from (de Vincenz, Hentschel 2010) such as *browar* 'brewer; brewery' from Middle High German *brouwer* 'brewer' (1). A preview of this entry is displayed in the main window (2). All hitherto produced excerpts of existing dictionary entries on East Slavic borrowings from the selected Polish word are listed as a tree structure in the editor (3); for each such entry, the tree shows all recorded (phonetic and diasystemic) variants, meanings, derivatives (with their own range of variants and meanings) and competing near-synonyms that have been input so far. Clicking on a tree item (here, on the variant *provar* of the entry *brovar* in the multivolume Belorussian Historical Dictionary *Histaryčny sloŭnik belaruskaj movy*) opens an input panel (4) for all pertinent lexicographical information, including an arbitrary number of records and quotations. A preview of the current state of the whole excerpt is also available (5). There is a separate input panel, not shown in figure 3, for editing cross-dictionary information on the possibly multiple borrowing paths within the East Slavic languages. Often Polish loanwords from German have formed compounds and derivatives; it is well possible that only one of these derived forms, but not the 'original' loanword, has been passed on into an East Slavic language. The editing software also offers convenient input options for such cases. There are additional tools for compiling the entries of the three new East Slavic loanword dictionaries from the excerpts.

There are many reasons why an off-the-shelf software solution would not have been suitable for the lexicographical tasks of the project. To begin with, it would have been next to impossible to customize a commercial dictionary editing application in order to incorporate cross-referencing functionality to an existing dictionary. In this particular case, cross-references are needed not only to whole entries of the Polish dictionary (de Vincenz, Hentschel 2010), but also to derivatives and compounds recorded in these entries, and, most important, to the different word senses given in the entries since they will serve as a *tertium comparationis* for word sense distinctions in the East Slavic loanwords. It would have been possible to customize a professional XML editor by implementing some kind of cross-referencing plugin. However, there is another layer in the editing process that cannot easily be managed in XML: After compiling excerpts of entries on a German loanword in, say, Ukrainian, in a number of Ukrainian loanword dictionaries, these excerpts have to be merged in a rather complex way to produce a new entry in the Ukrainian loanword dictionary of the portal. The excerpted loanword dictionaries (which may or may not cover different periods of the language) will have differing lemmatizations, list different variants of the word, use incompatible word sense distinctions and so on. On the other hand, there will usually be a lot of duplicate information. As a consequence, the amalgamation process of creating entries in the three new East Slavic portal dictionaries is far from trivial; doing this by cutting XML fragments from the excerpts and pasting them into the XML structure of the newly created entries would be an excessively tedious, error-prone and confusing task, even more so since word sense distinctions in parallel entries in the three dictionaries should be made in as uniform a manner as possible, based on the distinctions in the entries of (de Vincenz, Hentschel 2010). It is not a realistic goal to develop software tools for these tasks as simple XML editor plugins; even the very idea of using XML as the basic frame of reference is problematic in such a complex cross-resource editing context.

In fact, the software developed at the IDS is not directly XML-based, but uses a straightforward object-oriented data model for both the excerpts and the newly produced entries. This greatly simplifies the underlying cross-referencing and the implementation of tools for merging and validating lexicographical data from a large number of resources. The software produces XML serializations of the data that can be used both to construct HTML views of the data and to define the directed graph of the portal.



bretnal
brojter
brok
brotfas
browar
bruderszaft
bruk
brunna

BEL/HSBM: броваръ !
Varianten (2)
бровар
провар
Bedeutungen (1)
[1,2] 'Піваварня, ві... !
Ableitungen (4)
броварное
– Var. броварное
– Bed. 'Падатак з бровара'
броварня
– Var. броварня

Variante провар

Wort/Grammatik: провар (Substantiv [m.])

Belege

ИЮМ, X, 186 [(1695) 1871 bis 1906] *Erstbeleg*
'За перекриване и подрубане винницы мескіе, ведле провару меского новымъ хлудомъ заплатили'

КВЗС [1539] *Jüngstbeleg*

Warnung: Abweichung von der vorgesehenen Belegeanzahl 3

browar

subst. m., ab 1455; auch browarz.

Zu: mhd. *brouwer.

1. Betrieb, in dem Bier gebraut wird – zakład przemysłowy, w którym warzy się piwo

ca.1455 JA XIV 489, Stp Browar.

1544 LudWieś B3, Spxvi

броваръ

Eintrag 'броваръ' im Wörterbuch HSBM

Ausdrucksvarianten

- бровар (Substantiv [m.])
  Belege/Buchungsinformationen
  ○ (1510) 1903 КСД, 19 [Erstbeleg]
  мужъ мой небожчикъ покупилъ хоромы, избу норнȣю а броварь за

Figure 3: Screenshot: Preliminary user interface of the dictionary writing software.

2.3 General Lexicographical Issues Concerning Borrowing Chains

As explained above, the project presented in this paper strives to compile new loanword dictionaries that directly reference a Polish loanword dictionary already integrated into the portal. As the *Lehnwortportal* has been designed with a focus on leveraging existing resources, we expect to see other cases in the future where information from multiple existing loanword dictionaries is combined to reconstruct borrowing chains. Here, the data graph is an ideal means of abstracting from micro- and mediostructural idiosyncrasies of the dictionaries involved. Dutch may serve as a good example since it has served as a 'hub' that mediated German lexis into many languages, in particular those of colonized countries. Thus, we might try to combine information on Dutch loans from German – as represented in a traditional loanword dictionary (e.g., van der Sijs 2005) – with information on Dutch loanwords in other languages – as given in (van der Sijs 2010) – to (re)construct borrowing chains from German via Dutch into other languages. In these cases, an 'intermediate' Dutch loan corresponds to *two* connected vertices in the graph as it appears in two independent lexicographical resources – both as a loanword from German and as an etymon for Dutch loans in other languages. The etymological identification of these two 'instances' of the intermediate loanword is part of the lexicographical process to be carried out for the portal. There are many technical and lexicographical issues arising in borderline cases, e.g., when borrowing chains are directly specified in a loanword dictionary entry (such as "from German *Drucker* via Polish *drukarz*"). Here, multiple cross-resource etymological identifications with words in other resources might become necessary, even with the possibility of conflicting information, e.g. if another loanword dictionary of the portal gives a different German etymology for an intermediate loan.

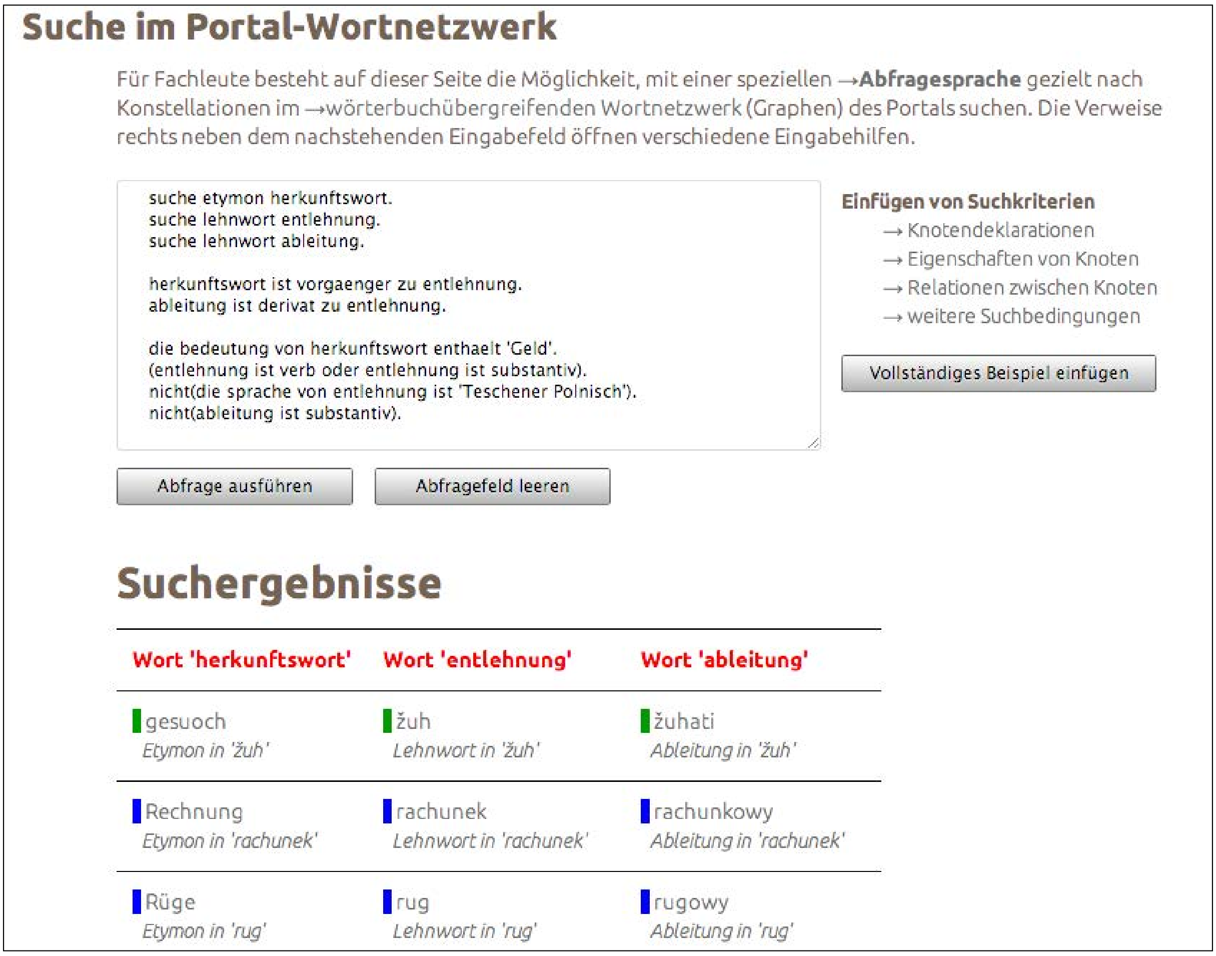
3 Access Structures for Borrowing Chains in the Portal

3.1 Online Entry Presentation

The information on borrowing chains is, in many cases, not present within the confines of a single loanword dictionary entry, but is instead distributed between different resources. The online presentation of individual dictionary entries should nevertheless make this information visible in all of the individual dictionary entries involved. At present, loanword dictionary entries and entries of the 'inverted loanword dictionary' of German metalemmata systematically cross-reference each other in the *Lehnwortportal*; in the case of 'indirect' loanwords from German, the web application will use the data graph to add another layer of information, viz. on the 'intermediate' or 'terminal' loanwords, to the existing entries in the Polish and East Slavic dictionaries. These additions only concern the presentation layer; the underlying entries remain unaltered. A special feature of the presently compiled East Slavic dictionaries will be the presence of cross-dictionary commentaries (including schematic visualizations of borrowing pathways) on all entries that refer to the same Polish loanword, since often German loanwords in an East Slavic language could also have been borrowed directly from German or were mediated by another East Slavic language (cf. figure 2 above).

3.2 Advanced Search Options

It is highly desirable that portal users can include search criteria concerning borrowing chains into their advanced queries such that, e.g., German loanwords in language X that were possibly mediated through language Y can be found. The *Lehnwortportal* offers fairly advanced and granular search options (cf. Meyer 2013a) that allow the inclusion of complex criteria concerning both German etyma (including metalemmata) and loanwords. With the inclusion of the Slavic dictionaries, search criteria will also be attributable to 'intermediate' loans in a borrowing chain. Search results will be ranked according to the weight of the edges of the graph path. Borrowing paths will be specifiable through a planned extension of the declarative domain-specific query language that is currently available for advanced portal users (figure 4); cf. (Wood 2012) for a general overview on graph database query languages and (Meyer 2013a) for more information on the portal's query language. Even a graphical search through an interactive visual query language for graph databases is conceivable (cf. Blau et al. 2002) and would allow users to literally draw the borrowing paths they are looking for.



Suche im Portal-Wortnetzwerk

Für Fachleute besteht auf dieser Seite die Möglichkeit, mit einer speziellen →**Abfragesprache** gezielt nach Konstellationen im →wörterbuchübergreifenden Wortnetzwerk (Graphen) des Portals suchen. Die Verweise rechts neben dem nachstehenden Eingabefeld öffnen verschiedene Eingabehilfen.

```
suche etymon herkunftswort.
suche lehnwort entlehnung.
suche lehnwort ableitung.

herkunftswort ist vorgaenger zu entlehnung.
ableitung ist derivat zu entlehnung.

die bedeutung von herkunftswort enthaelt 'Geld'.
(entlehnung ist verb oder entlehnung ist substantiv).
nicht(die sprache von entlehnung ist 'Teschener Polnisch').
nicht(ableitung ist substantiv).
```

Einfügen von Suchkriterien

→ Knotendeklarationen
→ Eigenschaften von Knoten
→ Relationen zwischen Knoten
→ weitere Suchbedingungen

Suchergebnisse

Wort 'herkunftswort'	Wort 'entlehnung'	Wort 'ableitung'
🟩 gesuoch *Etymon in 'žuh'*	🟩 žuh *Lehnwort in 'žuh'*	🟩 žuhati *Ableitung in 'žuh'*
🟦 Rechnung *Etymon in 'rachunek'*	🟦 rachunek *Lehnwort in 'rachunek'*	🟦 rachunkowy *Ableitung in 'rachunek'*
🟦 Rüge *Etymon in 'rug'*	🟦 rug *Lehnwort in 'rug'*	🟦 rugowy *Ableitung in 'rug'*

Figure 4: Screenshot: Example search using the portal's graph query language (http://lwp.ids-mannheim.de/search/prof).

For illustration purposes, here is a rough preview of how a simple query involving borrowing chains will look like in the portal's query language. Note that the original query language has a German-like context-free grammar; here, we present a corresponding English-like version. The query reads: "Find all Ukrainian or Belorussian words (including variants, derivatives etc.) in the database that have been borrowed through Polish from a German noun ending in *ung*." The query uses graph-theoretical terms where appropriate; thus, any loanword borrowed from German is represented by a node in the portal's directed graph that is a *descendant* of the node corresponding to the German etymon. The results of the query are ordered triples (germanWord, polishWord, eastSlavicWord) of those words recorded in entries of the portal's dictionaries that comply with all of the constraints specified in the query.

```
find etymon germanWord.
find loanword polishWord.
find loanword eastSlavicWord.
the language of germanWord is German.
the language of polishWord is Polish.
(the language of eastSlavicWord is Ukrainian OR the language of eastSlavicWord is Belorussian).

germanWord is a noun.
germanWord ends in 'ung'.
polishWord is descendant of germanWord.
eastSlavicWord is descendant of polishWord.
```

4 References

Blau, H., Immerman, N. & Jensen, D. (2002). *A Visual Language for Querying and Updating Graphs.* Technical Report 2002-037. University of Massachusetts, Amherst.

Burnard, L., Bauman, S. (eds.) (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Charlottesville, Virginia: TEI Consortium. Online: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html [04/11/2014].

de Vincenz, A., Hentschel, G. (2010). Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts. (= Studia slavica Oldenburgensia, vol. 20). Oldenburg: BIS-Verlag. Online edition: http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp [04/11/2014].

Engelberg, S. (2010). An inverted loanword dictionary of German loanwords in the languages of the South Pacific. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6-10 July 2010).* Ljouwert (Leeuwarden): Fryske Akademy, pp. 639-647.

Meyer, P. (to appear). Von XML zum DAG: Der lexikographische Prozess bei der Erstellung eines graphenbasierten Wörterbuchportals. In F. Mollica, M. Nied, M.J. Domínquez Vazquez (eds.) *Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik.* (to appear in: Lexicographica, Series Maior).

Meyer, P. (2013a). Advanced graph-based searches in an Internet dictionary portal. In I. Kosem, J. Kallas, P. Gantar, P. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 488-502. Accessed at: http://eki.ee/elex2013/proceedings/eLex2013_34_Meyer.pdf [04/11/2014].

Meyer, P. (2013b). Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation. In S. Kempgen, N. Franz, M. Jakiša, M. Wingender (eds.) *Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013.* München: Otto Sagner, pp. 233-242. (=Die Welt der Slaven. Sammelbände, vol. 50).

Meyer, P., Engelberg, S. (2011). Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In H. Hedeland, Th. Schmidt, K. Wörner (eds.) *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011.* Hamburg: Universität Hamburg, pp. 169-174 (=Arbeiten zur Mehrsprachigkeit/Working Papers in Multilingualism, Series B, No. 96).

van der Sijs, N. (2005). *Groot leenwoordenboek.* Utrecht: Van Dale Lexicografie.

van der Sijs, N. (2010). *Nederlandse woorden wereldwijd.* Den Haag: SDU Uitgever.

Wood, P. T. (2012). Query Languages for Graph Databases. In *SIGMOD RECORD*, 41(1) (March), pp. 50-60.

At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project)

Pavla Kochová, Zdeňka Opavská, Martina Holcová Habrová
Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i.
kochova@ujc.cas.cz, opavska@ujc.cas.cz, holcova@ujc.cas.cz

Abstract

The aim of the article is to present the new lexicographic project that is being implemented at the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i. Since 2012, its Department of Contemporary Lexicology and Lexicography has worked on the creation of a new medium-sized monolingual dictionary of Czech with the working title *Akademický slovník současné češtiny* (The Academic Dictionary of Contemporary Czech). With its size and method of treatment, the dictionary ranks among academic dictionaries, i.e. dictionaries with an elaborated, standardised and structured explanation of the meaning of lexical units, with an adequately rich exemplification documenting the typical use of lexical units, with a sufficiently elaborated description of the basic semantic relations, mainly synonymy and antonymy, with the appropriate description of the grammatical properties of lexical units and with usage labels (a description of the stylistic, temporal, spatial, frequency and pragmatic markedness) of lexical units.

Keywords: lexicography; monolingual dictionary; macrostructure of a dictionary; microstructure of an entry

1 Introduction

Akademický slovník současné češtiny (hereinafter as the *ASSČ*) builds on the tradition of the general monolingual dictionaries of Czech that emerged at the Institute for the Czech Language in the 20th century.[1] More than forty years have passed since the publication of a dictionary of a larger size, i.e. the

[1] This tradition has developed from the largest *Příruční slovník jazyka českého* (Reference Dictionary of the Czech Language; *PSJČ*, 1935–1957), through the medium-sized *Slovník spisovného jazyka českého* (Dictionary of the Standard Czech Language; *SSJČ*, 1960–1971) to the one-volume *Slovník spisovné češtiny pro školu a veřejnost* (Dictionary of Standard Czech for Schools and the Public; *SSČ*, 1st edition in 1978; 2nd, revised edition in 1994; 3rd, revised edition in 2003). The *PSJČ* is a scientific descriptive dictionary of a large size (ca 250,000 entries); it describes Czech vocabulary since 1880; it does not use run-on entries; examples are provided by quotations. The *SSJČ* is a medium-sized dictionary (192,908 entries); it captures the literary lexical standard of the time, but the range of the word list exceeds it (by including obsolete, infrequent words etc.); it describes contemporary Czech vocabulary (approximately from the 1930s, selectively from 1880); the SSJČ uses run-on entries; exemplification is mainly based on the minimal typical contexts. The *SSČ* is a smaller-size dictionary (2nd ed.: 45,366 entries) focusing on the widest range of users; it describes the central vocabulary of contemporary Czech (mainly from 1945) with an overlap to variously marked words; it has a normative character; exemplification is very limited and is based on the minimal typical contexts. On the history and characteristics of the modern Czech lexicography, see mainly the detailed study by Hladká (2007).

Slovník spisovného jazyka českého (since the publication of the first volume it has even been fifty years), which is very long considering vocabulary dynamics, linguistic methodology,[2] in terms of the platform for the creation of a dictionary as well as the medium for its publication.

2 The Basic Characteristics of the ASSČ

The *ASSČ* is a *medium-sized* dictionary,[3] with the expected number of 120–150 thousand lexical units. Its *aim* is to capture widespread contemporary Czech vocabulary used in public official and semi-official communication as well as in everyday (i.e. non-public, unofficial) communication. A natural part of the lexis described are terminological expressions, but not highly specialised terms. To a limited extent, the dictionary presents units utilised in professional and interest-group communication, namely if their use has been extended beyond their professional, interest milieu. Dialectal expressions have been included if they are common in a wider area and are used especially in oral communication or in literature. The expected *user* of the dictionary is a secondary-school educated native speaker; nevertheless, also those interested in Czech as a foreign language are marginally taken into account (since Czech is a language of a small nation, specialised monolingual dictionaries of a larger size for learners are not created). The dictionary being prepared will be continuously *published* on the Internet (on the website of the Institute of the Czech Language). After the work on the dictionary has been completed, it will be possible to publish the work as a whole in a book form.

3 The Dictionary Development Method: Selected Aspects

The essential *material basis* is the synchronic corpus of written texts SYN of the Institute of the Czech National Corpus of a size of 2.2 milliard words. Other material resources are the electronic archives of the company Newton Media, a. s. (the archives of both nationwide and regional printed periodicals and transcripts of current affairs television and radio programmes), the internet and the databases of the Institute of the Czech Language.[4]

[2] In connection with lexicography and lexicology, it is mainly the creation and development of computational and corpus linguistics and corpus lexicography (language corpora, excerpt databases, electronic archives, special software tools, eg. for an analysis of collocations the Word Sketch Engine – Kilgarriff et al. 2004). Cf. Čermák, Blatná 1995; Čermák 2010.

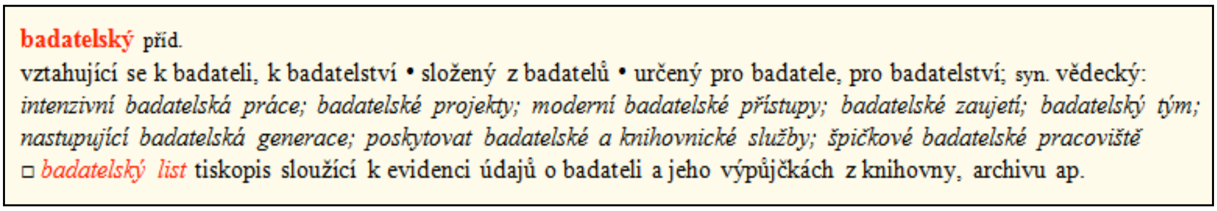
[3] It should be emphasised that the dictionary being developed is not a lexical database. The relation between a lexical database and a monolingual dictionary is understood in accordance with Hanks (2010: 581): "A lexical database is a fundamental background resource for use in the creation of many important linguistic artefacts – dictionaries, course books, computer programs for natural language processing among them. A great monolingual dictionary has a different function: it brings together speakers of a language, it has a socially integrative function, making explicit the basis of words and meanings and usage, which all uses of the language rely on."

[4] An excerpt database of neologisms (focused on new lexical phenomena), a database of specialised vocabulary, the Pralex – preparatory lexical database and the Modern Czech lexical archives created in 1911–1991.

4 The Macrostructure of the Dictionary

The word list of the *ASSČ* is built using a different lexicographic technique than before.[5] It draws on a set of three balanced corpora, SYN 2000, SYN 2005 and SYN 2010. The entries are selected from an automatically generated word list mainly based on the frequency criterion and the criterion of the commonness of their usage (i.e. only widespread lexical units are included; specialised terms, professional and slang expressions etc. are included only selectively). On the other hand, the word list has been expanded on the basis of word-formation relations (members of word-formation groups) and on the basis of co-hyponymic and other relations (members of lexical-semantic classes have been added).

Unlike in earlier dictionaries (*SSJČ, SSČ*), *derivatives* (relational adjectives, adverbs, names of properties), which used to be added to the lemmas as run-on entries, are listed as separate entries now. The new method (including the explanation of the meaning and exemplification) makes it possible to give an adequate lexicographic description, which however requires a detailed, often demanding analysis, cf. the explanation of the meaning in the entry for the relational adjectives (*badatel* n. "researcher" → *badatelský* adj. "vztahující se k badateli, k badatelství • složený z badatelů • určený pro badatele, pro badatelství" =pertaining to researchers, researching • consisting of researchers • intended for researchers, for researching), see figure 1.



badatelský příd.
vztahující se k badateli, k badatelství • složený z badatelů • určený pro badatele, pro badatelství; syn. vědecký: *intenzivní badatelská práce; badatelské projekty; moderní badatelské přístupy; badatelské zaujetí; badatelský tým; nastupující badatelská generace; poskytovat badatelské a knihovnické služby; špičkové badatelské pracoviště*
□ *badatelský list* tiskopis sloužící k evidenci údajů o badateli a jeho výpůjčkách z knihovny, archivu ap.

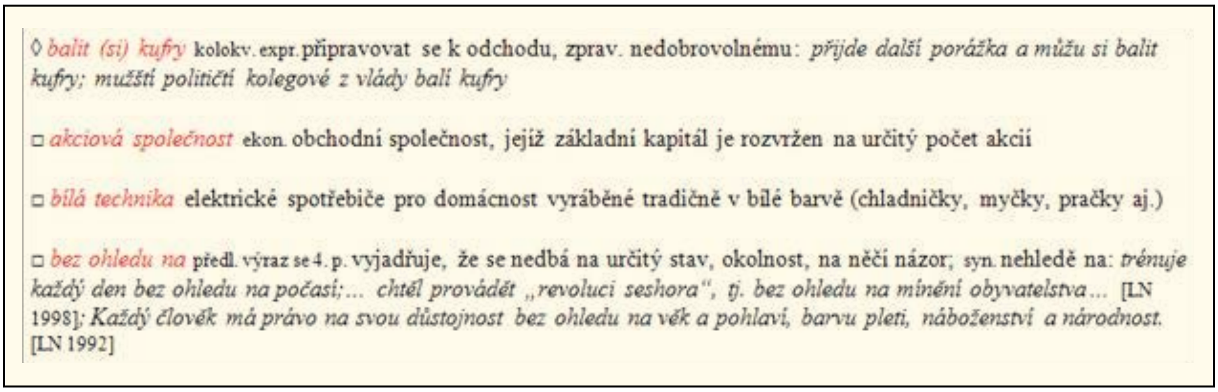
Figure 1: Entry *badatelský*.

Only some lexical types are treated as *run-on entries*. These include words derived by adding feminine suffixes (*herečka* ← *herec* "actress ← actor"), diminutives (*pejsek* ← *pes* "doggie ← dog") and frequentative verbs (*balívat* ← *balit* "to pack"), where the semantic structure of the derivative does not differ from the lemma. Cf. the lemma *bouček* "small beech" treated as a run-on entry in the entry buk "beech" (see figure 5).

In the ASSČ, greater autonomy has been given also to *multi-word lexical units*. The treatment distinguishes between: phraseological units (*balit si kufry* "to pack one's bags") and non-phraseological units (terminological – *akciová společnost* "joint-stock company"; non-terminological – *bílá technika* "white goods"; multi-word grammatical expressions – *bez ohledu na* "regardless of" preposition etc.). In the dictionary, these are listed in an one-word lemma entry, but it is taken into account that they are

[5] The word list of the PSJČ relied on a comprehensive and sophisticated excerption of 5 million excerpts. The word list of the SSJČ built on the word list of the preceding dictionary, i.e. PSJČ, and its own excerption. Similarly, the latest of these modern dictionaries, the one-volume SSČ, proceeded from the word list of the SSJČ and its own excerption.

independent formal-semantic lexical units; therefore, the meaning explanation and exemplification are provided for a large part of them (always for phrasemes; for non-phraseological units, the explanation is given where the meaning is not compositional). The independence of multi-word lexical units is indicated also by the method of their presentation in the entry (a highlighted multi-word lemma, labelling with special symbols), see figure 2 and figure 3.



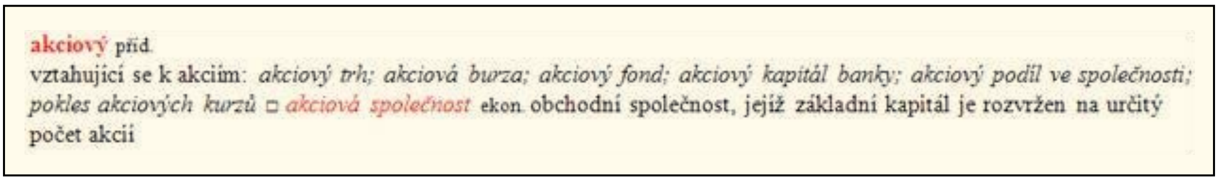
◊ *balit (si) kufry* kolokv. expr. připravovat se k odchodu, zprav. nedobrovolnému: *přijde další porážka a můžu si balit kufry; mužští političtí kolegové z vlády balí kufry*

□ *akciová společnost* ekon. obchodní společnost, jejíž základní kapitál je rozvržen na určitý počet akcií

□ *bílá technika* elektrické spotřebiče pro domácnost vyráběné tradičně v bílé barvě (chladničky, myčky, pračky aj.)

□ *bez ohledu na* předl. výraz se 4. p. vyjadřuje, že se nedbá na určitý stav, okolnost, na něčí názor; syn. nehledě na: *trénuje každý den bez ohledu na počasí;... chtěl provádět „revoluci seshora", tj. bez ohledu na mínění obyvatelstva...* [LN 1998]; *Každý člověk má právo na svou důstojnost bez ohledu na věk a pohlaví, barvu pleti, náboženství a národnost.* [LN 1992]

Figure 2: Multi-word lexical units balit (si) kufry, akciová společnost, bílá technika and bez ohledu na.



akciový příd.
vztahující se k akciím: *akciový trh; akciová burza; akciový fond; akciový kapitál banky; akciový podíl ve společnosti; pokles akciových kurzů* □ *akciová společnost* ekon. obchodní společnost, jejíž základní kapitál je rozvržen na určitý počet akcií

Figure 3: Multi-word lexical unit *akciová společnost* listed in the one-word entry *akciový.*

5 The Microstructure of an Entry

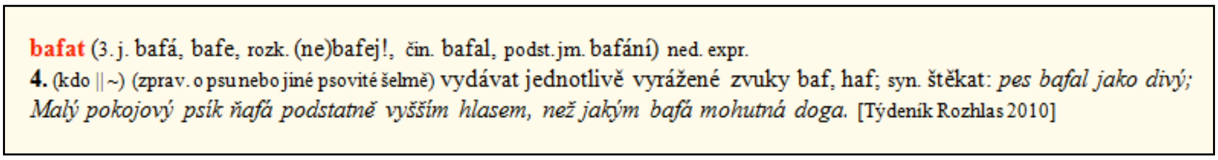
An entry in the *ASSČ* consists of the following parts: the lemma (including variant forms), information on homonyms, pronunciation, the etymology of the lexical unit, grammatical information (word class, morphology, valency), the usage label, the explanation of the meaning (including synonyms and antonyms), exemplification, notes (e.g. encyclopaedic information, further etymological information)[6] and cross-reference to (semantically, grammatically) related entries.

In the *ASSČ*, *grammatical information* (see figure 4) is treated more comprehensively than in previous monolingual dictionaries.[7] The morphological data in the ASSČ entries include mainly doublet forms and forms where the users may hesitate. The information on valency is systematically given for verbs

[6] On the usage of notes, see e.g. the Oxford Dictionary of English (Soanes, Stevenson 2005), in Czech lexicography the neological dictionaries (Martincová et al. 1998, 2004).

[7] In some respects, this transcends the genre of a general monolingual dictionary; on the other hand, it accommodates the users, who expect this type of information in a dictionary.

(both right and left valency), selectively also for nouns and adjectives. The valency information is semantically specified, if necessary, in the explanation of the meaning, or in the examples.



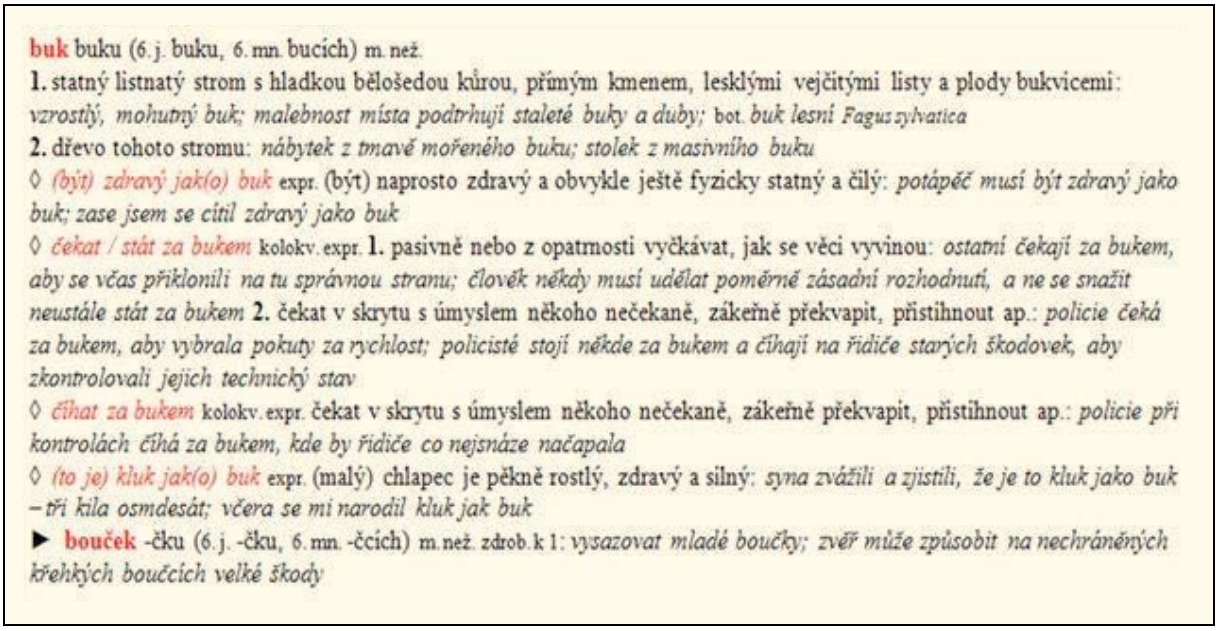
bafat (3. j. bafá, bafe, rozk. (ne)bafej!, čin. bafal, podst. jm. bafání) ned. expr.
4. (kdo || ~) (zprav. o psu nebo jiné psovité šelmě) vydávat jednotlivě vyrážené zvuky baf, haf; syn. štěkat: *pes bafal jako divý; Malý pokojový psík ňafá podstatně vyšším hlasem, než jakým bafá mohutná doga.* [Týdeník Rozhlas 2010]

Figure 4: Sense 4 of the lemma *bafat*.

When giving the *lexical meaning* of the entries, the *ASSČ* proceeds from the basic concept of determining the species classification – genus proximum – and differential semantic elements – differentia specifica (bearing in mind that besides notional elements also pragmatic elements need to be described). A part of the lexical meaning, however, are also those semantic elements that cannot be considered as necessary distinctive features but which mirror the complex of information on the denoted extra-linguistic reality that the language users have on the level of common knowledge. To a certain extent, the explanation of the meaning may hence contain "encyclopaedic" data (especially those that are objectively reflected in the word-formation structure of a word, in set similes and other phrasemes and in semantically derived meanings, on the basis of a metaphor).[8] In order to eliminate circularity, the explanation by means of synonyms is limited to a minimum, only to some slang, expressive or dialectal words.

The *exemplification part* of an entry includes both typical examples illustrating typical usage and extended examples that show less common, unusual and sometimes even authorial use of the word (mainly in the case of less frequent words and those belonging to peripheral areas of vocabulary). In addition, the examples are to illustrate grammatical information (especially on valency) and demonstrate (semantic) collocability. The exemplification may further contain those connotations which are not included in the explanation of the meaning but which the user (proto)typically connects with the unit concerned.

[8] Dolník (2012: 45); cf. Buzássyová, Jarošová (2006: 27–28).



buk buku (6. j. buku, 6. mn. bucích) m. než.
1. statný listnatý strom s hladkou bělošedou kůrou, přímým kmenem, lesklými vejčitými listy a plody bukvicemi: *vzrostlý, mohutný buk; malebnost místa podtrhují staleté buky a duby;* bot. *buk lesní Fagus sylvatica*
2. dřevo tohoto stromu: *nábytek z tmavě mořeného buku; stolek z masivního buku*
◊ *(být) zdravý jak(o) buk* expr. (být) naprosto zdravý a obvykle ještě fyzicky statný a čilý: *potápěč musí být zdravý jako buk; zase jsem se cítil zdravý jako buk*
◊ *čekat / stát za bukem* kolokv. expr. **1.** pasivně nebo z opatrnosti vyčkávat, jak se věci vyvinou: *ostatní čekají za bukem, aby se včas přiklonili na tu správnou stranu; člověk někdy musí udělat poměrně zásadní rozhodnutí, a ne se snažit neustále stát za bukem* **2.** čekat v skrytu s úmyslem někoho nečekaně, zákeřně překvapit, přistihnout ap.: *policie čeká za bukem, aby vybrala pokuty za rychlost; policisté stojí někde za bukem a číhají na řidiče starých škodovek, aby zkontrolovali jejich technický stav*
◊ *číhat za bukem* kolokv. expr. čekat v skrytu s úmyslem někoho nečekaně, zákeřně překvapit, přistihnout ap.: *policie při kontrolách číhá za bukem, kde by řidiče co nejsnáze načapala*
◊ *(to je) kluk jak(o) buk* expr. (malý) chlapec je pěkně rostlý, zdravý a silný: *syna zvážili a zjistili, že je to kluk jako buk – tři kila osmdesát; včera se mi narodil kluk jak buk*
▶ **bouček** -čku (6. j. -čku, 6. mn. -čcích) m. než. zdrob. k 1: *vysazovat mladé boučky; zvěř může způsobit na nechráněných křehkých boučcích velké škody*

Figure 5: Entry *buk* "beech".

6 Software

Unlike earlier monolingual dictionaries created in the Institute of the Czech Language, the *ASSČ* has been compiled since the very beginning by means of specialised lexicographic software for dictionary creation (DWS). After various possibilities were considered, it was decided that new, special software needs to be developed as the creation of the dictionary has its significant specifics and the programme must be flexible. The software development has received grant support from the Ministry of Culture of the CR within the National and Cultural Identity (NAKI) applied research and development programme. (For more details on the software, see Barbierik et al. 2013; Barbierik et al. 2014.)

7 Conclusion

In the creation of the *ASSČ*, we are seeking new paths for the resolution of the issues that lexicographers have always faced as well as those of the modern present. Although the preparation of a good monolingual dictionary is a Sisyphean task – "the pursuit of perfection in lexicography is doomed to constant failure" –,[9] it must be attempted. "A dictionary of the national language is one of the basic needs of an educated man." (J. Jungmann, the preface to *Slovník česko-německý* (A Czech-German Dictionary)).

[9] We borrowed the metaphor at the end from Hanks (2005: 254).

8 References

Barbierik, K. et al. (2013). A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System. In K. Gajdošová, A. Žáková (eds.), *Natural Language Processing, Corpus Linguistics, E-learning, Proceedings of the conference Slovko 2013, Bratislava, 13–15 November 2013.* Lüdenscheid: RAM-Verlag 2013, pp. 9–26.

Barbierik, K. et al. (2014). Simple and Effective User Interface of Dictionary Writing System. In *Euralex 2014 Proceedings, Bolzano 15–19 July 2014.*

Buzássyová, K., Jarošová, A. (eds.) (2006). *Slovník súčasného slovenského jazyka A–G.* (First edition.) Bratislava: Veda.

Czech National Corpus – SYN2000. Institute of the Czech National Corpus, Prague 2000. Accessed at: http://www.korpus.cz [07/04/2014].

Czech National Corpus – SYN2005. Institute of the Czech National Corpus, Prague 2005. Accessed at: http://www.korpus.cz [07/04/2014].

Czech National Corpus – SYN2010. Institute of the Czech National Corpus, Prague 2010. Accessed at: http://www.korpus.cz [07/04/2014].

Czech National Corpus – SYN. Institute of the Czech National Corpus, Prague. Accessed at: http://www.korpus.cz [07/04/2014].

Čermák, F. (2010). Notes on Compiling a Corpus-Based Dictionary. In *Lexikos* 20 (*AFRILEX-reeks/series* 20:2010), pp. 559–579.

Čermák, F., Blatná, R. (eds.) (1995). *Manuál lexikografie.* Jinočany: H & H.

Dolník, J. (2012). Lexikálna pragmatika. In K. Buzássyová, B. Chocholová & N. Janočková (eds.), *Slovo v slovníku. Aspekty lexikálnej sémantiky – gramatika – štylistika (pragmatika). Na počesť Alexandry Jarošovej.* Bratislava: Veda, pp. 41–49.

Hanks, P. (2005). Johnson and Modern Lexicography. In *International Journal of Lexicography*, 18(2), pp. 243–266.

Hanks, P. (2010). Compiling a Monolingual Dictionary for Native Speakers. In *Lexikos* 20 (*AFRILEX-reeks/series* 20:2010), pp. 580–598.

Hladká, Z. (2007). Lexikografie. In J. Pleskalová, M. Krčmová et al. (eds.), *Kapitoly z dějin české jazykovědné bohemistiky.* Prague: Academia, pp. 164–198.

Jungmann, J. (1835 (1834) – 1839). *Slovník česko-německý.* (5 vol.) Prague: Knížecí arcibiskupská knihtiskárna.

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

Martincová, O. et al. 1998. *Nová slova v češtině. Slovník neologizmů 1.* Prague: Academia.

Martincová, O. et al. 2004. *Nová slova v češtině. Slovník neologizmů 2.* Prague: Academia.

Příruční slovník jazyka českého 1935–1957. Prague: Státní pedagogické nakladatelství / SPN.

Slovník spisovné češtiny pro školu a veřejnost (1978). (Second, revised edition 1994; third, revised edition 2003.) Prague: Academia.

Slovník spisovného jazyka českého (1960–1971). (First edition.) Prague: Nakladatelství ČSAV.

Soanes, C., Stevenson, A. (eds.) (2005). *Oxford Dictionary of English* (Second, revised edition.) Oxford: Oxford University Press.

Acknowledgements

The article has been written within the grant project of the National and Cultural Identity (NAKI) applied research and development programme A New Path to a Modern Monolingual Dictionary of Contemporary Czech (DF13P01OVV011).

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

Carolin Ostermann
Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
carolin.ostermann@fau.de

Abstract

Frame semantics has so far been neglected or even been rejected in the context of EFL-lexicography, although lexicographic description within a frame semantics approach would have advantages for learners, e.g. the coherent presentation of several relevant lexical items at a time, as well as their conceptual connection, both of which would also further vocabulary acquisition. This proposal will detail how a frame semantics approach for the example section in English monolingual learner's dictionaries can contribute to the notion of cognitive lexicography, i.e. a lexicography that puts an emphasis on how users process language, which would in turn facilitate a user's understanding of an entry. For this purpose, so-called *frame example sections* were developed on agentive nouns (e.g. *bridegroom, plaintiff*); these are small coherent text passages that define and exemplify the noun in relation to its whole frame. The frame example sections mention related frame elements, collocating verbs and describe the typical scenario underlying a semantic frame, in order to promote decoding, i.e. understanding the meaning of lexical items, as well as encoding, i.e. learning words and finding related language material. The paper will be rounded off by presenting the results of a small user-study that was conducted on the frame example sections.

Keywords: frame semantics; learner's dictionaries; cognitive lexicography; user-study

1 Frame Semantics and Learner's Dictionaries

Frame semantics in Fillmore's terms (1982) has come to be a widely accepted notion of semantic description, and in relation to lexicography, it has inspired the FrameNet online project (cf. Fontenelle 2003). In traditional lexicography, however, the approach has been neglected so far and even deemed useless (Bublitz and Bednarek 2004: 50). This paper will, however, demonstrate how frame semantics can be used in English monolingual learner's dictionaries. The approach is part of the larger concept of cognitive lexicography (cf. Ostermann 2012 and fthc.), in which theories and semantic analyses of cognitive linguistics are used in common lexicographic practice in order to create dictionary features and entries which are more accessible to the dictionary user, since they use and describe language in the same way the users process it.

Frame semantics is a very useful tool for meaning description in lexicography: Fillmore and Atkins (1992, 1994, 2000) have demonstrated several times in how far a frame approach can help with distinguishing meanings of polysemous items ('*risk*') and ensure a more realistic display of their different senses. This is one example of what Geeaerts (2007: 1168) refers to by stating generally that cognitive linguistics can enrich lexicography by a more realistic conception of semantic structure.

The feature proposed here aims at a more vivid exemplification of lexemes within the context of their frame, enabling the user to acquire new vocabulary from the frame and find important collocates for encoding, e.g. writing purposes. The feature replaces or complements example sentences in traditional dictionary entries as a so-called *frame example section (FE-section).* In the following, the structure and composition of FE-sections will be outlined, illustrating how they fit into a dictionary entry while at the same time offering an onomasiological access to the dictionary's macro-structure. A few remarks on a user-study conducted will round the paper off.

2 Frame Example Sections

2.1 Theory and Structure

For the application of frame semantics to a traditional dictionary entry the example section has been selected. Example sentences are especially suitable for being replaced or supported by *FE-sections* since they do not carry the main burden of rendering meaning but complement the definition by showing the meaning in context and offering typical collocations (cf. Drysdale 1987: 218-222). Since the FE-section is a small coherent text passage on a lexical item and mentions the frame with its frame elements and most important collocations, it additionally allows the user to grasp the meaning better. Regarding its language, the style of FE-sections is natural and typical, informative and intelligible, as good examples are supposed to be (Atkins and Rundell 2008: 458). This generally follows Fillmore's demands (2003: 283) that we should define "not words but only families of words that jointly express one frame".

For the writing of FE-sections, a suitable lexeme has to be chosen; in addition to the lexical items treated, the frame semantic content for each FE-section has to be established. The relevant frame and its frame elements are determined by using elicitation techniques, i.e. simple questions such as 'who', 'where', 'what', 'which aim'?. Superordinate place and collocating verbs are determined, and information from FrameNet is taken into consideration if the frame also figures there. Authentic language material is also collected from the BNC web, especially for collocations and related lexical items. Rundell (1988: 135) observed here very early that "(...) any account in a learner's dictionary of the word *problem* should at the very least mention as significant collocates the verbs *pose* and (especially) *solve*" and this can be ensured by an analysis of authentic language material. The FE-sections are written with the help of this collective input. Once the text has been produced, various perspectives in accordance with the various frame elements are created in order to be able to enter the FE-section at all the lexemes in the dictionary that are part of the frame. Finally, the potential for a related frame is checked, i.e. synonyms, antonyms or related semantic fields. The figure below summarizes the process of writing FE-sections.

SET-UP OF FRAME EXAMPLE SECTIONS:

1. Choice of the lemma: person-denoting noun.
2. Identification of the frame and frame elements.
3. Collection of authentic language material from the BNC, esp. of collocations.
4. Writing of the (main) frame example section with its annotations.
5. Check for perspectives of the frame example section.
6. Check for semantic 'spin-offs', i.e. related frames.
7. Decision of places to enter in the dictionary (in line with perspectives).

Table 1: Set-up of frame example sections.

2.2 The Set of Frame Example Sections

The FE-sections developed in this proposal centre on so-called person-denoting nouns. These are nouns that occupy an agentive slot in a frame, denoting a person and its habitual activities, and therefore provide a good perspective as a start, especially since they comprise actions and objects, as well as people or places that interact. The table below lists all the lexemes with their respective frame for which FE-sections have been produced. These 17 lexemes can also be divided into three groups: EVENT-frames (where something happens, usually starting with a preposition of time), ACTIVITY-frames (starting with *when* and introducing the setting of the frame), and PLACE-frames (taking place at typical locations).

bridegroom WEDDING ▪ *caretaker* BUILDING ▪ *conductor* ORCHESTRA ▪ *conductor* TRAIN ▪ *landlord* RENT ▪ *librarian* LIBRARY ▪ *mayor* CITY ▪ *midwife* BIRTH ▪ *pawnbroker* MONEY ▪ *plaintiff* COURT ▪ *striker* FOOTBALL ▪ *surgeon* OPERATION▪ *suspect* POLICE▪ *umpire* SPORT ▪ *undertaker* FUNERAL ▪ *usher* PERFORMANCE ▪ *waiter* RESTAURANT

Table 2: Person-denoting nouns and their frames.

For the lexeme *bridegroom*, the FE-section is reproduced below with annotations: the WEDDING-frame is an event frame, i.e. one where something happens. *Bridegroom, bride, husband* and *wife* are frame-constitutive elements, i.e. those which are necessary to understand the frame, and are printed in small capitals. Frame-supportive elements, i.e. those which are optional for an understanding of the frame and rather expand it, here *priest/pastor, church, reception*, are underlined. Collocations (*on the wedding day, to get married*) are given a dotted underlining. The full annotations including sources of authentic language material (here from BNC web and FrameNet) and perspectives for the FE-section on *bridegroom* can be found in the appendix, as well as the FE-sections for all the other items.

***bridegroom* - WEDDING**
On their WEDDING day, the BRIDE and the BRIDEGROOM get married and become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
ANNOTATION: [On$^{\text{Coll}}$ their wedding day]$^{\text{Event}}$, the bride$^{\text{Partner1/WhoColl}}$ and the bridegroom$^{\text{Partner 2/Who}}$ get married$^{\text{CollActivity}}$ and become$^{\text{[change relationship] / Goal}}$ husband and wife$^{\text{Partners}}$. A priest or pastor in church$^{\text{where}}$ traditionally marries them with family and friends present. Afterwards, there often is a wedding reception$^{\text{Coll}}$.

Table 3: FE-section for *bridegroom.*

2.3 A Cognitive Macro-structure

Many of the person-denoting nouns are rather rare items (cf. *pawnbroker, plaintiff, usher*) or only supposedly transparent lexical items (*caretaker, landlord*), which makes them very interesting in a language-learning perspective. It would be ideal if, once these items are looked up, they would become attached to a user's mental lexicon, possibly via familiar material and links within their frame. This is also the reason why the various perspectives of FE-sections are written und the FE-sections should be entered repeatedly at the entries of all their participating lemmata.

In this way, all the person-denoting nouns also contribute to a macrostructure that exhibits more links between single items than traditional dictionaries do, and one that also allows for onomoasiological access. Every FE-section spans a small net over the macro-structure with its single frame elements; all FE-sections together span an even larger net since they often share elements or deal with polysemy (cf. the two FE-sections for *conductor*).

This is also in accordance with Geeraerts' assumption (2007: 1169) that "Cognitive Linguistics may also suggest ways of dealing with the links between the senses of lexical items that go beyond common practice". If we suppose that the FE-section is – whether incorporated within the dictionary entry or in a box nearby – clearly delimited regarding its layout (e.g. use of colours, etc.), it almost automatically leads the user to related entries, especially since the same information of one frame can be found in all the places of the frame in the dictionary. In an electronic, online or CD-ROM-version of a dictionary, this could even be achieved more effectively by hyperlinking. FE-sections therefore also fulfill lexicographically the function of signposts (the capital print of the frame as a meaning indication via synonym, cf. DeCesaris 2012) and of component-internal implicit cross-references (cf. Svensén 2009: 388 and 391), in which many entries of one frame, but also across frames are linked.

3 A User-Study on Frame Example Sections

3.1 Methodology

In order to determine the usefulness of FE-sections, a small-scale user-study was conducted with 50 university students of English. The hypothesis was that in a two-part production-oriented primed vocabulary task, the group of students in the target group (n^t=25) who received dictionary entries of the respective lexemes, complemented by FE-sections, would perform better than those in the control group (n^c=25) who worked with traditional dictionary entries only.

The participants received in the first part of the experiment a randomised reading booklet with the LDOCE5-dictionary entries of 12 of the above-mentioned lexemes as a prime (two groups of six items: *caretaker, midwife, pawnbroker, plaintiff, umpire, usher* and *conductor*[1], *conductor*[2], *landlord, striker, surgeon, undertaker*). The participants in the target group worked with reading-booklets in which the dictionary entries were complemented by the FE-sections; the participants in the control group received dictionary entries complemented by reading material on the lexemes taken from the BNC, so that both groups had the same amount of reading material to master. On each page of the booklet, they found one entry and were supposed to read it carefully within ca. 25 seconds, turning the page only when being told to do so and not going backwards. This session was devised as primed input for the second part of the experiment, which followed after a break of approximately 45 minutes. In this second part, the test subjects received a worksheet on the 12 person-denoting nouns, on which they were supposed to give a German translation of each noun, define it in their own words and tick off in a list whether they had known the word before.

It must be noted generally, however, that the hypothesis could not be verified, since the experiment yielded inconclusive, statistically non-significant results.

3.2 Results and Discussion

Regarding the knowledge of the test items, it can be concluded that the test was conducted in a homogeneous group with approximately the same level of knowledge of all the items across the participants. The items from the first group, such as *pawnbroker, plaintiff, usher* and *umpire*, were rated very low and were fairly unknown, whereas the items from the latter six received higher ratings.

For the results of the translation task (reproduced in the chart below), the scores of correct translations were counted for each item in both groups and compared; the significance of difference was checked with the help of the χ^2-test. The numbers of correct translations are approximately equal for all items, with the exceptions of *pawnbroker* and *landlord*, which proved to be statistically significant (χ^2= 2.01, p<0.20 and χ^2= 3.57, p<0.10). It should be noted, however, that many students seemed to have had problems in coming up with a good translation, since a certain number of the participants suggested e.g German 'Torschütze' instead of 'Torjäger' for *striker* and did not even seem to be aware of the semantic difference. Therefore, demanding a German translation might not have been the best measure, as it yielded problems of its own, even when the concepts behind the lexical items were apparently understood, since in many cases, correct paraphrases were given.

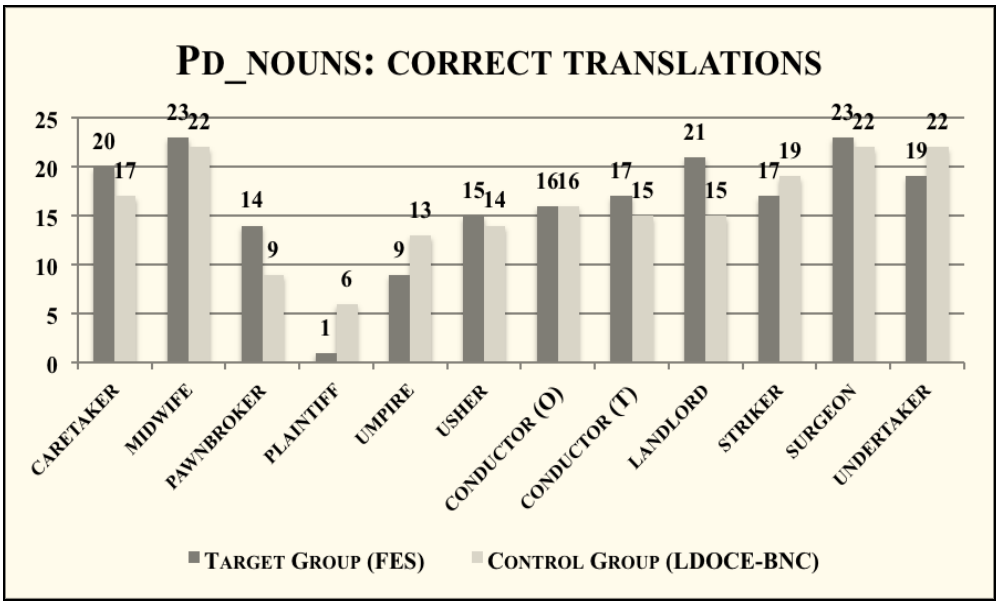


Figure 1: Results of translation task

In order to evaluate the results of the paraphrasing task, a point system was devised. Points were assigned in the participants' paraphrases to a correct paraphrase in general, to the frame mentioned and to all the frame elements reproduced. Generally, the participants in the target group scored higher for each item and in general (36.22 points on average compared to 29.34 points for the participants in the control group), and their paraphrases were also longer (14.97 words on average compared to 11.48 words in the control group). It has to be admitted, however, that there is a certain correlation between the amount of input and output, which, on the other hand, admits the conclusion that more input in the form of FE-sections is indeed beneficial. It should be noted that a learning effect (i.e. when people indicated that they had not known the word before but gave a correct definition) could be achieved more often in the target group and that the number of paraphrases given compared to the number of correct paraphrases given was equal in more instances in the target group. The non-transparent item *landlord* ('Vermieter' in German, but its parts often translated literally as 'Landherr' / 'Lehensherr') caused fewer misunderstandings among the participants in the target group, the effect of which can also be attributed to the cognitive FE-sections. Therefore, the FE-sections did score an effect, even if it was small and statistically not significant.

Overall, it can be concluded that the complexity of the task probably made it difficult to measure the effect that FE-sections can have. The reading time might not have been sufficient for vocabulary acquisition, especially since "lexical acquisition is not immediate" (Béjoint 1988: 145), and vocabulary items will not get a real foothold in one's mental lexicon through decoding alone (Atkins and Rundell

2008: 410). The more difficult the items were (e.g. *plaintiff* compared to the simpler *midwife*), the poorer the results were, or the more blanks could be found on the worksheets; only single instances of a better performance with one item or another, or cases of real vocabulary acquisition in the target group could be ascertained. Possibly, the wealth of information in the FE-sections also hindered immediate acquisition with difficult items. These effects could be elucidated in another test condition or in a longer testing phase with repeated tasks or dictionary training of the participants.

4 Conclusion

All in all, it can be concluded that FE-sections are a new approach for EFL-lexicography which would probably work best in an individual look-up situation. Although no superior results over traditional dictionary entries could be proven statistically, the benefits still come into play, and this is one step on the way to a more cognitive and more onomasiological dictionary of encyclopaedic nature.

5 References

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

Béjoint, H. (1988) Psycholinguistic evidence and the use of dictionaries by L2 learners. In: M. Snell-Hornby (ed.). *ZüriLEX `86 Proceedings.* Tübingen: Francke. pp. 139-148.

Bublitz, W. and M. Bednarek. (2005) Nur im begrenzten Rahmen. *Frames* im Wörterbuch. In: T. Herbst, G. Lorenz, B. Mittmann, and M. Schnell (eds.) Lexikographie, ihre Basis- und Nachbarwissenschaften. (Englische) Wörterbücher zwischen "common sense" und angewandter Theorie. Tübingen: Niemeyer. pp. 35-52.

DeCesaris, J. (2012) On the Nature of Signposts. In: *Proceedings of the 15th EURALEX International Congress.* Universitetet i Oslo. pp. 532-540.

Drysdale, P.D. (1987) The role of examples in a learner's dictionary. In: A. P. Cowie (ed.). The dictionary and the language learner: papers from the EURALEX seminar at the University of Leeds, 1-3 April 1985. Tübingen: Niemeyer. pp. 213-223.

Fillmore, Ch. J. (1975) An Alternative to Checklist Theories of Meaning. In: C. Cogen, H. Thompson, G. Thurgood, K. Whistler and J. Wright (eds). *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society.* Berkeley: Berkeley Linguistics Society. pp. 123-131.

- 2003. Double-Decker Definitions: The Role of Frames in Meaning Explanation. In: *Sign Language Studies* 3(3): pp. 263-295.

Fillmore, Ch. J. and B.T.S. Atkins (1992) Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors. In: A. Lehrer and E. Feder Kittay (eds.). Frames, Fields and Contrasts. New Essays in Semantic and Lexical Organization. Hillsdale, New Jersey / London: Erlbaum. pp. 75-102.

- 1994. Starting where the dictionaries stop: The challenge of Corpus Lexicography. In: B.T.S. Atkins and A. Zampolli (eds.). Computational Approaches to the Lexicon. Oxford: Oxford Univ. Press. pp. 349-393.
- 2000. Describing Polysemy: The Case of 'Crawl'. In: Y. Ravin and C. Leacock (eds.). Polysemy. Theoretical and Computational Approaches. Oxford: Oxford University Press. pp. 91-110.

Fontenelle, Th. (ed.) (2003) Special issue on FrameNet and frame semantics. In: *International Journal of Lexicography* 16(3).

FrameNet. Accessed at: https://framenet.icsi.berkeley.edu [10/04/2014].

Geeraerts, (2007) Lexicography. In: D. Geeraerts and H. Cuyckens (eds.). The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford University Press. pp. 1160-1174.

LDOCE5 = *Longman Dictionary of Contemporary English.* **5**th ed. 2009. Ed. director Michael Mayor. Harlow: Pearson – Longman.

Ostermann, C. 2012. Cognitive Lexicography of Emotion Terms. In: *Proceedings of the 15th EURALEX International Congress.* Universitetet i Oslo. pp. 493-501.

- fthc. Cognitive Lexicography. In: S. Niemeier and C. Juchem-Grundmann, with D. Schönefeld (eds.). Dictionaries of Linguistics and Communication Science (WSK) online, Vol 14: Cognitive Grammar*.* Berlin: Mouton de Gruyter. [online. 1 page].

Rundell, M. (1988) Changing the rules: Why the monolingual learner's dictionary should move away from the native-speaker tradition. In: M. Snell-Hornby (ed.). *ZüriLEX `86 Proceedings.* Tübingen: Francke. pp. 127-137.

Svensén, B. (2009) A Handbook of Lexicography. The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

The BNC web. Accessed at: http://bncweb.lancs.ac.uk [10/04/2014].

Appendix 1: An Annotated Example for ***bridegroom***

1. Lemma: *bridegroom*	
2. Frame: WEDDING	
2. Frame elements	bride, bridegroom, husband, wife, church, priest / pastor ⇨ superordinate place: church ⇨ collocating verb: marry ⇨ kind of frame: EVENT
2.a Elicitation techniques	⇨ who, where, activity, goal?
2.b FrameNet Frame FEs from FN FrameNet definition	Forming_relationships ⇨ Partner 1, Partner 2, Partners; Epistemic stance ⇨ Partner 1 interacts with Partner 2 (also collectively called Partners) to change their social relationship.
3. Authentic language material	
Collocations from BNC	to get married; (on their) wedding day, wedding reception, bride
4. Frame example section	
On their WEDDING day, the BRIDE and the BRIDEGROOM get married and become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.	
ANNOTATION	
[On$^{\text{Coll}}$ their wedding day]$^{\text{Event}}$, the bride$^{\text{Partner1/WhoColl}}$ and the bridegroom$^{\text{Partner 2/Who}}$ get married$^{\text{CollActivity}}$ and become$^{\text{[change relationship] / Goal}}$ husband and wife$^{\text{Partners}}$. A priest or pastor in church$^{\text{where}}$ traditionally marries them with family and friends present. Afterwards, there often is a wedding reception$^{\text{Coll}}$.	
5. Different perspectives	
BRIDE	On their WEDDING day, the BRIDE gets married to her BRIDEGROOM and they become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
GROOM	On their WEDDING day, the BRIDEGROOM gets married to his BRIDE and they become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
WIFE	On their WEDDING day, the BRIDE and the BRIDEGROOM got married and became HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
HUSBAND	On their WEDDING day, the BRIDE and the BRIDEGROOM got married and became HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
6. Semantic spin-off	
antonym	divorce
7. Place(s) in the dictionary	
wedding ▪ bridegroom ▪ bride ▪ husband ▪ wife	

Table 4: Full annotation for *bridegroom*.

Appendix 2: The set of frame example sections

bridegroom WEDDING	On their WEDDING day, the BRIDE and the BRIDEGROOM get married and become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and fri ends present. Afterwards, there often is a wedding reception.
caretaker BUILDING	In a public BUILDING, e.g. a school, a CARETAKER (or also JANITOR) is the person who looks after the BUILDING. S/he takes care of the BUILDING's maintenance and makes sure that everything is in order, that broken things are repaired or that rules are obeyed. The CARETAKER usually has his or her own OFFICE in the BUILDING where s/he can be found.
conductor TRAIN -------------- ORCHESTRA	In a TRAIN, a CONDUCTOR (or also GUARD) is responsible for checking and collecting or also selling the PASSENGERS' TICKETS; s/he furthermore is in charge of the train, making sure everything is in order or answering the passengers' questions. Conductors also travel on BUSES where they collect the fare. -------------- When an ORCHESTRA OR CHOIR performs, either as a rehearsal or in front of an AUDIENCE, a CONDUCTOR stands in front on a podium and conducts, i.e. directs the MUSICIANS' PERFORMANCE with a baton (small thin stick). The MUSICIANS follow the CONDUCTOR's movements so that all play in a coordinated way and the PERFORMANCE sounds good.
landlord RENT	When you RENT A PLACE TO LIVE, i.e. an apartment/flat or a house, you pay MONEY, the RENT, to the LANDLORD who owns the building and lets you live there. You are then the TENANT and a formal contract, the lease, guarantees your rights as a TENANT.
librarian LIBRARY	In a LIBRARY, a LIBRARIAN is the person who is in charge of running the institution, i.e. lending BOOKS or other materials to LIBRARY users. People can read the BOOKS there or they can borrow them. Schools and universities usually have their own libraries and their use is often free of charge.
mayor CITY	In a CITY OR TOWN, the MAYOR is the head of the local GOVERNMENT. S/He is elected directly by the citizens and resides in a city or town hall. S/he fulfils official duties and functions and makes decisions in local politics.
midwife BIRTH	When a pregnant WOMAN goes into labour and is about to give BIRTH to a BABY, she usually goes to hospital. There, she gets help from a MIDWIFE, who is a nurse helping women to get through labour pains and who also takes care of the MOTHERS and their BABIES before and after birth.
pawnbroker MONEY	When you are in urgent need of MONEY, but cannot or do not want to borrow money from a bank, you may turn to a PAWNBROKER in a PAWNSHOP. S/he will lend you money in exchange for valuable OBJECTS, e.g. jewellery or electronic devices. If you cannot pay back the MONEY after a certain while, the pawnbroker will sell what you have PLEDGED.
plaintiff COURT	In COURT, a PLAINTIFF brings a CASE against another person, the defendant. The PLAINTIFF is usually supported by a LAYWER (in Britain a solicitor in the lower courts of law) to fight the case successfully, and the judge or a jury has to decide on the verdict.
striker FOOTBALL	In a FOOTBALL MATCH, the STRIKER is the PLAYER whose main task on the PITCH it is to score a GOAL and help his team to win, which the other team's PLAYERS and especially the goalkeeper try to prevent.
surgeon OPERATION	During an OPERATION, a SURGEON is the doctor who cures and rescues PATIENTS by performing surgery, i.e. by operating on patients in a HOSPITAL in an OPERATING THEATRE with nurses and other doctors assisting. Patients who undergo surgery are seriously ill and usually stay in hospital to recover.
suspect POLICE	When the POLICE think that a person took part in a CRIME, they arrest this person, who is a SUSPECT. After the ARREST, the SUSPECT is taken into custody at the POLICE STATION for a police interview / an interrogation.
umpire SPORT	During a SPORTS COMPETITION in an arena, an UMPIRE is the person who makes sure that RULES are obeyed. There is an UMPIRE present in e.g. baseball, tennis, cricket, hockey, or athletics COMPETITIONS; s/he also calls the score, decides on penalties, starts races, or reports irregularities to chief UMPIRES (depending on the discipline).
undertaker FUNERAL	After somebody's DEATH, a FUNERAL is held at a CEMETERY. An UNDERTAKER OR FUNERAL DIRECTOR prepares the deceased person's burial or cremation and arranges the FUNERAL service, so that people can attend the ceremony and mourn the loss of the deceased.
usher PERFOR-MANCE / EVENT	When people go to see a public PERFORMANCE OR EVENT, e.g. in a theatre, a cinema, a concert hall, or a sports stadium, they show their TICKETS to an USHER (or USHERETTE) who shows them their SEATS or even guides them there. Often, the USHER also keeps order during a show.
waiter RESTAURANT	In a RESTAURANT, people sit at TABLES and eat a MEAL for which they have to pay the bill at the end. A WAITER OR WAITRESS brings customers the MENU first and later serves the food they ordered.

Table 5: The Set of Frame Example Sections.

Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology

Alessandro Panunzi°, Irene De Felice*, Lorenzo Gregori°, Stefano Jacoviello†, Monica Monachini*, Massimo Moneglia°, Valeria Quochi*, Irene Russo*
°University of Florence, *ILC CNR (Pisa), †University of Siena
alessandro.panunzi@unifi.it, irene.defelice@ilc.cnr.it, lorenzo.gregori@unifi.it, stefano.jacoviello@gmail.com, monica.monachini@ilc.cnr.it, moneglia@unifi.it, valeria.quochi@ilc.cnr.it, irene.russo@ilc.cnr.it

Abstract

Action verbs have many meanings, covering actions in different ontological types. Moreover, each language categorizes action in its own way. One verb can refer to many different actions and one action can be identified by more than one verb. The range of variations within and across languages is largely unknown, causing trouble in all translation tasks. IMAGACT is a corpus-based ontology of action concepts, derived from English and Italian spontaneous speech corpora, which makes use of the universal language of images to identify the different action types extended by verbs referring to action in English, Italian, Chinese and Spanish. This paper presents the IMAGACT search interface and the various kinds of linguistic information the user can derive from it. IMAGACT makes explicit the variation of meaning of action verbs within one language and allows comparisons of verb variations within and across languages. Because the action concepts are represented with videos, extension into new languages beyond those presently implemented in IMAGACT is done using competence-based judgments by mother-tongue informants, without intense lexicographic work involving underdetermined semantic descriptions.

Keywords: Action verbs; Image ontology; Multilingual dictionary; Computer-aided translation

1 Introduction

In all language modalities, action verbs bear the basic information that should be understood in order to make sense of a sentence. Moreover when we communicate, we have to refer to actions very often. Native speakers do not have a problem finding the right verb for a specific action in their own language. However, in a foreign language, they often have difficulty choosing the appropriate verb. The reason is that the more common action verbs, in their own meaning, refer to many different actions: in this sense, they are "general" verbs. Moreover, each language categorizes actions in its own way. These facts imply that there are not one-to-one translation relationships between different general verbs in different languages (Majid et al. 2007; Kopecka & Narasimhan 2012). If we take the English verb *to*

cross, for instance, we could argue that it can refer to at least two different action types, as in the sentences:
(1) John crosses the street
(2) John crosses his arms
On the contrary, in Italian we must use two different verbs to translate the previous sentences, namely *attraversare* (for *crossing the street*) and *incrociare* (for *crossing arms*):
(3) Gianni attraversa la strada
(4) Gianni incrocia le braccia

The problem is a significant one because reference to action is very frequent in ordinary spoken communication (Moneglia & Panunzi 2007) and specifically high-frequency verbs can each refer to many different action types (Moneglia in press).
The IMAGACT project has now delivered a corpus-based language ontology covering the set of actions most frequently referred to in everyday language. Using English and Italian spoken corpora, we have identified 1010 distinct action concepts and visually represented them by means of prototypical scenes, either animated (3D) or filmed (Moneglia et al. 2012; Frontini et al. 2012). The cross-linguistic correspondences to action concepts of 521 Italian verbs and 550 English verbs (i.e., the verbal lexicon most likely to be used when referring to action) are stored in a database. The action concepts in IMAGACT have already been extended to Chinese and Spanish (included in the first IMAGACT release). Perhaps more importantly, the action concepts can be easily identified by speakers of any language, since they are represented in an ontology of animated and filmed scenes.
This paper presents the IMAGACT online interface and how queries are made to the database. The user can search in IMAGACT in three main ways: a) as a bilingual dictionary, based on concept selection; b) through explicit comparison of the range of actions that can in principle be referred to by two lexical entries, of the same language or of different languages; c) through the direct selection of an action concept in the gallery of prototypic scenes, independently of the language of the user. In the last section, the paper also introduces an initiative aimed at the extension of the IMAGACT database to other languages.

2 Dictionary

If the user wonders how an English action verb translates into Italian or into another target language (Spanish and Chinese in the IMAGACT first release), IMAGACT can be used as a multilingual dictionary of images. Figure 1 shows the thumbnail images of the main types of actions identified by the English verb *to cross*.

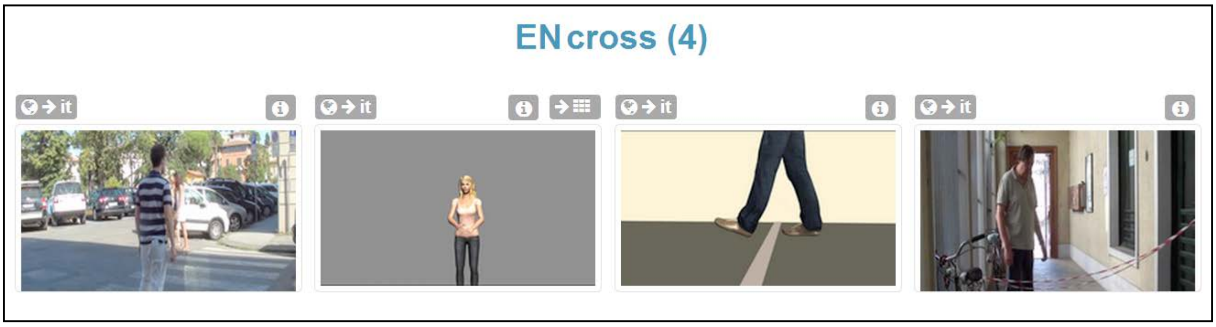
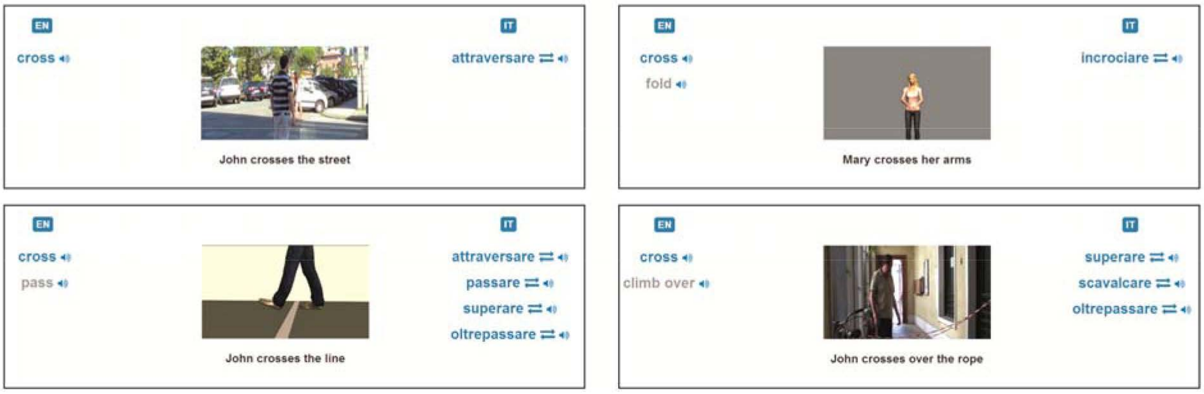


Figure 1: The variation of *to cross* across action types.

Looking at the various action types this verb expresses, the user can:

- select the action type he is interested in
- look at the animation to clarify the referred action
- see how this action is identified in the target language

IMAGACT returns one main verb and an additional set of verbs which equally identify this specific type of action. For each scene, which represents a distinct action type, Italian gives different translation for this verb, as shown in Figure 2.



John crosses the street

Mary crosses her arms

John crosses the line

John crosses over the rope

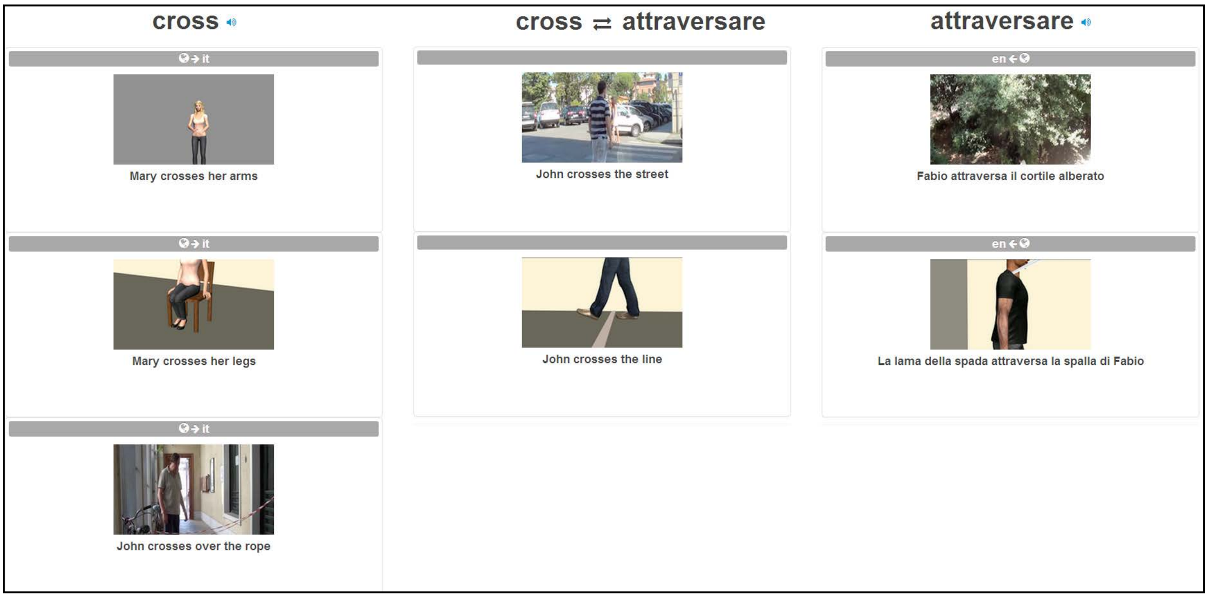
Figure 2: The cross-linguistic relation of verb(s) to action types.

3 Comparison

The user can compare verbs that in principle should translate between each other from two different languages. Searching with this function, the system illustrates the set of action types in which both verbs can be respectively applied. The result of such a search for *to cross* and *attraversare* (see Figure 3) supports the intuition that the two verbs can translate to each other, at least with respect to some of the action types they can refer to. At the same time, however, the system shows which actions can be indicated by one verb but not by the other, and *vice versa*. As a consequence, the difference between the

Italian verb *attraversare* and the English verb *to cross* becomes explicit. The Italian user will learn that, in English, *to cross* cannot be applied to the types on the right column in Figure 3. In this case, he can go directly to the English translation of verb *attraversare*, as shown in Figure 4: for these two actions he has to use, respectively, *to traverse / to pass* and *to stab / to pierce*.

Comparison between two verbs can also be requested within the same language, to allow the user exploring more deeply the differences in meaning between the lexical entries suggested by the system. For instance, an English user can learn that both the verbs *passare* and *attraversare* can be applied to the action type illustrated by the scene on the left column, second row, in Figure 4. The user may wonder what the difference is between the two Italian lemmas suggested by the system. So, he can then compare the two verbs (of the same target language), clarifying the differences between their referential properties (Figure 5).



Mary crosses her arms

Mary crosses her legs

John crosses over the rope

John crosses the street

John crosses the line

Fabio attraversa il cortile alberato

La lama della spada attraversa la spalla di Fabio

Figure 3: Comparison of *turn* vs. *girare* (results of the query interface with graphic adaptations).

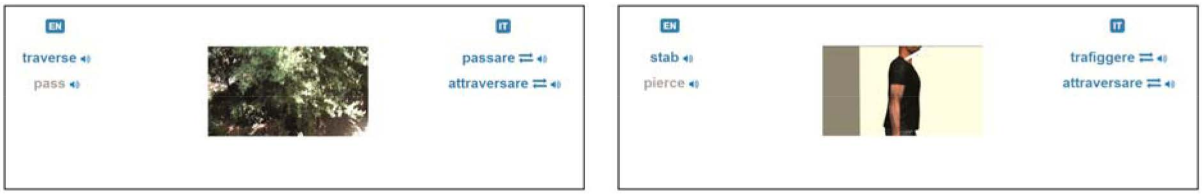
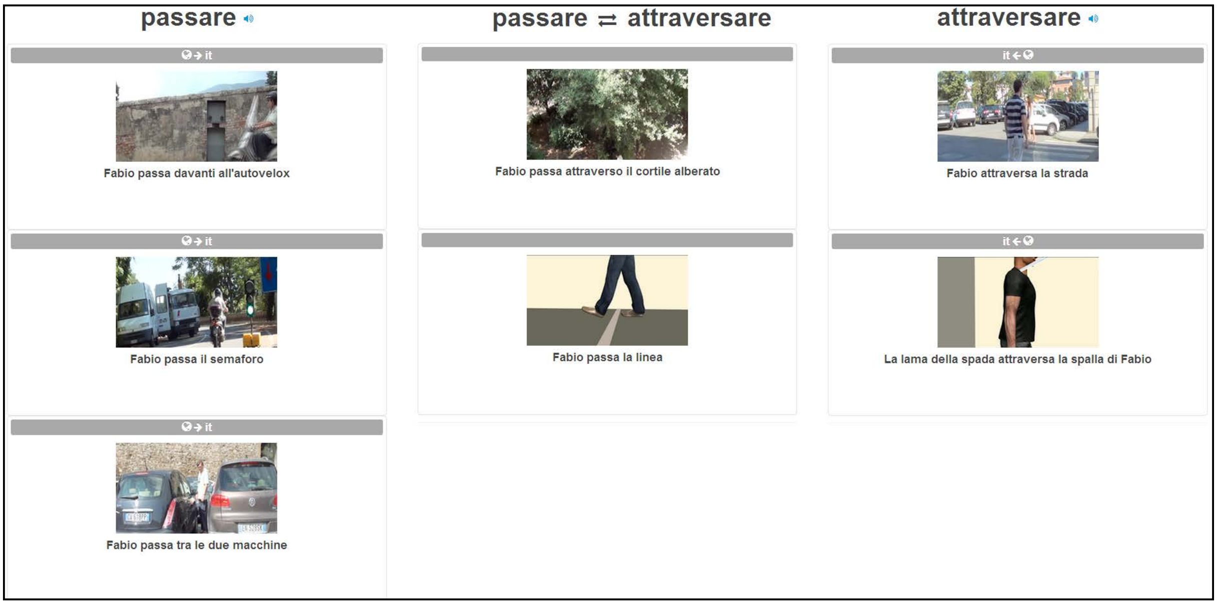


Figure 4: From comparison to linguistic categorization.



passare | passare ⇄ attraversare | attraversare

Fabio passa davanti all'autovelox

Fabio passa il semaforo

Fabio passa tra le due macchine

Fabio passa attraverso il cortile alberato

Fabio passa la linea

Fabio attraversa la strada

La lama della spada attraversa la spalla di Fabio

Figure 5: Intra-linguistic comparison.

4 Gallery

If the language of the user is not represented in IMAGACT, he can use the system directly as a gallery of scenes. This may be of special interest to users who speak minority languages.

The system works through the selection of one “meta-category” of action among the ones proposed by the interface. Such meta-categories are represented by a series of 3D animations, which are continuously played in loop, as the thumbnails in Figure 6 suggest.

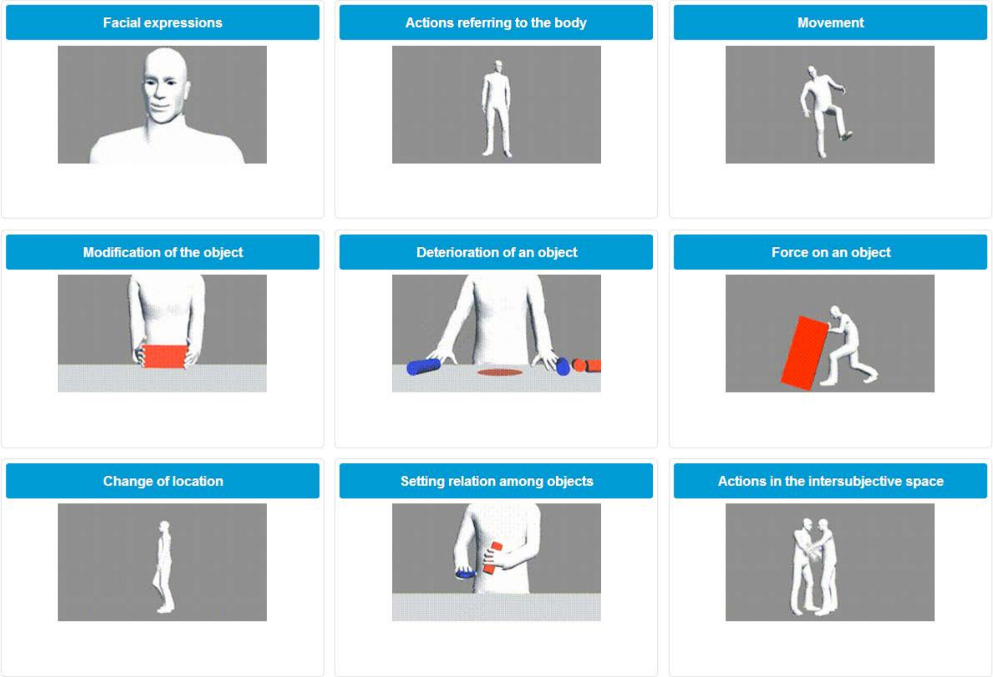

Facial expressions
Actions referring to the body
Movement
Modification of the object
Deterioration of an object
Force on an object
Change of location
Setting relation among objects
Actions in the intersubjective space

Figure 6: Representation of action meta-categories through avatars.

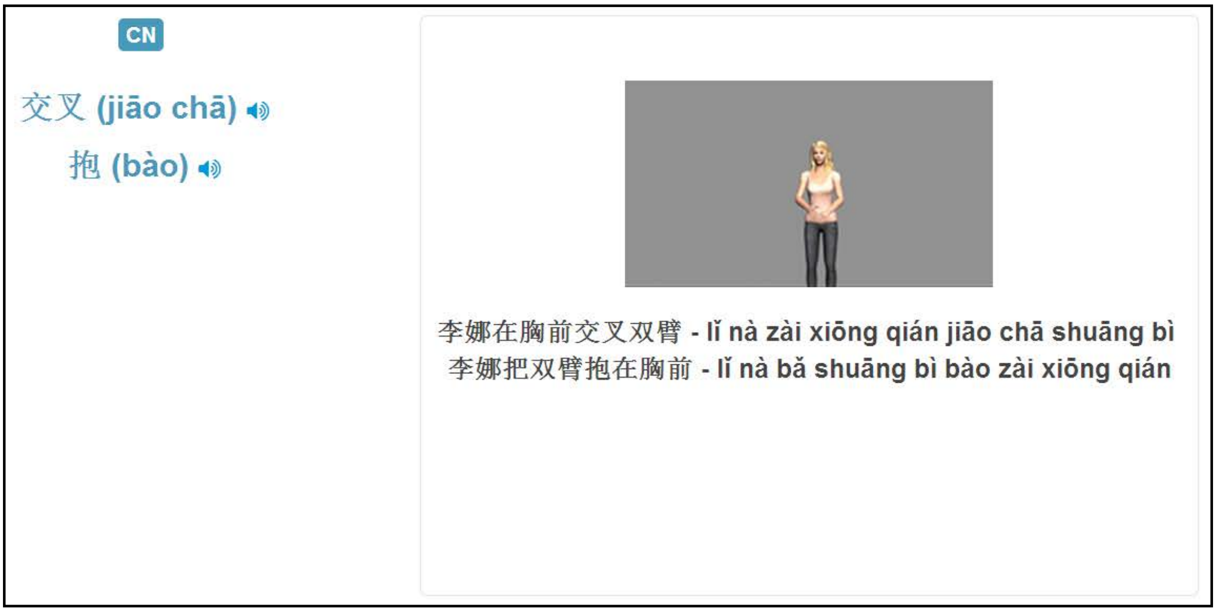
The numerous actions covered by IMAGACT are gathered into 9 macro-classes, which have high relevance in human categorization of action. Meta-categories are ordered according to criteria which take into account the informative focus of the action, as reported in Table 1.

Perspective centered on the Actor	**Perspective centered on the Actor-Theme relation**	**Perspective centered on the Theme-Destination relation**
Actions referring to facial expression	Modifications of the object	Change of location of the object
Actions referring to the body	Deterioration of the object	Setting relations among objects
Movement in space	Forces on the object	Actions in inter-subjective space

Table 1: Criteria for meta-categories.

The user can figure out what kind of action these stand for by looking at the abstract representation heading each class, and of course through a quick look at the actions gathered under each one. The user identifies the action he is interested in independently of the word he has for that action in his own language; after choosing the action via its visual representation, he is able to reach its linguistic categorization in the required target language. From this point of view, the IMAGACT gallery reverses the ordering of the dictionary: it goes from concepts to language instead of from language to concepts.

Once the user has understood the meaning of the action groups, it will be easier to search for the specific action he is interested in. He will click on one scene in the gallery headed by one category and get the linguistic categorization of the concept in one of the possible target languages in IMAGACT. For instance, Figure 7 is what the system returns when asked for the Chinese verb for the action corresponding to the verb *to cross* under the category *Actions referring to the body* (i.e., *crossing the arms*).



交叉 (jiāo chā)

抱 (bào)

李娜在胸前交叉双臂 - lǐ nà zài xiōng qián jiāo chā shuāng bì

李娜把双臂抱在胸前 - lǐ nà bǎ shuāng bì bào zài xiōng qián

Figure 7: From gallery to linguistic categorization (Chinese).

5 Extending the dictionary

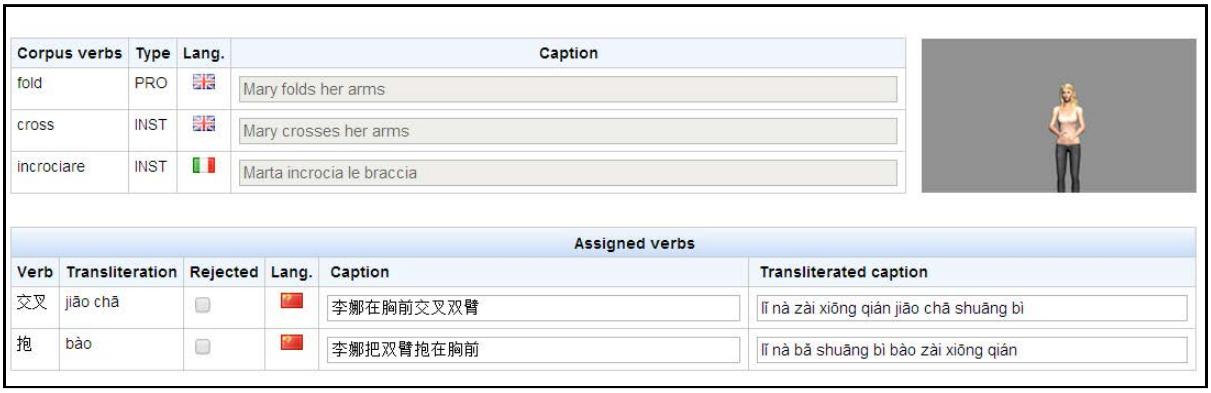
Because IMAGACT's direct representation of actions through scenes can be interpreted independently of language, the system allows the mapping of lexicons from different languages onto the same cross-linguisitic ontology. On this basis, it is possible to ask mother-tongue informants which verb(s) in their languages should be applied to each scene, thus extending the ontology to any language (IMAGACT4ALL).

In the simplified interface for the Competence Based Extension of the IMAGACT database to other languages (called *CBE light*), the set of action concepts represented by the IMAGACT prototypic scenes is assumed as a fixed-reference universe, and the work starts directly from such scenes.

An informant receives the action types as input. Figure 8 shows the interface the informant would use for processing one action type and how this has been done in the case of Chinese. The interface presents the informant with the scene prototype and the matching English and Italian verbs derived from corpus analysis. The informant assesses the action represented in the video and provides the verb or verbs in his language that can be used to refer to that specific action.

Lemmas are annotated in its citation form, as it is commonly reported in dictionaries, in the box corresponding to his language. For each lemma he then writes in the caption box a simple sentence in the present tense, third-person singular, filling all the arguments of the verb that properly describes the action. This sentence will be used as the caption of the scene in the language of the informant.

Both the verb and the caption should be written in the current writing system of the language of the informant. If this system does not use Latin characters, the informant also provides the verb and its caption in Latin characters, as can be seen for Chinese.



Corpus verbs	Type	Lang.	Caption
fold	PRO	🇬🇧	Mary folds her arms
cross	INST	🇬🇧	Mary crosses her arms
incrociare	INST	🇮🇹	Marta incrocia le braccia

Assigned verbs

Verb	Transliteration	Rejected	Lang.	Caption	Transliterated caption
交叉	jiāo chā	□	🇨🇳	李娜在胸前交叉双臂	lǐ nà zài xiōng qián jiāo chā shuāng bì
抱	bào	□	🇨🇳	李娜把双臂抱在胸前	lǐ nà bǎ shuāng bì bào zài xiōng qián

Figure 8: Simplified Competence Based Extension (CBE light).

Given that verbs with different meanings can identify the same action, the informant is asked to find multiple lemmas allowed by his language for each action. However, simply viewing one short clip may be not sufficient to elicit all the alternatives. The infrastructure provides one simple means to stimulate the thinking of the informant. More specifically, corpus-based annotation generated English and Italian alternatives that fit with the represented scene. These verbs will function as suggestions for figuring out alternatives in the language of the informant. Therefore, after the first lemma has been determined, the annotator is requested to judge whether or not the alternatives suggested have translations in his language, translations that can be used in referring to the event in question. If so, he will report a new verbal lemma and a new caption by adding a line to his language options. The work of the informant must be supervised by a mother-tongue expert linguist before the language is mapped onto the IMAGACT database. More specifically, an annotation can be rejected by the supervisor during revision if considered inappropriate. Spanish and Chinese have already been implemented through IMAGACT4ALL, and various initiatives are currently being pursued to extend the database to a number of different languages.

6 References

Frontini, F., De Felice, I., Khan, F., Russo, I., Monachini, M., Gagliardi, G., Panunzi, A. (2012). Verb Interpretation for Basic Action Types: Annotation, Ontology Induction and Creation of Prototypical Scenes. In M. Zock, R. Rapp (eds) *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex III), Mumbay, 15 December 2012.* Red Hook (NY): Curran Associates, pp. 69-80.

IMAGACT. Accessed at: http://www.imagact.it [04/04/2014].

Kopecka, A., Narasimhan, B. (2012). *Events of Putting and Taking, A Cross-linguistic perspective.* Amsterdam/Philadelphia: John Benjamins.

Majid, A., Bowerman, M., van Staden, M., Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. In *Cognitive Linguistics*, 18(2), pp. 133-152.

Moneglia, M. (in press). The semantic variation of action verbs in multilingual spontaneous speech corpora. To appear in T. Raso, H. Mello (eds) *Spoken Corpora and Linguistic Studies.* Amsterdam/Philadelphia: John Benjamins.

Moneglia, M., Monachini, M., Calabrese, O., Panunzi, A., Frontini, F., Gagliardi, G., Russo, I. (2012). The IMAGACT Cross-linguistic Ontology of Action. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds) *Proceedings of Eighth Language Resources and Evaluation Conference (LREC 2012), Istanbul, 23-25 May 2012.* Paris: ELRA, pp. 2606-2613.

Moneglia M., Panunzi A. 2007. Action Predicates and the Ontology of Action across Spoken Language Corpora. The Basic Issue of the SEMACT Project. In M. Alcántara Plá, Th. Declerk (eds) *Proceeding of the International Workshop on the Semantic Representation of Spoken Language, SRSL7, at Conferencia de la Asociación Española de Inteligencia Artificial, CAEPIA 2007, Salamanca, 12-16 November 2007.* Salamanca: Universidad de Salamanca, pp. 51-58.

Acknowledgements

The IMAGACT project is funded by the PAR/FAS program of the Tuscan region in Italy. Further research on the IMAGACT database, including this paper, is being accomplished with the contribution of the MODELACT project (2013-2016), funded by the national programme Futuro in Ricerca.

Degrees of Synonymity as the Basis of a Network for German Communication Verbs in the Online Reference Work *Kommunikationsverben* in OWID

Kristel Proost, Carolin Müller-Spitzer
Institut für Deutsche Sprache, Mannheim
proost@ids-mannheim.de, mueller-spitzer@ids-mannheim.de

Abstract

This contribution presents the procedure used in the *Handbuch deutscher Kommunikationsverben* and in its online version *Kommunikationsverben* in the lexicographical internet portal OWID to divide sets of semantically similar communication verbs into ever smaller sets of ever closer synonyms. *Kommunikationsverben* describes the meaning of communication verbs on two levels: a lexical level, represented in the dictionary entries and by sets of lexical features, and a conceptual level, represented by different types of situations referred to by specific types of verbs. The procedure starts at the conceptual level of meaning where verbs used to refer to the same specific situation type are grouped together. At the lexical level of meaning, the sets of verbs obtained from the first step are successively divided into smaller sets on the basis of the criteria of (i) identity of lexical meaning, (ii) identity of lexical features, and (iii) identity of contexts of usage. The stepwise procedure applied is shown to result in the creation of a semantic network for communication verbs.

Keywords: communication verbs; speech act verbs; synonymity; synonymy; conceptual field; semantic network; access structures; advanced search options

1 Kommunikationsverben in OWID

This contribution deals with the synonymy relations of German communication verbs and the way in which they were used to create a semantic network for these verbs in the *Handbuch deutscher Kommunikationsverben* (cf. Harras et al. 2004, Harras/Proost/Winkler 2007) and in its online version *Kommunikationsverben*, which has recently been integrated into the lexicographical internet portal OWID ('Online Wortschatz- und Informationssystem Deutsch') of the Institut für Deutsche Sprache. In both the print and the online reference work, the meaning of German communication verbs is described on two levels: a conceptual and a lexical level. The distinction between these two levels of meaning derives from two-levels-semantic (cf. Bierwisch & Lang 1989, Bierwisch & Schreuder 1992, Lang 1994). On the conceptual level, communication verbs are described as referring to different types of situations in which a speaker utters something to a hearer. The situation types referred to by communication verbs are represented as consisting of several components relating to the attitudes of the speaker, to properties of the propositional content of the speaker's utterance and to specific aspects of the situation in question. Verbs used to refer to the same specific situation type constitute a "paradigm" or conceptual field. On the lexical level, communication verbs belonging to the same field are differentiated with respect to their lexical meaning and their lexical features. Verbs which are identical with respect to their lexical meaning are subsumed under the same lemma and hence appear in the same dictionary entry. Of a set of verbs having the same lexical meaning, only one is lemmatised – usually the one which is least specific with respect to its contexts of usage – while the others are listed as synonyms of the verb lemmatised. *Kommunikationsverben* contains about 800 verbs, 241 of which are lemmatised and appear with an entry of their own. All other verbs are listed as synonyms of the verbs lemmatised. On the whole, *Kommunikationsverben* lists 170 fields of German communication verbs.

In *Kommunikationsverben* in OWID, the conceptual and the lexical level of the meaning of communication verbs have each been implemented in different types of access structures (cf. Müller-Spitzer & Proost 2013). Particularly, the online version provides some advanced search options allowing the user (i) to combine components of situations to "create" many different situation types and find the verbs matching them, and (ii) to search for verbs sharing a smaller or larger number of lexical features.

Since the conceptual and the lexical level of the meaning of communication verbs are each associated with different degrees of semantic specification, verbs grouped together on each of these two levels are synonymous to different degrees. In this contribution, we will show that the notion of the graded nature of synonymity may be used to divide sets of semantically similar communication verbs into ever smaller sets of increasingly closer synonyms, a procedure which ultimately results in the creation of a semantic network for communication verbs. By providing the two advanced search options, not available in the print version, the online version facilitates the user's access to the different degrees of similarity in meaning among synonymous communication verbs, thereby enhancing the structure of *Kommunikationsverben* as a semantic network.

2 Synonymity as a Graded Feature

Synonymy is a relation of similarity or identity of meaning among the senses of different lexical items (cf. Cruse 1986: 267; Cruse 2002: 486). Since similarity of meaning is a matter of degree, different types of synonymy relations have been distinguished, depending on the degree of similarity of the senses of the lexical items compared. Absolute synonymy involves complete identity of meaning and forms one end-point on the scale of synonymity (cf. Cruse 1986: 268). All other types of synonymy proposed encompass not only similarity of meaning, but also some degree of semantic difference between the senses of two or more lexical items. Difference in meaning is involved, for example, in the relation between propositional synonyms (e.g. *begin-commence*) and that between plesionyms or near-synonyms (e.g. *giggle-chuckle*), the difference between these two types of synonym being that substitution of one item by the other yields sentences with equivalent truth-conditions in the case of the former but not in that of the latter (cf. Cruse 1986: 270-289; Cruse 2002: 489-490). On the scale of synonymity, propositional synonymy occupies a position in between that of absolute synonymy and that of plesionymy. The latter shades off into non-synonymous difference of meaning, which constitutes the zero-point on the scale of synonymity (cf. Cruse 1986: 268).

3 The Meaning of Communication Verbs

3.1 Communication Verbs

Communication verbs are verbs used to refer to different types of situations in which a speaker (henceforth: S) utters something to a hearer (henceforth: H). In the default case, the speaker's utterance also contains a proposition (henceforth: P). Some but not all of these verbs lexicalise combinations of speaker attitudes such as the speaker's propositional attitude, i.e. the attitude of the speaker towards the proposition of his/her utterance, the speaker's intention and the speaker's presuppositions. This smaller set of communication verbs is called "speech act verbs" (cf. Proost 2006: 65; 2007: 8-9). Examples of German speech act verbs include *behaupten* ('assert'), *mitteilen* ('inform'), *lügen* ('lie'), *auffordern* ('demand'), *versprechen* ('promise'), *loben* ('praise'), *kritisieren* ('criticise'), *schimpfen* ('scold'), and *klagen* ('complain'). Examples of German communication verbs which are not part of the narrower set of speech act verbs in the sense outlined above are *sagen* ('to say'), *sprechen* ('to speak'), *brüllen* ('to scream'), *unterbrechen* ('to interrupt'), and *faxen* ('to fax'). *Kommunikationsverben* focuses on speech act verbs.

3.2 Representing the Meaning of Communication Verbs

3.2.1 The Conceptual Level of the Meaning of Communication Verbs

All situations referred to by communication verbs are characterised by the presence of four features or situational roles: a speaker, a hearer, a set of speaker attitudes, and an utterance (mostly) containing a proposition. Since these four elements are part of any situation referred to by communication verbs, they constitute the unifying feature of the meaning of these verbs (cf. Verschueren, 1980: 51-57; 1985: 39-40; Wierzbicka, 1987: 18; Harras et al. 2004: Introduction; Proost, 2006: 651). The type of situation referred to by all speech act verbs is therefore called the 'general resource situation type'.

Two of the roles of the general resource situation type, the role of the speaker attitudes and that of the utterance, may be specified in different ways. The role of the speaker attitudes may be specified as consisting of the speaker's attitude to the proposition of his/her utterance, the speaker's intention, and the speaker's presuppositions. The speaker's propositional attitude may be further specified as S's taking P to be true, S's knowing P, S's wanting P, S's evaluating P positively or negatively, and so on. Specifications of the speaker's intention include S's intention to make H recognise S's propositional attitude (for example, to make H recognise that S knows P or takes P to be true) or to get him/her to do something. The speaker's presuppositions may concern an attitude of H (whether H takes something to be true, whether he/she knows something), the interests of S and H concerning P (whether P is in the interest of S or in the interest of H), or properties of P (for example, whether P is the case). The role of the utterance is specified by properties of the propositional content. These include the event type of P (whether P is an action, event, or state of affairs), the temporal reference of P (whether P precedes, coincides with, or follows the time of S's uttering P) and, in the case that P is an action, the agent of P (S, H, S & H, and so on).

Different combinations of specifications of the different types of speaker attitudes and of the properties of the propositional content constitute special resource situation types. These are referred to by distinct types of verbs. For example, verbs like *mitteilen* ('inform'), *lügen* ('lie') and *loben* ('praise') and their synonyms are used to refer to the situation types characterised by the specifications listed in Tables 1-3:

Special Resource Situation Type: Representatives.Information.mitteilen	
Propositional Content (P)	
Event Type	not specified
Temporal Reference	not specified
Agent	not specified
Speaker Attitudes	
Propositional Attitude	S knows: P
Intention	S wants: H know: P
Presuppositions	H does not know: P

Table 1: Situation type referred to by mitteilen ('inform'), informieren ('inform'), instruieren ('advise') and unterrichten ('advise').

Special Resource Situation Type: Representatives.Assertives.lügen	
Propositional Content (P)	
Event Type	not specified
Temporal Reference	not specified
Agent	not specified
Speaker Attitudes	
Propositional Attitude	S does not take to be true: P
Intention	S wants: H recognise: S takes to be true: P
Presuppositions	H does not know: P

Table 2: Situation type referred to by lügen (‘lie’), schwindeln and flunkern (both ‘fib’) and their prefixed forms anlügen (‘lie to sb.’), belügen (‘lie to sb.’), erlügen (‘lie about sth.’), rumlügen (‘tell lies’), vorlügen (‘lie to sb. about sth.’), anflunkern (‘fib to sb.’), rumflunkern (‘tell fibs’), vorflunkern (‘fib to sb. about sth.‘) , anschwindeln (‘fib to sb.’), beschwindeln (‘fib to sb.’), rumschwindeln (,tell fibs‘), vorschwindeln (‘fib to sb. about sth.‘).

Special Resource Situation Type: Expressives.evaluative.positive.loben	
Propositional Content (P)	
Event Type	action
Temporal Reference	past
Agent	H of 3^{rd} person
Speaker Attitudes	
Propositional Attitude	S considers: P good
Intention	S wants: H recognise: S considers: P good
Presuppositions	P is the case

Table 3: Situation type referred to by loben (‘praise’), huldigen (‘pay tribute to’), ehren (‘honour’), würdigen (‘acknowledge’) and honorieren (‘appreciate’).

The combinations of the specifications of the speaker attitudes and of the properties of the propositional content lexicalised by verbs like *mitteilen*, *lügen* and *loben*, respectively, may also be conceived of as the concepts lexicalised by these verbs. Thus, *mitteilen* lexicalises the concept of a verbal action performed by a speaker who knows P and assumes that H does not know P with the intention that H know P, P being an action, event or state of affairs preceding, co-occurring with or following the time of S’s utterance. The information in Table 2 captures the idea that verbs like *lügen* express the concept of a verbal action whereby a speaker who does not take P to be true and assumes that H does not know P intends the hearer to recognise that he/she – i.e. the speaker – takes P to be true. The verb *loben* lexicalises the concept of a verbal action performed by a speaker who evaluates P, a past action by H or a 3^{rd} person, positively and intends the hearer to recognise this.

Verbs which are used to refer to the same special resource situation type constitute a "paradigm" or conceptual field. With respect to the examples in Tables 1-3, this means that the sets {*mitteilen, informieren, instruieren, unterrichten*}, {*lügen, anlügen, belügen, erlügen, rumlügen, vorlügen, flunkern, anflunkern, schwindeln, anschwindeln, beschwindeln, rumflunkern*} and {*loben, huldigen, ehren, würdigen, honorieren*} each represent a conceptual field.

3.2.2 Methods Used to Describe the Conceptual Level of the Meaning of Communication Verbs

Following a procedure proposed by Baumgärtner (1977: 260-264), the different specifications of the role of the speaker attitudes and the role of the utterance as well as the lower-level specifications of each of these are obtained from a comparison of sentences containing speech act verbs. The well-formedness of some of these and the ill-formedness of others show which elements are relevant to the meaning of the verbs they contain. For example, a comparison of the sentences in (1) and (2) shows that *to order* lexicalises the values 'future', 'action' and 'hearer' for the specifications of the temporal reference, the event type and the agent of P, respectively, while *to promise* lexicalises the values 'future', 'action' and 'speaker', respectively, for these specifications:

(1) a. I order youi to PROi leave the room.
b. *I order youi to PROi have left the room.
c. *I order youi for mej to PROj leave the room.

(2) a. Ii promise you to PROi leave the room.
b. *Ii promise you to PROi have left the room.
c. *Ii promise youj to PROj leave the room.

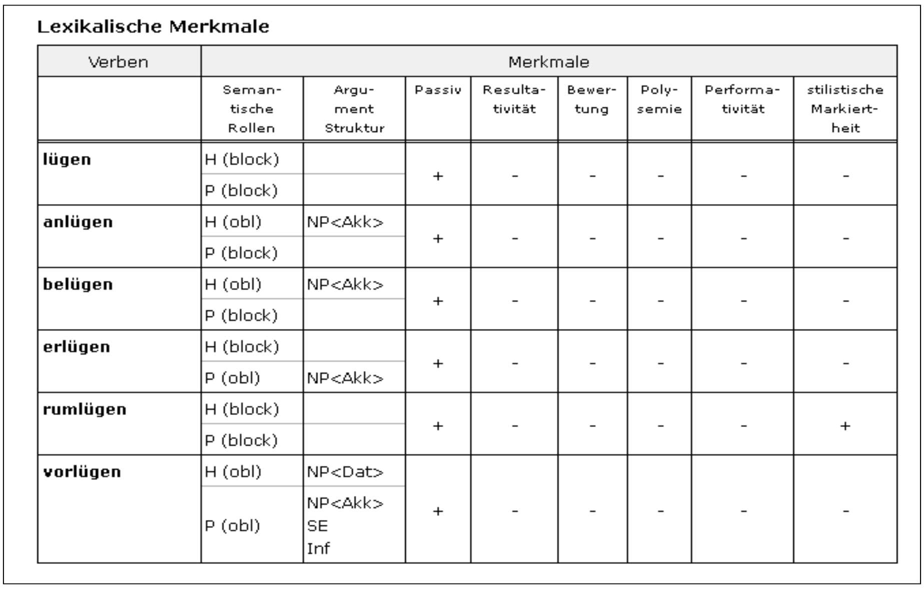
The introspective analysis exemplified in (1)-(2) has shown that the higher-level specifications of the speaker's propositional content, the speaker's intention, the speaker's presuppositions and the propositional content are essential aspects of the meaning of speech act verbs. These four aspects correspond to five of the seven components of illocutionary force which Searle & Vanderveken (1985: 12-20) and Vanderveken (1990: 103-136) have argued to determine the conditions under which a particular type of speech act is both successful and non-defective. Particularly, the aspect of the speaker's propositional attitude corresponds to the component of the sincerity conditions, the aspect of the speaker's intention to the component of the illocutionary point, the aspect of the speaker's presuppositions to the components 'mode of achievement of the illocutionary point' and 'preparatory conditions', and the aspect of the propositional content to the component of the propositional content conditions (cf. Harras 2001: 26-31, Proost 2006: 654-655).

While the higher-level specifications of the speaker's propositional attitude, the speaker's intention, the speaker's presuppositions and the propositional content are obtained from the type of analysis exemplified in (1)-(2), the lower-level specifications of each of these are calculated systematically, i.e. irrespective of any existing lexicalisations. For example, the specification 'temporal reference of P' is assumed to have the specifications 'Past', 'Present' and 'Future', the specification of the event type of P the specifications 'action', 'state' and 'event', and so on. The question of which values are lexicalised by a particular verb was decided on the basis of examples from the Mannheim German Reference Corpus DeReKo ("Deutsches Referenzkorpus"). Methodological issues are dealt with in detail in the introductions to both volumes of the Handbuch deutscher Kommunikationsverben (cf. Harras et al. 2004, Harras 2007), which are also available in the online version.

3.2.3 The Lexical Level of the Meaning of Communication Verbs

Verbs which belong to the same conceptual field but differ from each other with respect to their lexical meaning appear with an entry of their own. In the *Handbuch deutscher Kommunikationsverben* and its online version in OWID, lexical meanings were differentiated on the basis of examples from the IDS-corpora of written German. All other verbs are listed as synonyms of the verbs lemmatised. With respect to the *lügen*-field, this means that *lügen* ('lie') and *schwindeln* ('fib') each appear with a separate entry. These verbs differ from each other in that *schwindeln* but not *lügen* expresses an evaluation by a discourse situation speaker, i.e. a speaker who uses this verb to comment on the utterance of the resource situation speaker. Particularly, a speaker who uses the verb *schwindeln* to refer to the resource situation speaker's act of lying thereby indicates that he/she does not consider S's act of lying to have serious consequences for H. In the *Handbuch deutscher Kommunikationsverben* and its online version *Kommunikationsverben* in OWID, this difference in the lexical meaning of *lügen* and *schwindeln* is reflected by the meaning paraphrases of these verbs in their respective entries. Since the evaluation expressed by *schwindeln* is an evaluation by a discourse situation speaker, it is not an element of the resource situation referred to by this verb. Hence, within the framework of *Kommunikationsverben*, it is not part of the conceptual component of its meaning. Rather, it is an essential part of the lexical meaning of this verb. With respect to the *lügen*-field, this means that this contains two lemmatised verbs, *lügen* and *schwindeln*. The verb *flunkern* is subsumed under the lemmatised verb *schwindeln,* because it has the same lexical meaning.

Verbs which belong to the same field *and* have the same lexical meaning are differentiated with respect to the following lexical features: (i) expression of thematic roles, (ii) syntactic realisation of thematic roles, (iii) passivisation, (iv) resultativity, (v) evaluation by a discourse situation speaker (a speaker describing the speech act performed by the reference situation speaker), (vi) polysemy, (vii) performativity (the possibility for a verb to be used performatively), and (viii) stylistic markedness. Each member of a field is characterised as possessing or lacking each of these features as exemplified for the verb *lügen* und its prefixed forms *anlügen*, *belügen*, *rumlügen* and *vorlügen* by the screenshot in Figure 1:



Lexikalische Merkmale

Verben	Merkmale							
	Semantische Rollen	Argument Struktur	Passiv	Resultativität	Bewertung	Polysemie	Performativität	stilistische Markiertheit
lügen	H (block)		+	-	-	-	-	-
	P (block)							
anlügen	H (obl)	NP<Akk>	+	-	-	-	-	-
	P (block)							
belügen	H (obl)	NP<Akk>	+	-	-	-	-	-
	P (block)							
erlügen	H (block)		+	-	-	-	-	-
	P (obl)	NP<Akk>						
rumlügen	H (block)		+	-	-	-	-	+
	P (block)							
vorlügen	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE Inf						

Fig. 1: Lexical Features of lügen ('lie') and its prefixed forms. "block" ("blocked") means that the thematic role in question either cannot be realised at all or can be realised only by a prepositional phrase headed by vor ('before') or by an adpositional phrase headed by gegenüber ('in front of'); "obl" (obligatory") means that the thematic role in question must be realised.

In addition to being differentiated with respect to their lexical features, verbs with the same lexical meaning may be distinguished with respect to their typical contexts of usage. Information on the range of contexts the non-lemmatised verbs may occur with is provided in the section *Kommentar* ('commentary') in the dictionary entry of the corresponding lemmatised verb. *Schwindeln* and *flunkern*, for example, are identical with respect to the specific type of situation they are used to refer to and regarding their lexical meaning but differ with respect to the contexts in which they are typically being used. Particularly, *flunkern* is used more frequently than *schwindeln* when reference is made to situations involving children telling lies, as illustrated in (1):

(1) Fast jeder fünfte Schüler (19 Prozent) verschweigt seinen Eltern schlechte Noten. 32 Prozent der Kids flunkern, wenn es allgemein um das Thema Schule geht. [Frankfurter Rundschau, 03.02.1999] ['Almost every fifth pupil (19 Percent) keeps quiet to his parents about bad marks. 32 percent of the kids fib when the topic school is dealt with in general.']

Because of this restriction on the range of contexts in which it is typically used, *flunkern* is not lemmatised and hence does not appear with an entry of its own. Rather, it is mentioned as a synonym of the verb *schwindeln*, which is less restricted then *flunkern* with respect to its contexts of usage and is therefore lemmatised and appears with an entry of its own.

4 Criteria for the Synonymy of Communication Verbs

As shown in the previous section, both the *Handbuch deutscher Kommuniationsverben* and its online version *Kommunikationsverben* in OWID describe communication verbs on different levels of analytical detail. Communication verbs may be grouped together on each of these levels. Depending on the analytical level on which they are grouped together, communication verbs may be regarded as being synonymous in either a broader or a narrower sense. As an illustration of how the different criteria apply, they will be explained with respect to the verbs of the *lügen*-field, which has been introduced in the previous section. Additional examples will be discussed in section 5.

4.1 The Criterion for Synonymy in the Broader Sense: Membership in the Same Field

On the lowest level of specification, verbs which are used to refer to the same special resource situation type and hence constitute a field in the sense outlined in section 3.2.1 may be regarded as synonyms in a broader sense. The corresponding criterion for synonymy on this level is membership within the same conceptual field. This means that all of the verbs mentioned underneath Tables 1-3 are synonyms in a broader sense. The degree of synonymity relating these lexical items is low. Membership within the same conceptual field is the minimum requirement for communication verbs to be considered synonyms at all. All other criteria for synonymy concern the lexical level of meaning and/or restrictions of usage. Verbs grouped together by these criteria are synonyms in a narrower sense.

4.2 Criteria for Synonymy in the Narrower Sense

4.2.1 Identity of Lexical Meaning

The first criterion relating to the lexical level is identity of lexical meaning. When applied to the verbs of the *lügen*-field, this criterion groups together *schwindeln* and *flunkern* as synonyms, distinguishing them from *lügen* by virtue of the fact that they both express an evaluation by a discourse situation speaker not part of the meaning of the latter verb (see section 3.2.3).

4.2.2 Number of Shared Lexical Features

The degree of synonymity of communication verbs is additionally determined by the number of their shared lexical features. For example, *anlügen* and *belügen* are identical regarding all lexical features including their argument structure properties (see Figure 1). By contrast, *anlügen* and *belügen* on the one hand and *lügen* on the other differ in their argument structure properties while being identical with respect to all other lexical features. Specifically, *lügen* blocks the realisation of the roles of H and P while *anlügen* and *belügen* both obligatorily realise the role of the hearer as an NP in the accusative case and block the realisation of the role of P (see Figure 1). Due to these differences in the argument structure properties of *anlügen* and *belügen* on the one hand and *lügen* on the other, the degree of synonymity between the former two verbs is higher than that between either of them and *lügen*.

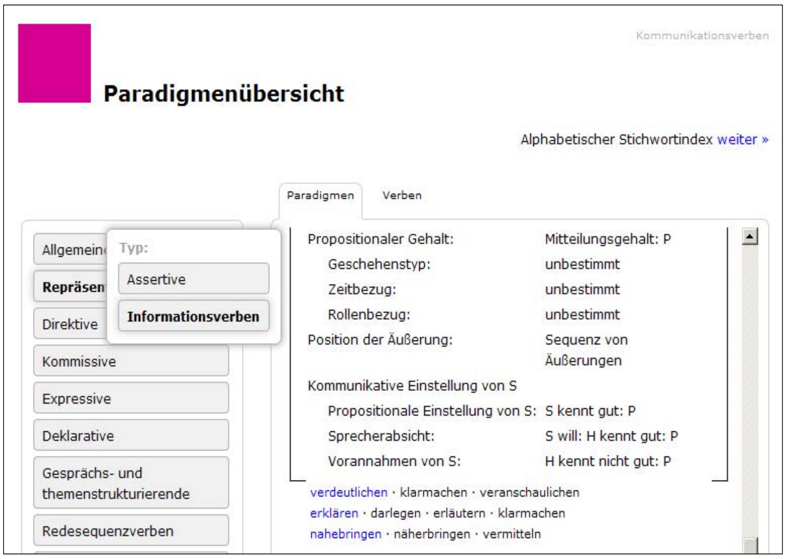
4.2.3 Substitutability *salva veritate*

The verbs *schwindeln* and *flunkern* are identical with respect to (i) the specific type of situation they are used to refer to, (ii) their lexical meaning, and (iii) their lexical features. As discussed in section they differ merely with respect to the contexts in which they are typically used: *flunkern* is used more frequently than *schwindeln* when reference is made to situations involving children telling lies as illustrated by example (1). Since substitution of *flunkern* by *schwindeln* in (1) does not yield a sentence with different truth-conditions, *flunkern* and *schwindeln* are substitutable *salva veritate*. Substitutability *salva veritate* is commonly regarded as an essential condition for propositional or cognitive synonymy. For a more detailed discussion of this particular type of synonymy, see Harras (2007b: 329-365).

5 Some Applications

5.1 Example I: Representatives of the Type 'verdeutlichen' ('explain')

Different degrees of synonymity may also be observed among the verbs of the field containing the lemmatised verbs *verdeutlichen* ('explain'), *erklären* ('explain') and *nahebringen* ('bring sth. home to sb.'). These verbs and the synonyms of each of these all express the concept of a verbal action whereby a speaker who knows something (P: a past, present or future action, event or state of affairs) well and assumes that H does not have sufficient knowledge of P makes several utterances to make H know P well. In *Kommunikationsverben* in OWID, information about special resource situation types is represented in the section *Paradigmenübersicht* ('overview of paradigm'). The screenshot in Figure 2 represents the special resource situation type referred to by *verdeutlichen*, *erklären* and *nahebringen* and the synonyms of each of them:



Propositionaler Gehalt:	Mitteilungsgehalt: P
Geschehenstyp:	unbestimmt
Zeitbezug:	unbestimmt
Rollenbezug:	unbestimmt
Position der Äußerung:	Sequenz von Äußerungen
Kommunikative Einstellung von S	
Propositionale Einstellung von S:	S kennt gut: P
Sprecherabsicht:	S will: H kennt gut: P
Voraannahmen von S:	H kennt nicht gut: P

verdeutlichen · klarmachen · veranschaulichen
erklären · darlegen · erläutern · klarmachen
nahebringen · näherbringen · vermitteln

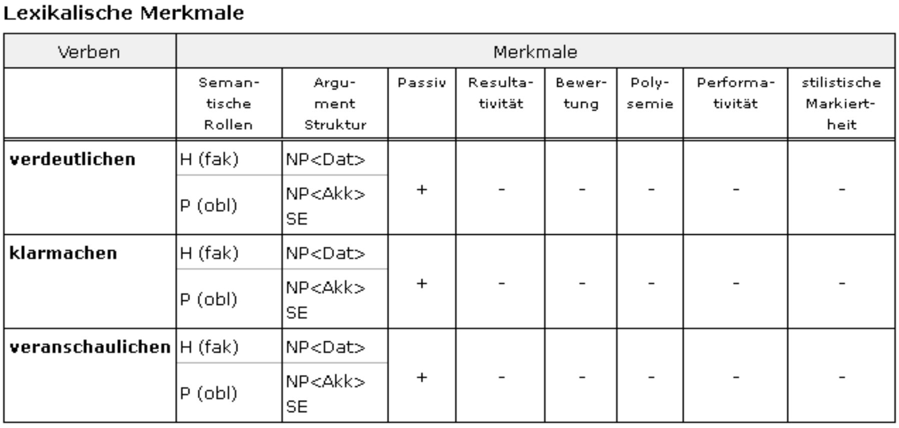
Fig. 2: Situation type referred to by verdeutlichen (‚explain'), erklären (‚explain') and nahebringen ('bring sth. home to sb.') and their synonyms.

Since the verbs *verdeutlichen* (‚explain‘), *klarmachen* (‚make sth. clear to sb.‘), *veranschaulichen* (‚illustrate‘), *erklären* (‘explain‘), *darlegen* (‘explain‘), *erläutern* (‘explain’), *nahebringen* (‘bring sth. home to sb.’), *näherbringen* (‘bring sth. home to sb.’) and *vermitteln* (‘pass on knowledge’) are all used to refer to the same special resource situation type, they are synonyms in a broader sense.

On the lexical level of meaning, *verdeutlichen*, *erklären* and *nahebringen* are differentiated on the basis of their lexical meaning as follows:

- *verdeutlichen*: ‘to make sth. more comprehensible; to explain the crucial aspects of an issue or problem to sb. in order to make that person understand this issue or problem well’. Since *klarmachen* and *veranschaulichen* have the same lexical meaning as *verdeutlichen*, these three verbs are synonyms in a narrower sense.
- *erklären*: ‘to represent difficult and/or complex facts exactly and comprehensibly to sb. in order to make that person understand them well’. The verbs *darlegen*, *erläutern* and *klarmachen* are listed as having the same lexical meaning. *Erklären*, *darlegen*, *erläutern* and *klarmachen* may therefore be regarded as synonyms in a narrower sense.
- *nahebringen*: ‘to make sb. familiar with sth., usually with knowledge concerning a specific field, in order to arouse that person’s interest’. Since *näherbingen* and *vermitteln* are listed as having the same lexical meaning, these two verbs and *nahebringen* may be considered synonyms in a narrower sense.

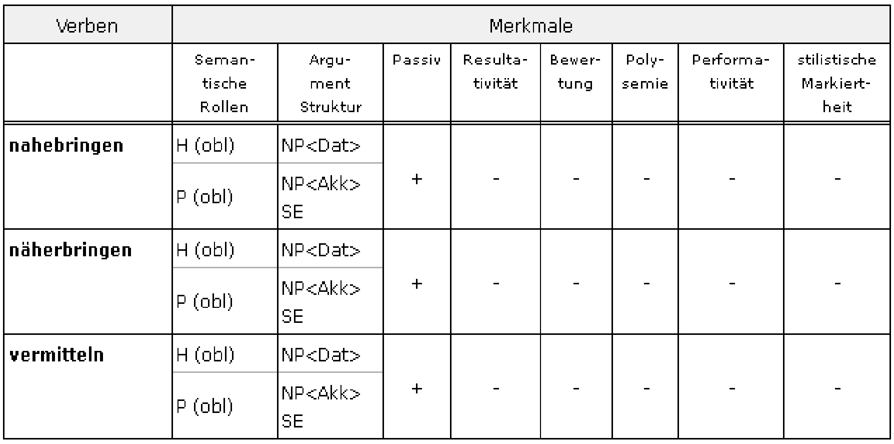
On a more detailed level of analysis, verbs which are synonymous in as far as they have the same lexical meaning may be further differentiated by their lexical features. As indicated by Figures 3 and 4, *verdeutlichen*, *klarmachen* and *veranschaulichen* on the one hand and *nahebringen*, *näherbingen* and *vermitteln* on the other are completely identical with respect to all of their lexical features, including the syntactic realisation of their arguments:



Lexikalische Merkmale

Verben	Merkmale							
	Semantische Rollen	Argument Struktur	Passiv	Resultativität	Bewertung	Polysemie	Performativität	stilistische Markiertheit
verdeutlichen	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
klarmachen	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
veranschaulichen	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						

Fig. 3: Lexical features of verdeutlichen(‘explain‘), klarmachen (‘make sth. clear to sb.‘) and veranschaulichen (‘illustrate’).

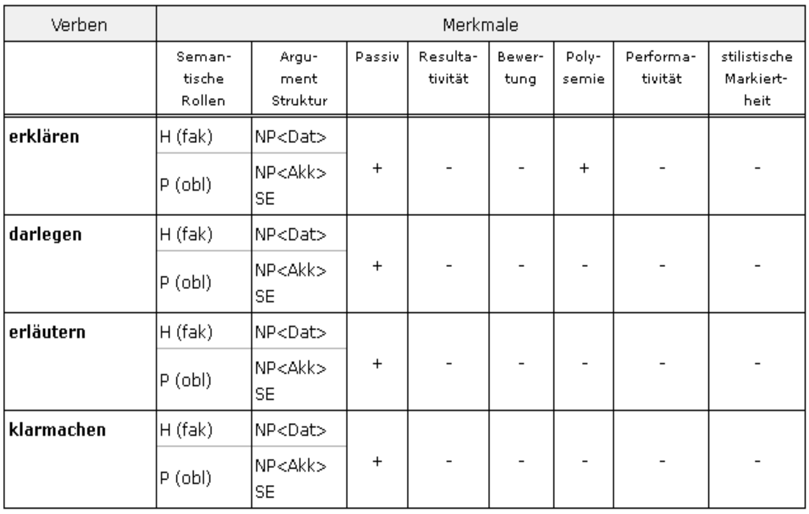


Verben	Merkmale							
	Semantische Rollen	Argument Struktur	Passiv	Resultativität	Bewertung	Polysemie	Performativität	stilistische Markiertheit
nahebringen	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
näherbringen	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
vermitteln	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						

Fig. 4: Lexical features of nahebringen, näherbringen (both ‘bring sth. home to sb.‘) and vermitteln (‘pass on knowledge’).

Since all of the verbs mentioned in Figures 3 and 4 are identical with respect to all of their lexical features, they are very close synonyms.

Within the set {*erklären, darlegen, erläutern, klarmachen*}, *darlegen, erläutern* and *klarmachen* are also identical with respect to all of their lexical features. Like the verbs of the two other sets mentioned above, these three verbs are therefore synonymous to a very high degree. Since *erklären* differs from each of the three other verbs in that it is polysemous while the others are not, the degree of synonymity between this verb and each of the other three verbs is lower than among the other three verbs. Figure 5 lists the lexical features of *erklären* and its synonyms *darlegen, erläutern* and *klarmachen*:

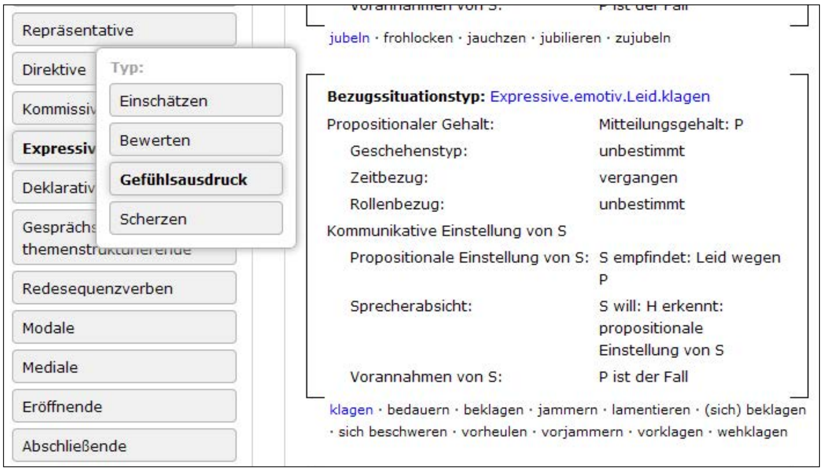


Verben	Merkmale							
	Semantische Rollen	Argument Struktur	Passiv	Resultativität	Bewertung	Polysemie	Performativität	stilistische Markiertheit
erklären	H (fak)	NP<Dat>	+	-	-	+	-	-
	P (obl)	NP<Akk> SE						
darlegen	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
erläutern	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
klarmachen	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						

Fig. 5: Lexical features of erklären (‘explain‘), darlegen (’explain‘), erläutern (‘explain‘) and klarmachen (‘make sth. clear to sb.’)

5.2 Example II: Expressives of the Type 'klagen' ('complain').

None of the verbs which are part of the field represented in Figures 2-5 share any special contextual restrictions on the basis of which they may be claimed to be even closer synonyms. Synonymy relations of this kind may be observed from a comparison of the contextual restrictions associated with the use of the verbs *klagen* ('complain'), *jammern* ('moan') and *lamentieren* ('lament'). These verbs are used to refer to situations in which a speaker who feels sorrow because of something (P: a past action, event or state of affairs) makes one or more utterances with the intention that the hearer recognize that he/she, i.e. the speaker, feels sorrow because of P. The situation referred to by *klagen* and its synonyms is represented by the screenshot in Figure 6:



jubeln · frohlocken · jauchzen · jubilieren · zujubeln

Bezugssituationstyp: Expressive.emotiv.Leid.klagen

Propositionaler Gehalt:	Mitteilungsgehalt: P
Geschehenstyp:	unbestimmt
Zeitbezug:	vergangen
Rollenbezug:	unbestimmt
Kommunikative Einstellung von S	
Propositionale Einstellung von S:	S empfindet: Leid wegen P
Sprecherabsicht:	S will: H erkennt: propositionale Einstellung von S
Vorannahmen von S:	P ist der Fall

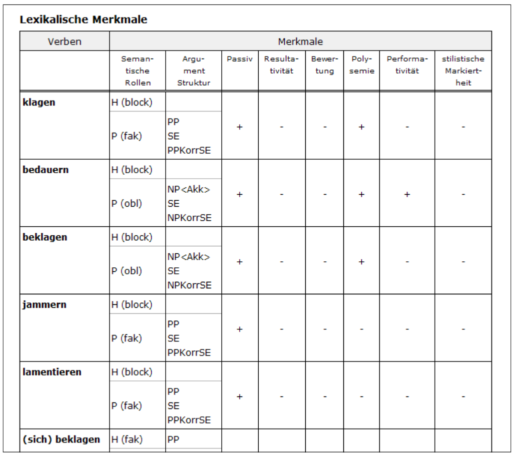
klagen · bedauern · beklagen · jammern · lamentieren · (sich) beklagen · sich beschweren · vorheulen · vorjammern · vorklagen · wehklagen

Fig. 6: Situation referred to by klagen ('complain') and its synonyms.

Zooming in, for the sake of brevity, on the verbs *klagen*, *jammern* and *lamentieren*, these verbs differ regarding the contexts in which they are typically used. Though *klagen* and *lamentieren* may be used in most of the contexts in which *klagen* is used, *klagen* is used more frequently in combination with expressions designating diseases:

(2) Seltener wird über Kopfschmerzen geklagt. ['People rarely complain about a headache.']
(3) ?Seltener wird über Kopfschmerzen gejammert. ['People rarely moan about a headache.']
(4) ?Seltener wird über Kopfschmerzen lamentiert. ['People rarely lament about a headache.']

To the extent that *jammern* and *lamentieren* are less restricted with respect to their typical contexts of usage than *klagen*, the degree of synonymity between them is higher than that between either of them and *klagen*. As indicated by Table 7, the degree of synonymity between *jammern* and *lamentieren* is also higher than that between either of them and *klagen* by virtue of the fact that the former two verbs are identical with respect to all lexical features while *klagen* differs from *jammern* and *lamentieren* in that it is polysemous, which the latter two verbs are not:



Lexikalische Merkmale

Verben	Merkmale							
	Seman-tische Rollen	Argu-ment Struktur	Passiv	Resulta-tivität	Bewer-tung	Poly-semie	Performa-tivität	stilistische Markiert-heit
klagen	H (block)		+	-	-	+	-	-
	P (fak)	PP SE PPKorrSE						
bedauern	H (block)		+	-	-	+	+	-
	P (obl)	NP<Akk> SE NPKorrSE						
beklagen	H (block)		+	-	-	+	-	-
	P (obl)	NP<Akk> SE NPKorrSE						
jammern	H (block)		+	-	-	-	-	-
	P (fak)	PP SE PPKorrSE						
lamentieren	H (block)		+	-	-	-	-	-
	P (fak)	PP SE PPKorrSE						
(sich) beklagen	H (fak)	PP						

Fig. 7: Lexical features of klagen ('complain'), jammern ('moan') and lamentieren ('lament').

6 Conclusion: The Creation of a semantic network

The procedure whereby sets of verbs used to refer to the same special resource situation type are divided in a stepwise fashion into ever smaller sets of ever closer synonyms ultimately results in the creation of a semantic network for communication verbs. Fig. 8 represents the section of this network comprising representatives of the type 'verdeutlichen' ('explain'):

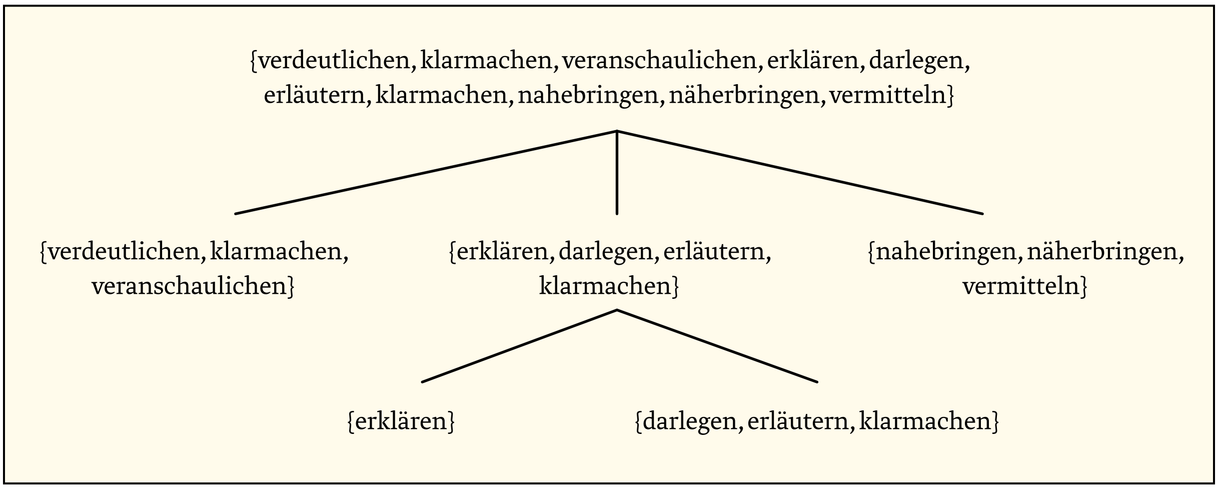

{verdeutlichen, klarmachen, veranschaulichen, erklären, darlegen, erläutern, klarmachen, nahebringen, näherbringen, vermitteln}
{verdeutlichen, klarmachen, veranschaulichen}
{erklären, darlegen, erläutern, klarmachen}
{nahebringen, näherbringen, vermitteln}
{erklären}
{darlegen, erläutern, klarmachen}

Fig. 8: Section of the network for communication verbs comprising representatives of the type 'verdeutlichen' ('explain').

Searching for verbs with varying degrees of synonymity is significantly facilitated by the online version, which provides two advanced search options allowing the user to automatically search for verbs sharing a smaller or larger number of conceptual and/or lexical features by selecting them from an input mask.

7 References

Barwise, J. & Perry, J. (1983). *Situations and Attitudes*. Cambridge/MA: The MIT Press.

Baumgärtner, K. (1977). Lexikalische Systeme möglicher Performative. In *Zeitschrift für Germanistische Linguistik*, 5, pp. 257-276.

Bierwisch, M. & Lang, E. (1989). Somewhat Longer – Much Deeper – Further and Further: Epilogue to the Dimensional Adjective Project. In M. Bierwisch & E. Lang (eds.) *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation*. Springer Series in Language and Communication; 26. Berlin: Springer, pp. 471-514.

Bierwisch, M. & Schreuder, R. (1992). From Concepts to Lexical Items. In *Cognition*, 42, pp. 23-46.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: CUP.

Cruse, D. A. (2002). Paradigmatic relations of inclusion and identity III: Synonymy. In D. A. Cruse, F. Hundsnurscher, M. Job, P. R. Lutzeier (eds.) *Lexikologie-Lexicology: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen – An international handbook on the nature and structure of words and vocabularies*. Vol. 2. Berlin/New York: Walter de Gruyter, pp. 485-497.

DeReKo - Das Deutsche Referenzkorpus. http://www1.ids-mannheim.de/kl/projekte/korpora/

Harras, G. (2001). Performativität, Sprechakte und Sprechaktverben. In G. Harras (ed.) *Kommunikationsverben: Konzeptuelle Ordnung und Semantische Repräsentation*. Studien zur deutschen Sprache; 24. Tübingen: Narr, pp. 11-32.

Harras, G. (2007a). Lexikalische Strukturen der Repräsentative. In G. Harras, K. Proost, E. Winkler *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: de Gruyter, pp. 73-124.

Harras, G. (2007b). Partielle und totale Synonymie von Sprechaktverben. In G. Harras, K. Proost, E. Winkler *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: de Gruyter, pp. 329-365.

Harras, G., Winkler, E., Erb, S. & Proost, K. (2004): *Handbuch deutscher Kommunikationsverben. Teil I: Wörterbuch*. Schriften des Instituts für Deutsche Sprache; 10.1. Berlin/New York: de Gruyter.

Harras, G., Proost, K. & Winkler, E. (2007): *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: de Gruyter.

Kommunikationsverben. Electronic version of the *Handbuch deutscher Kommunikationsverben*. Adaptation for the online version by Kristel Proost. Accessed at: http://www.owid.de/wb/komvb/start.html. [06/04/2014].

Proost, K. (2006). Speech Act Verbs. In K. Brown (ed.-in-chief) *Encyclopedia of Language & Linguistics*. 2nd ed. Vol XI. Oxford: Elsevier, pp. 651-656.

Proost, K. (2007). Conceptual Structure in Lexical Items: The Lexicalisation of Communication Concepts in English, German and Dutch. Pragmatics & Beyond New Series; 168. Amsterdam/Philadelphia: Benjamins.

Lang, E. (1994). Semantische vs. konzeptuelle Struktur: Unterschneidung und Überschneidung. In M. Schwarz (ed.) *Kognitive* Semantik: *Ergebnisse, Probleme, Perspektiven*. Tübinger Beiträge zur Linguistik; 395. Tübingen: Narr, pp. 9-40.

Müller-Spitzer, C., Proost, K. (2013): Kommunikationsverben in OWID: An Online Reference Work with Advanced Access Structures. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituu, pp. 296-309.

Searle, J. R. & Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge, UK: Cambridge University Press.

Vanderveken, D. (1990). *Meaning and Speech Acts I: Principles of Language Use*. Cambridge, UK: Cambridge University Press.

Verschueren, J. (1980). *On Speech Act Verbs*. Amsterdam: John Benjamins.

Verschueren, J. (1985). What people say they do with words: Prolegomena to an empirical-conceptual approach to linguistic action. Norwood, NJ: ABLEX.

Wierzbicka, A. (1987). *English speech act verbs: A semantic dictionary*. Sydney: Academic Press.

Job-hunting in Italy: Building a glossary of “English-inspired” job titles

Virginia Pulcini, Angela Andreani[1*]
Università degli Studi di Torino
virginia.pulcini@unito.it, angela.andreani@unito.it

Abstract

This paper reports on a study of “English-inspired” job titles retrieved from a specialized corpus of job advertisements posted on Italian web pages. This corpus was created using the WebBootCat tool in the Sketch Engine, following the methodology described by Baroni and Bernardini (2004) and Baroni et al. (2006). The aim is to build a glossary of English job titles to be published online as a tool for prospective job applicants. Checking their status in English and Italian dictionaries, we will establish whether the titles collected are current English terms, false Anglicisms, or “English-inspired” creations. The preliminary findings consist of a list of 30 job titles which are analyzed in terms of form and meaning, and grouped into categories depending on whether an Italian equivalent is available or not. The corpus of job postings is used to analyze the lexical profile of job titles, their meaning and/or possible covert manipulative intent. In fact, data shows that some English job titles may be preferred to Italian equivalents to attribute greater status to the actual job designation and description. Moreover, some job titles are characterized by complex pre-modification which may confuse the ultimate users, i.e. job hunters themselves.

Keywords: job title; Anglicism; Anglicization

1 Introduction and research aims

Owing to the process of internationalization and globalization of business and trade, the job market is one of the many areas in which the influence of the English language is quite strong. A growing number of multinational companies have adopted English as a company language and most of them use English as a lingua franca on a regular basis for business communication. An emblematic case is the recent transformation of the historic Turin-based FIAT car company into a multinational through the merger with the American Chrysler and its adoption of a new “non-Italian” name – FCA – which stands for Fiat-Chrysler Automobiles. By the same token, small and medium-sized enterprises, even though operating domestically but aspiring to expand beyond national borders, also find it advisable to take on an international profile by using English for branding and product advertising. Today a

[1] * Both authors are responsible for the overall planning of this research. V. Pulcini drafted sections 1, 2, 3, and 4. A. Andreani drafted sections 3.1 (3.1.1, 3.1.2, 3.1.3), 3.2, 3.3 (3.3.1, 3.3.2).

good level of competence in English is an indispensable asset to hold a high-level job in the world of business. A working knowledge of spoken and written English is normally requested also for lower level occupations such as technical and clerical jobs, as emerges from advertisements in the national and international job market.

The key role of English in professional settings has greatly enhanced its desirability as a foreign language to learn. As a result, today there are more learners of English and competent non-native speakers of English than ever before, and the vocabularies of many languages have adopted a large stock of English words and terms, especially in specialized domains (Furiassi et al. 2012). The use of English is dictated primarily by practical reasons but also because non-native speakers have a favourable perception of it. As pointed out by Pulcini:

> What is crucial in favouring the adoption of English loanwords are speakers' positive attitudes towards Anglicisms [...]. For better or for worse, the English language enjoys status and prestige, and Anglicisms are perceived by most speakers as modern, dynamic, fashionable and are thought to convey a higher level of competence and professionalism. (Pulcini et al. 2012: 16)

This study focuses on the influence of English on the designation and description of job titles in Italy, which appears to be a widespread and growing phenomenon in all non-English-speaking countries. Previous research carried out by Van Meurs et al. (2006; 2011) on job advertisements in the Netherlands has highlighted and described by means of quantitative data several particular aspects of the presence of English in Dutch job postings. For example, the use of English is greater in adverts posted by multinational companies and organizations than by domestic ones. English is more pervasive in ads for higher-level and academic jobs rather than medium-level vacancies and more frequent in specific domains such as transport, storage, communications and commerce as compared to, for example, the financial sector. Van Meurs et al. (2014) have also studied the perception of English loanwords with respect to Dutch equivalents in job advertisements, showing that English and Dutch terms have different associative meanings in the minds of the users. Another study carried out by Taavitsainen and Pahta (2003) points out that English is mandatory for recruitment in Finnish companies, and many Nordic companies have chosen English as their official language, abandoning their domestic names in favour of "English-inspired" ones in order to favour internationalization and, at the same time, sound young, modern and trendy. As for job postings, sometimes they are written entirely in English, and the use of English is quite frequent in vacancies for Scandinavian and Finnish companies, as well as for Swiss ones, as pointed out by Watts (2002). Taavitsainen and Pahta mention a recent campaign at the University of Helsinki against the use of "this odd form of business jargon", arguing that English job titles "blur the job description and unnecessarily mystify functions in the business world." (Taavitsainen & Pahta 2003: 8)

The research study illustrated in this paper is part of a wider project focussed on the influence of the English language in Italy[2], including its impact on the world of business. The focus is on the use of English or "English-inspired" job titles retrieved from a corpus of job advertisements posted on Italian web pages. The aim of this research is to build a glossary of English-looking job titles to be published online as a tool for job hunters in Italy. Using dictionaries and corpora in order to observe the lexical profile of these job titles, we will try and establish which of these are current Anglicisms, false Anglicisms, or "English-inspired" creations. We will argue that some terms are rather opaque to the Italian user and their adoption is motivated by the intention to give "higher status" to a particular job or to camouflage its real nature and thus confuse or deceive the prospective applicant.

2 Methodology

The collection of English-looking job titles began with a preliminary survey of the websites of some Italian online job finding agencies[3] and of the websites of the Italian branch of some multinational human resource consulting companies.[4] On the websites, the user can select, among other options, a professional category (*categoria professionale* or *funzione aziendale,* e.g. retail, HR, banking), an industry sector (*settore*), and, in some instances, a specific role or job position (*mansione* or *funzione aziendale*, e.g. receptionist). The dropdown menus often include, alongside Italian ones, professional categories and functions already in English, which formed our preliminary list of English job titles.[5]

This was then expanded by querying a domain-specific corpus of Italian job advertisements, which we built using the WebBootCat tool in the Sketch Engine (Kilgarriff et al. 2004). Drawing on the methodology described in Baroni and Bernardini (2004) and Baroni et al. (2006), we selected a number of seedwords from among the most frequent terms and phrases in job postings: *annunci di lavoro*; *offerte di lavoro*; *si offre*; *si propone*; *si richiede*; *annuncio*; *lavoro*; *azienda*; *contratto*; *candidato*; *settore*; *profilo*; *esperienza*; *competenze*.[6] The corpus was then compiled using the TreeTagger for Italian (Baroni's model) and opened in the Sketch Engine to compare it with the itTenTen10 corpus and extract further key terms to be used as seeds. The procedure was iterated twice, and then repeated at approximately three

[2] The project is "The English language in Italy: linguistic, educational and professional challenges", promoted by the University of Turin in conjunction with the *Compagnia di San Paolo* (2013-2015) and coordinated by Virginia Pulcini. www.englishinitaly.wordpress.com

[3] Accessed at http://www.adhr.it; http://www.alispa.it; http://www.carrieraefuturo.com; http://www.eurointerim.it; http://www.gigroup.it; http://www.humangest.it; http://www.obiettivolavoro.it; http://www.orienta.net; http://it.quanta.com; http://www.umana.it/it-IT/home-page [13/10/2013]

[4] Accessed at http://www.adecco.it; http://www.manpower.it; http://www.randstad.it; http://www.synergie-italia.it [13/10/2013]

[5] Data entry, development engineer, hostess, order entry, promoter, receptionist, telemarketer, visual merchandiser, web designer.

[6] The total number of seeds, 14, was set following Baroni and Bernardini: "For well-defined specialized domains, a small list of seeds (in the 5-to-15 range) is typically sufficient" (2004: 1314). The additional parameters (tuple size, minimal and maximal file size, max URLs per query, etc.) were set according to the default settings of the WebBootCat in the Sketch Engine.

weeks' distance, obtaining a final corpus of 241,021 tokens.[7] The corpus was queried to retrieve additional English or English-looking job titles in context.

3 Preliminary Findings

The preliminary findings consist of a list of 30 job titles which are analyzed in terms of form, meaning and Italian equivalents in English and Italian general and specialized dictionaries and in our corpus. The English dictionaries considered are the *Collins English Dictionary* online (CED) and the *Cambridge Business English Dictionary* online (CBED); the Italian dictionaries are *Zingarelli 2014* (ZING) and the bilingual encyclopaedic dictionary *Economics&Business* (Picchi 2011, henceforth E&B).

Table 1 shows the attestation of the terms in the reference dictionaries. The cells highlighted in grey indicate Anglicisms with a current Italian equivalent. Items in italics are dictionary headwords that slightly diverge in form from our listed titles though they retain the same expected meaning.

As several terms were not recorded in dictionaries, we also browsed through the specialised glossary of job types published by the UK job finding website *Prospects*,[8] and through the International Standard Classification of Occupations elaborated by the International Labour Organization (ISCO08).[9] Also available online are the *Classificazione delle Professioni* (Classification of Occupations) produced by the Italian National Institute for Statistics (ISTAT CP2011) and the ISCO-ISTAT table of correspondences, the *Raccordo* ISCO08-CP2011, issued by the same Institute.[10] In order to account for the currency of Anglicisms in Italian we referred to the online historical archives of the Italian newspapers *La Stampa* (1867-2000) and *la Repubblica* (1984-present).[11]

[7] In order to increase visibility, a single job advertisement is normally posted on several websites; therefore, queries run within a short time span from one another will tend to retrieve many duplicates.

[8] www.prospects.ac.uk

[9] www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm

[10] http://www.istat.it/it/archivio/18132

[11] http://www.archiviolastampa.it/; http://ricerca.repubblica.it/

	CED	CBED	ZING	E&B
accountant	accountant	accountant		accountant
area manager	area manager		area manager	area manager
baby sitter	baby-sitter	babysitter	baby-sitter	
barman	barman	barman	barman	
beauty sales agent	agent	agent	agent	sales agent
data entry		(other meaning)	(other meaning)	
deejay	deejay	deejay	deejay	
development engineer	engineer		mod+engineer	mod+engineer
electrical practical instructor	instructor	instructor		
export area manager	export manager			export manager
(financial) controller	(financial) controller	(financial) controller	controller	controller
first article inspector	inspector	inspector		
hostess	hostess	hostess	hostess	
instrument practical instructor	instructor	instructor		
mistery shopper	mystery shopper	mystery shopper		
order entry				
(junior) programmer	programmer	programmer		
project manager	project manager	project manager	project manager	project manager
promoter	promoter	promoter	promoter	promoter
receptionist (junior)	receptionist	receptionist	receptionist	
retail sales manager		retail manager		
runner	runner		(other meaning)	
sales account		(other meaning)		(other meaning)
sales manager	sales manager	sales manager	sales manager	sales manager
shop assistant	shop assistant	shop assistant		shop assistant
store manager				store manager
store specialist				
telemarketer	telemarketer	telemarketer		telemarketer
visual merchandiser	merchandiser	merchandiser	merchandiser	merchandiser
web designer	web designer	web designer		

Table 1: Job titles in CED, CBED, ZING and E&B.

Formally, a small group of job titles are polymorphemic one-word items, characterized by endings such as -er, -ist, -or, -ant, -man, that typically realize the {noun agent} morphemic function, i.e. denote the agent of the action indicated by the root element (e.g. promote ® promoter). In fact, most job titles are typically complex words, either solid compounds, like *barman* or 2- or 3-word compounds, characterized by the modifier+head structure, in which the head element indicates the job function and the left-hand modifying element functions as classifier, i.e. it indicates a sub-class of the head element (e.g. sales manager = a person in charge of a company's sales activities and its sales force, CBED). This word-formation mechanism can trigger even more complex items if further classification of duties or skills need to be specified (e.g. beauty sales agent*).* However, since Italian is a language that typically modifies on the right of the head element, in complex job titles the order of the elements may be changed, as in *project manager junior* (instead of *junior project manager*): "Stiamo ricercando un *project manager junior* per gestione progetti C++/C#."

In the following paragraphs, we present a sample of the analysis carried out on the start-list of job titles of our glossary, distinguishing between: a) Anglicisms which coexist with Italian equivalents; b) Anglicisms which do not have Italian equivalents; c) English (inspired) job titles which are not recorded in the selected dictionaries or are recorded with another meaning (false Anglicisms).

3.1 Anglicisms with Italian equivalents

The English job titles with a current Italian equivalent are *accountant*, *area manager*, *baby sitter*, *barman*, (*financial*) *controller*, *programmer*, *project manager*, *sales manager*, *shop assistant* and *telemarketer*. The presence of these terms in Italian and English monolingual and bilingual dictionaries online makes it theoretically viable for job hunters in Italy to check the meaning of unknown or unfamiliar terms.

3.1.1 Area manager and sales manager

The head *manager* of these compounds is recorded as part of the core Italian lexicon in ZING (ultimately from It. *maneggiare* according to the *Oxford English Dictionary*). The first attestation of this Anglicism in Italian is 1895. It coexists alongside Italian *direttore* and *dirigente*, which are recorded as equivalents in E&B. This is a highly productive loanword in Italian, which functions as the head of numerous occupational titles.[12]

E&B defines *area manager* as the title used especially by American businesses and organizations to denote the person responsible for the sales force and for the marketing and distribution of products within a specific geographical area. The Italian equivalent proposed is *direttore di zona*. The ZING definition is consistent with E&B, but the recorded equivalent is the Italian *capoarea*. There are no occurrences of the Italian *direttore di zona* in our corpus, which contains instead 9 occurrences of the Anglicism and 3 occurrences of *capoarea*. The examples below show the use of this Anglicism in context:

(1) Per piccola e solida azienda di prodotti per l'edilizia ricerchiamo 1 *Area Manager* Italia. La figura si interfaccerà con la proprietà per seguire e consolidare i clienti acquisiti e espandere il pacchetto clienti. Viaggerà spesso e farà da referente per la rete agenti su tutto il territorio nazionale.

[12] Other titles recorded in the selected Italian dictionaries are account manager, office manager, risk manager, and property manager.

(2) La risorsa sarà inserita come *Area Manager* per il Mercato Italia e si occuperà del contatto e gestione degli agenti e dei clienti; della ricerca e sviluppo di nuovi contatti; studi di settore; redazione di offerte commerciali e partecipazione a fiere di settore.

(3) La posizione, che riporta all'*Area Manager* della zona di competenza, ha la funzione di presidiare dei punti vendita specializzati del canale di riferimento, garantendo il raggiungimento degli obiettivi qualitativi e quantitativi stabiliti.

Examples (1) and (2) are extracts of job ads for *area managers*, which provide a general description of the tasks required by the position, such as maintaining contacts with existing customers and acquiring new ones, liaising with other agents and representatives on the assigned territory, creating proposal documents and representing the company at trade exhibitions. Example (3) is the extract of a job vacancy for a *commercial hostess*, whose tasks will include "reporting to the area manager of the assigned (geographical) area."

In the Standard Classification of Occupations "managers responsible for specialized functions within a specific geographic area" are clearly distinguished from "managing directors and chief executives." (ISCO08: 15). The Italian *direttore o dirigente di dipartimento* are provided in the EN-IT table as the standard equivalents of area manager-level occupational titles (irrespective of the department and specialization, e.g. sales or HR).[13]

With the exception of one occurrence of HR area manager, all instances of the Anglicism in our corpus refer to sales department area managers. The second compound analysed here is, in fact, *sales manager*, translated by both E&B and ZING as *direttore delle vendite*. E&B also records *direttore commerciale*. While this Anglicism is found 5 times in our corpus, *direttore commerciale* and *direttore vendite* also occur 7 times each. These job titles are shown in context in the examples below:

(4) Nell'ambito del potenziamento dell'organico della filiale Svizzera in Ticino di multinazionale americana in costante crescita ricerchiamo *Sales Manager* / Sales Account da inserire all'interno della nostra struttura. Dimestichezza ed interesse per la tecnologia. Requisiti richiesti: -residenza a 25-30 km dal confine svizzero -età compresa tra 25 e 35 anni -esperienza di vendita o simile in servizi business to business di 2 anni -conoscenza lingua inglese.

(5) In un'ottica di potenziamento della rete commerciale, il Gruppo ricerca nuove risorse per il ruolo di *Sales Manager*, da inserire all'interno della filiale di Milano. La funzione prevede lo sviluppo del portafoglio clienti corporate [...].

(6) Importante gruppo tedesco, attivo nella commercializzazione di materiale elettrico e sistemi di fissaggio per il settore fotovoltaico, in un'ottica di forte sviluppo, ricerca un/una *sales manager*.

Example (4) seems to suggest that *sales manager* and *sales account* might be treated as equivalent roles; a more detailed discussion of this pair is provided in section 3.3 below. All job advertisements point to

[13] At a higher level of a company structure we find *direttore generale*, *imprenditore*, *dirigente* e *amministratore* as the standard Italian equivalents to Chief executives or managing directors.

the commercial development of the company through the expansion of its customer base as one of the key responsibilities of the position.

3.1.2 Accountant and (financial) controller

(7) *Accountant / Financial controller </p> <p>* Veneto, Veneto / Permanenti *</p><p>* Per nostra azienda cliente, realtà multinazionale, ricerchiamo un *Accountant / Financial Controller* per la loro sede in Pennsylvania (USA). Il candidato si dovrà occupare di tutta la gestione contabile, fiscale, tesoreria, crediti, liquidità.

Example (7) shows the only occurrence of the Anglicism *accountant*, which, in our corpus, has been superseded by its Italian equivalent *contabile*, with 14 occurrences. The choice of the Anglicism might, in fact, depend on the type of company advertising the vacancy, e.g. the American branch of a multinational corporation. *Accountant* appears to be regarded as a synonym for *financial controller* in the posting, although the two denote different level positions in English: "an executive who is the head of a company's finance or accounts department" the former, and "a person or company whose job is preparing the financial records of people, companies, or organizations" the latter (CBED). The CBED lists *controller* and *comptroller* as alternative forms of this compound. These are also recorded in ZING, which marks them as business terms, and E&B, in which *controller* and Italian *controllore della gestione* are recorded as current equivalents. There are no occurrences of *controllore della gestione* in our corpus, which contains instead 1 occurrence of *controller* alongside the Italian *responsabile amministrativo*:

(8) *Controller* Filiale Svizzera. Dinamico gruppo metalmeccanico italiano ci ha incaricato di selezionare un responsabile amministrativo-*controller* per la sede svizzera di un'azienda.

In fact, the lack of further information in the job advertisement makes it difficult to ascertain whether the position advertised in example (8) is exactly the same as the one described in example (7).

3.1.3 Baby-sitter, barman, telemarketer and programmer

The Anglicism *baby-sitter*, borrowed in the mid-20th century (1950 according to ZING), occurs 20 times in the corpus vs. 2 occurrences of its Italian equivalent *tata*. No occurrences are found for the other Italian equivalent *bambinaia*, which has registered a steady decline in use since the second half of the 20th century (*La Stampa*):

(9) [...] ricerca urgentemente una *babysitter* per attività didattiche e ludiche con bimbo di 6 anni. Si richiede: -Esperienza pregressa nella mansione -Preferibile titolo di studio ed esperienze in pedagogia -Ottima conoscenza lingua inglese o madrelingua inglese

Barman and *barista* are both attested in our corpus, where the false Anglicism *barlady* is also found for the feminine form instead of English *barmaid*, or the gender-neutral form *bartender*.[14]

(10) [...] ricerca per noto locale del fossanese un/a *barman /barlady* con esperienza documentabile nella mansione per inserimento con contratto di somministrazione.

[14] Furiassi records the false Anglicism barwoman (2010: 110-11, 145).

The Italian equivalent *barista* may denote both the person serving drinks in a bar and a bar owner (ZING). The polysemy of the Italian term might obscure the intended meaning of such a job vacancy as the one offered in example (11), where the job requirements: "experience of at least 5 years in the management of bar activities", are generic, and could be read as experience in serving and dealing with customers as well as experience in the actual management of a bar:

(11) Agenzia per il lavoro [...] ricerca per importante bar un barista / *barman*. Requisiti richiesti: esperienza di almeno 5 anni nella gestione delle attività di bar.

Telemarketer is rare in Italian, quoted in the *la Repubblica* daily newspaper 3 times from 2003 but recorded in E&B and translated as *televenditore*. The term *telemarketer* appears in the menu of one of the job finding agencies considered. In our corpus it appears as *telemarketing* preceded by the Italian nouns *operatori*, *operatrici*, *risorse* or *addetti* (equivalent to the English worker/s, workforce) and produces the hybrid compounds *operatrici telemarketing*, *risorse di telemarketing* and *addetto telemarketing*:

(12) Ricerchiamo Operatori *telemarketing* per fissaggio appuntamenti telefonici per conto di Consulente certificato Telecom/Tim. Il lavoro potrà essere svolto da casa.

Finally, an anomalous case in this first group is represented by the Anglicism *programmer*. *Programmatore*, derived from the Italian verb *programmare,* is well established in Italian, occurring in our corpus 57 times vs. a single instance of *programmer*. In fact, as shown in example (13), the job posting in which *programmer* occurs features an unusually high frequency of Anglicisms, italicized below:

(13) *Job Title*: Stage - *Junior Programmer Job ID*: 143142 *Location*: Milano *Organization*: Siemens S.p.A. *Mode of Employment*: Stage, *Full time*. Per il nostro ufficio *Energy Automation Solution Operation* del settore *Infrastructures & Cities* di Siemens Italia, nella sede di Milano (Vipiteno) cerchiamo un *Junior Programmer*. Scopo formativo dello stage è l'affiancamento al nostro personale che si occupa dello sviluppo di sistemi informatici per la gestione di reti di pubblica utilità con l'obiettivo di acquisire la conoscenza per sviluppare applicazioni relative ai sistemi e alle soluzioni progettate nella divisione *Smart Grid*.

3.2 Anglicisms with no Italian equivalent

This group includes *deejay*, *hostess*, *mystery shopper*, *promoter*, *receptionist*, *runner* and *web designer*. *Deejay* is a well-established loan, first attested in Italian in 1987. More recent is the Anglicism *mystery shopper* (CED= "a person who is employed, often by the owners, to visit shops, hotels, etc, incognito, and assess the quality of the service offered"), not recorded in Italian dictionaries but quoted in the *la Repubblica* newspaper (single instance in 1994, then occasionally from 2001 both in its English spelling and with <y> graphically adapted to <i>). Alongside the job details, the advertisement in the corpus also provides a definition of this Anglicism:

(14) Cerchiamo urgentemente una *mistery shopper* per veloce lavoro nel mese di settembre [...] Il *mystery shopper* è il cosiddetto cliente misterioso ossia una persona che fingendosi cliente effettua una visita presso un punto vendita.

Also *web designer*, though unrecorded in Italian dictionaries, appears to be quite transparent in meaning for the Italian user, as clearly denoting "someone whose job is to design websites" (CBED). In some advertisements it appears that the role is treated as an equivalent to the Italian *grafico* (graphic designer) omitting the website-specific function of the role, as in the following example:

(15) Si richiede esperienza consolidata nella mansione di grafico e/o *web designer*. Il candidato ideale è in possesso di conoscenza approfondita dei principali applicativi di grafica (Flash, Illustrator, Coreldraw, DreamWeaver, ecc).

The meanings and lexical profile of the remaining titles are more complex. Beginning with *receptionist*, the examples retrieved from the corpus indicate that the term might also be used as an equivalent of the Italian *centralinista* (telephone operator), or even of *telemarketer*:

(16) Per azienda nel settore moda ricerchiamo centralinista/ *receptionist* che abbia già maturato esperienza di almeno 2 anni presso aziende strutturate e modernamente organizzate.

(17) Centro Fitness vicino a Padova, cerca una *Receptionist* con mansioni di vendita interna abbonamenti, gestione clienti acquisiti, utilizzo del telefono (*telemarketing* in e out).

Another interesting example in this second group is *runner*, defined in the CED as "a messenger for a bank or brokerage firm" (meaning 2) or "a person who operates, manages, or controls something" (meaning 7). The synonyms proposed are "messenger, courier, errand boy, dispatch bearer", thus describing an unskilled, entry-level position and a possible equivalent of the Italian *fattorino, addetto allo spostamento merci*.[15] The only occurrence in the corpus describes this position as follows:

(18) Per azienda moda lusso ricerchiamo 1 *runner*. Si richiede esperienza all'interno di negozi di moda e abbigliamento in qualità di venditore e di magazziniere. La risorsa si occuperà del ricevimento merce e preparazione dei prodotti per la vendita e supporterà in caso di necessità i colleghi venditori.

Thus, to some extent, the job description might be deceptive: *runner* is translated as *venditore e magazziniere* (=salesperson and runner, note that *salesperson* occurs first), although the tasks are the reception and preparation of products (*ricevimento merce e preparazione prodotti*) in support, if necessary, of the actual salespersons. The analysis shows that the Anglicism might in fact find a current Italian equivalent in *fattorino*, and thus typically an unskilled job, although the employer – a luxury-fashion house (*azienda moda lusso*) – requires specific experience in fashion and clothing (*si richiede esperienza*). It should also be pointed out that ZING records the Anglicism with different meanings: a) a person who runs, Italian *corridore* and b) strip of linen placed across a table, which has no current Italian equivalent.

Even more complex are the lexical profiles of the Italian *promoter* and *hostess*. The wordsketch of *promoter* indicates that this term might be used to refer to marketing positions, as an equivalent to *salesperson*, sometimes preceded by "sales", as in "[...] offre la posizione di *Sales Promoter* e un percorso di

[15] cf. *OED*: "A person employed to perform various (generally menial or unskilled) tasks, typically involving moving from place to place. Also more generally: an assistant." (meaning 2d). See also ISCO-08: 543 and 564 "Transport and storage labourers".

crescita professionale" and as an equivalent to *telemarketer*, as in "Azienda settore Telecomunicazioni seleziona operatori / *promoter* telefonici da casa per servizio di promozione e vendita abbonamenti". The term *hostess*, borrowed in 1948, is used in Italian to denote a) women flight attendants or b) conference assistants. In the corpus, however, the meanings found referred to either conference assistants and nightclub hostesses, as illustrated in example (19):

(19) Ragazza *hostess* per lavoro di figurante di sala night club [...] Cerchiamo *hostess* da assumere con regolare contratto per il lavoro di figurante di sala per eleganti ed esclusivi night club di alto livello.

3.3 "English-inspired" job titles

The remaining job titles are complex job titles characterized by a modifier+head structure, in which the head element, generally recorded in English dictionaries, indicates the job function. None of these compounds is recorded in Italian or English dictionaries, although they might indeed sound plausible or acceptable both in form and in meaning, especially considering that they are sometimes accompanied by a description of duties and functions in job advertisements. This third group also features the false Anglicisms *data entry*, *order entry* and *sales account*, which will be treated separately in section 3.3.2.

3.3.1 Complex job titles

This group includes the following titles: *beauty sales agent*, *development engineer*, *electrical practical instructor*, *export area manager, first article inspector*, *instrument practical instructor*, *retail sales manager*, *store manager*, *store specialist* and *visual merchandiser*.

The Anglicism *engineer* makes a particularly interesting candidate for our analysis. This is in fact a highly productive head for new job titles, as shown in both Italian and English dictionaries which record a wide variety of occupational titles like *safety engineer*, *civil engineer* and *mechanical engineer*. While Italian *ingegnere* denotes professionals with a University degree, the English *engineer* might also refer to a technician with specialist competence, but not necessarily a graduate, that the Italian would translate as *tecnico*. In the job market, this is a crucial difference with respect to the salary offered to the prospective candidate, and to the perceived prestige of the position. Corpus analysis might help to profile this term and its usage in a larger corpus of job postings.

Our corpus features one posting for a *testing engineer*. In fact, the search for an entry level position is clarified in the actual job description: the job title is repeated in the first lines of the advertisement, preceded by the adjective *giovane* (young), and the job requirements include details pertaining to experience ("even limited") and level of education ("degree in engineering or other technical qualification").

(20) Requisiti: - laurea in ingegneria meccanica o altro titolo di studio di formazione tecnica. - esperienza pregressa, anche minima, nella mansione maturata all'interno di aziende operanti nell'automotive, su motori a benzina con competenze in particolare su Fuel Injection.

Beauty sales agent might be perceived as more economical and effective than its longer Italian equivalent *agente di vendita (di prodotti) di bellezza*. In fact, even in the job description the Italian standard equivalent of *sales agent*, *agente di vendita* is avoided and replaced by the euphemism *animatrice commerciale* (commercial performer, entertainer or catalyst). Yet, this is the only occurrence of *sales agent* in our corpus, which tends to prefer *agente di vendita* and to specify by means of the job description the business sector of the position advertised (supplies for coffee makers, real estate business, telephone market, electricity, etc.).

Electrical practical instructor and *instrument practical instructor* are plausible English compounds considering "(practical) instructor" as the head of the compound and "electrical" and "instrument" as modifiers. The current Italian equivalent for the head of the Anglicism would be *istruttore* or *formatore*, and denote an instructor for electrical technicians in the one example, and an instructor for instrument technicians in the other ("Vocational education teachers" in ISCO08: 112). Perhaps a strategic function underlies the creation of the "English-inspired" titles for the job posting, which advertises positions for an international training centre which will require fluency in English:

(21) Stiamo cercando un *Instrument Practical Instructor* per il Training Center ECU di Cortemaggiore. Il nostro cliente è l'Eni Corporate University. Sarà un On the Job Training Indirizzato a Instrument Technician iracheni, quindi il corso sarà tenuto in inglese.

(22) Stiamo cercando un *Electrical Practical Instructor* per il Training Center ECU di Cortemaggiore. Il nostro cliente è l'Eni Corporate University. Sarà un On the Job Training Indirizzato a Electrical Technician iracheni, quindi il corso sarà tenuto in inglese.

3.3.2 False Anglicisms

As a job title, the compound *sales account* seems to be an innovation typical of the Italian job market. The examples in the corpus refer to such activities as fostering business to business commercial transactions and expanding the customer base of a company, as example (23) shows:

(23) Si cerca un *sales account* con esperienza nel settore e predisposizione alle attività commerciali per inserimento in importante azienda che opera nella progettazione e realizzazione di prodotti e macchine speciali. La risorsa, rispondendo al responsabile commerciale si occuperà di attività consulenziale tecnica pre e post vendita e di implementazione del portafoglio clienti

Sales account may be considered as equivalent to *sales manager*, as already pointed out with reference to example (4) above. In English, *sales account* indicates "a record of the total cash or credit sales for a particular period" or "a customer that a business sells its products to" (CBED). It is not present in the CED. In fact, *sales manager*, "a person in charge of a company's sales activities and its sales force" (CBED), and *account manager*, "someone employed by a company to be responsible for one or more of its customers, especially someone in the banking or advertising industry" (CBED), are the best candidates as quasi-equivalents to Italian *sales account*, which might in fact be an ellipsis of *sales account manager*, or denote a lower – i.e. not managerial – professional level. This third group also includes such false Anglicisms as *data entry* and *order entry* used in the websites of job finding agencies (cf. footnote 4) to refer to the agent rather than to the activity. While *order entry* is not recorded in the selected dictionaries, *data entry* is recorded in ZING to denote the activity, and not as an agent noun. Examples of the use of either *data* or *order entry* as agents were not retrieved in our corpus – which contains instead one occurrence of the correct usage of the Anglicism in the hybrid compound *impiegata data entry* ("data entry clerk") – though an advanced Google search for the terms in Italian pages published in Italy can easily produce results like "Per importante società multinazionale ricerchiamo un *data entry* con pregressa esperienza nel ruolo da inserire con contratto di somministrazione", "agenzia per il lavoro ricerca per azienda cliente un *data entry*", "Per importante azienda cliente ricerchiamo un *order entry*. La risorsa si occuperà dell'inserimento degli ordini esteri."

4 Conclusion

The Anglicisation of the job market gives the opportunity to linguists to observe language change and lexical innovation and reflect on the underlying mechanisms that trigger the introduction of new job titles. As has emerged from the present corpus-driven research, there is a growing habit of using Anglicisms or English-looking coinages to refer to functions or positions in Italian job postings. As a phenomenon of lexical innovation, the adoption of loanwords is motivated by the need to fill a lexical gap in the recipient language, but, especially in the case of Anglicisms, the main reason is to comply with international terminology in global business, and to express modernity and professionalism.

Our start list contains a few instances of "necessary" Anglicisms, i.e. *deejay*, *hostess*, *mystery shopper*, *promoter*, *receptionist*, *runner* and *web designer.* For these terms there are no competing Italian equivalents. Their success in the recipient language can be ascribed to several characteristics, such as brevity and conciseness for *deejay*, modernity for *web designer*, *promoter* and *baby-sitter* (taking over the old-fashioned *bambinaia* and *balia*, or the childish *tata*). When a domestic equivalent exists, the preference for English is dictated primarily by pragmatic and stylistic reasons, since English terms better answer the need for monoreferentiality and conciseness (e.g. beauty sales agent / *agente di vendita di prodotti di bellezza*). However, the coexistence of a foreign term along with a native equivalent can be regarded as a case of multiple terminology (controller/*controllore della gestione*), which violates the terminological principle according to which a term identifies a single concept (Pulcini 2012). As the job market develops giving rise to new jobs or professional profiles, a new term may in fact describe different duties as in the case of *receptionist*, whose tasks consist not only in answering to incoming calls (It. *centralinista*) but to attend to a wider range of services, including telemarketing. Finally, multinational companies may opt for an English job title to comply with the established international profile of the company, as in the case of *accountant/ financial controller*, which is advertised by a company based in

Pennsylvania, USA. An example of a term which has been successfully assimilated into Italian and also displays great productivity is *manager*. Although many equivalent terms exist to identify different levels of managerial statuses (*direttore*, *dirigente*, etc.), *manager* seems to be an "all-purpose" term, lending itself to a variety of pre-modifications to indicate the management area involved (e.g. sales manager, area manager). We may add that *manager* is a long-standing and very productive Anglicisms in Italian, ultimately a re-borrowing from Italian *maneggiare*, which is the source of the English term.
On the other hand, several advertised jobs may indeed be deceptive for job seekers. The very productive term *engineer*, for example, which resembles Italian *ingegnere* because of the common classical source, may refer to a technician with specialist competence and not necessarily to a professional with a degree in engineering. The former meaning may slowly be filtering into Italian as well, to attribute greater prestige to the actual job designation.
Among the terms discussed in this paper, some may be deceptive for the prospective applicant for different reasons. For instance, the term *hostess* has extended its meaning from air hostess, which has been replaced by the gender-neutral *assistente di volo* in Italian, to other jobs for which a female assistant or attendant is sought. , e.g. in the meeting and event industry. In our data, however, many job positions also referred to nightclub hostesses. For the term *runner*, instead, both job designation and description were obscure, referring to functions as salespersons or storage labourers. Finally, the terms that we labelled as false Anglicisms were possibly derived from the ellipsis of multi-word compounds, e.g. *data entry* for *data entry clerk*.
In conclusion, the adoption of English and "English-inspired" job titles within the context of the Italian job market is a growing phenomenon, partly dictated by the need to name new occupations but especially to comply with the Anglicization of the job market and specialized terminology. Therefore, in this research we aimed to provide the theoretical framework on which to ground the compilation of a glossary of English (or English-looking) job titles – and their potentially misleading nature – to be made publicly available online as a dedicated tool for prospective job hunters in Italy.

5 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: Instant domain-specific corpora to support human translators. *Proceedings of EAMT-2006*, pp. 247-252.

Cambridge Business English Dictionary Online. Accessed at: http://dictionary.cambridge.org/dictionary/business-english/ [07/04/2014]

Classificazione delle Professioni CP2011. Accessed at http://www.istat.it/it/archivio/18132 [26/03/2014]

Collins English Dictionary Online. Accessed at: http://www.collinsdictionary.com/dictionary/english [07/04/2014]

Furiassi, C. (2010). *False Anglicisms in Italian.* Monza: Polimetrica.

Furiassi, C., Pulcini, V., Rodríguez González, F. (eds.) (2012). *The Anglicization of European Lexis*. Amsterdam: John Benjamins.

International Standard Classification of Occupations. Group definitions. Accessed at www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm [26/03/2014]

Kilgarriff, A., Rychly, P. Smrz, P., Tugwell, D. (2004). The Sketch Engine. In *Proceedings EURALEX 2004*. Lorient: Franc, pp 105-116. Accessed at: http://sketchengine.co.uk [13/10/2013]

la Repubblica. Accessed at http://ricerca.repubblica.it/ [02/04/2014]

La Stampa. Accessed at http://www.archiviolastampa.it/ [02/04/2014]

Oxford English Dictionary Online. Accessed at: http://www.oed.com [02/04/2014]

Picchi, F. (2011) *Economics & Business. Dizionario enciclopedico economico e commerciale inglese-italiano con glossario italiano-inglese*. Bologna: Zanichelli. Accessed at: http://dizionarionline.zanichelli.it/dizionariOnline/#economics [07/04/2014]

Prospects. Types of Jobs. Accessed at http://www.prospects.ac.uk/types_of_jobs.htm [07/04/2014]

Pulcini, V. (2012). Register variation in tourism terminology. In R. Facchinetti (ed.) *A Cultural Journey through the English Lexicon*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 109-132.

Pulcini V., Furiassi C., Rodríguez González, F. (2012). The lexical influence of English on European languages: From words to phraseology. In C. Furiassi, V. Pulcini & F. Rodríguez González (eds.) *The Anglicization of European Lexis*. Amsterdam: John Benjamins, pp. 1-24.

Raccordo ISCO08-CP2011. Accessed at http://www.istat.it/it/archivio/18132 [26/03/2014]

Taavitsainen, I., Pahta, P. (2003). English in Finland: globalisation, language awareness and question of identity. In *English Today,* 4, pp.3-15.

Van Meurs, F., Korzillius, H., den Hollander, A. (2006). The use of English in job advertisements on the Dutch job site Monsterboard.nl and factors on which it depends. In *ESP across cultures*, 3, pp. 103-123.

Van Meurs, F., Planken, B., Gerritsen, M., Korzillius, H. (2011). Reasons given by Dutch makers of job ads for placing all-English, partly English or all-Dutch job advertisements in Dutch newpapers: An interview-based study. In C. Degano, G. Garzone (eds.) *Discursive practices and textual realizations in organizational communication: Product and process, frontstage and backstage*. Trezzano sul Naviglio: Arcipelago Edizioni, pp. 53-57.

Van Meurs, F., Hornikx, J., Bossenbroek, G. (2014). English loanwords and their counterparts in Dutch job advertisements: an experimental study in association overlap. In E. Zenner, G. Kristiansen (eds.) *New Perspectives in Lexical Borrowing*. Boston/Berlin: De Gruyter Mouton, pp. 171-190

Watts, R. J. (2002). English in Swiss job adverts: A Bourdieuan perspective. In A. Fischer, G. Tottie, H. M. Lehmann (eds.) *Text Types and Corpora*. Tübingen: Gunter Narr Verlag, pp. 103-122.

Zingarelli, N. (2013) *loZingarelli2014*, Bologna, Zanichelli. Accessed at: http://dizionarionline.zanichelli.it/dizionariOnline/#zingarelli [07/04/2014]

Acknowledgements

We wish to acknowledge the financial contribution of the *Compagnia di San Paolo* (Torino) for the project “The English language in Italy: linguistic, educational and professional challenges” of the University of Torino (2013-2015).

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

Věra Schmiedtová
The Institute of the Czech national corpus, Charles University in Prague
vera.schmiedtova@ff.cuni.cz

Abstract

The intention is to preserve the vocabulary of the period for the younger generation; also to remind the older generation of vocabulary that they used to encounter, but are gradually forgetting. The dictionary is specific in that it is made up of two types of vocabulary – the language of communist propaganda and the spoken language emerging from how people reacted to the pressure of propaganda, often including popular humour. The first type of vocabulary has been collated through Corpus of Totalitarianism, for the second type a corpus-based source does not exist (it is the language as spoken, which it was possible to collate through quotes from fiction, journalistic writings and from the author's own observations. This language has been is checked in contemporary written corpuses, on some occasions it is to be found in the Corpus of Totalitarianism or on the internet).

Keywords: dictionary of communist propaganda; czech language

1 Historical context

The Czechoslovak Republic underwent huge political changes in 1989. The period of communist totalitarian rule ended (1948-1989) and the country returned to democracy. This paper attempts to show how language changed with the change of political discourse. We have to bear in mind that totalitarian language changed its character over time. In this sense, three main historical periods can be identified: The fifties: major ideological pressures dominate Czech society. The focus is on building a new (socialist, communist) society and defining the conflict between the system and its real and alleged opponents. The focus is on the future and the prevailing tone is one of enthusiasm. Young people and children are targeted to represent these values. In some speakers, a full identification between their identity and the ideology of the system - and thus its language - can be observed. The sixties: This period is one of sobering up. Language reflects two main themes: (1) the attempt to escape from the restraints of the communist regime (socialism with a human face), (2) the end of all hope for political change after the Prague Spring, starting with the Soviet occupation of the former Czechoslovakia on August 21, 1968. The seventies and the eighties: The time of disillusionment and so-called normalization. Typical for speakers is not to identify themselves with their language.

2 Vocabulary collected

2.1 “Language of the rulers” – language of propaganda

This was gathered on the basis of the Corpus of Totalitarianism
“Totalitarian Corpus”
This is composed from journalistic texts. It includes three samples of Rudé právo (Red Justice), the daily newspaper of the Communist Party of Czechoslovakia, which reflected the ideological standpoints of the communist government:
The period 1948-1989 can be divided into three periods

2.1.1 The 1950s (1952, a total of 926 texts, from 16.6 to 31.12.1952)

Examples of vocabulary

Building a new order

agitátoři a propagandisté	*agitators* and *propagandists*
v *agitačních střediscích*	*at “agitation centres”*
v rudých koutcích pomocí *agitek*	*in red cells* helped by *propaganda songs*
stěngazety a desky cti	*“wall newspapers”, “lists of honour”*
agitace tlampači	*agitation through loudspeakers.*
komunisté, nestraníci	*communists, non-party members .*
reakcionáři, vykořisťovatelé, kulaci, fabrikanti	*reactionaries, exploiters, “kulaks”, factor owners*
podvracení republiky, velezradu, vlastizradu	*subverting the republic, treason,*
psovi psí smrt	*“a dog’s death for a dog”*

Atmosphere of the time

průvody	*marches*
manifestace s transparenty a s *mávátky*	*demonstrations, banners, flags*
alegorické vozy	*floats*
Sovětský svaz, náš vzor	*The Soviet Union, our model*
akademik Lysenko	*The academic Lysenko*
Mičurin, generalissimus Stalin	*Micurin, Generalissimo Stalin*
stachanovci	*Stakhanovites*
Zlobinova metoda	*the Zlobin Method*
etc.	

Agriculture

kolektivizace	*collectivization*

združstevňování vesnice	*"cooperativizing" a village*
scelování pozemků	*rationalizing parcels of land*
rozorávání mezí	*ploughing in the gaps between fields*

etc.

Industry

havíři/ horníci, úderníci, novátoři a vynálezci	*miners/colliers, "shock-workers", innovators, inventors*
budování socialismu	*building socialism*
pětiletky	*five-year plans*
stavby socialismu	*socialist constructions*
stavby mládeže	*youth constructions*
závazky	*commitments*
zlepšovatelského hnutí	*the "innovation movement"*

etc.

2.1.2 1960s (1969, a total of 1038 texts, from 1.4 to 31.7.1969) Prague Spring

Examples of vocabulary

reformátoři	*reformers*
socialismus s lidskou tváří	*socialism with a human face*
ekonomická reforma, obrodný proces	*economic reform, process of renewal*
demokratizace, pluralita	*democratization, plurality*
deformace	*deformation*
dogmatik, konzervativec, kolaboranti	*dogmatic, conservative, collaborators*
bratrská pomoc, internacionální pomoc	*fraternal support, international help*

etc.

2.1.3 1970s and 1980s (1977, a total of 800 texts, from 3.1 to 31.3.1977)

Continuation of the period of "normalization"

Examples of vocabulary

normalizace	*normalization*
konsolidac	*consolidation*
agent, banda, bdělost, diverze	*agent, band, vigilance, diversion*
oportunista, područí, reakcionář	*opportunist, bondage, reactionary*
spiklenecká banda	*conspiratorial gang*
američtí váleční paliči	*American warmongers*
dřít kůži s těl dělníků	*tearing the skin from the workers' backs*
grandiózní stavba socialismu	*a grandiose socialist construction*

krvavý pes Tito	*that bloodstained dog, Tito*
šťastné zítřky	*a happy future*
zahnívající kapitalismus	*decaying capitalism*
kontrarevoluce, krizové období	*counter-revolution, period of crisis*
prověrky	*screening/vetting*
výměna členských legitimací	*renewal of party membership cards*
pomýlený	*misguided*
vyloučení nebo *vyškrtnutí ze strany*	*expelled or deleted from the party*
zdravé jádro	*the healthy core*
exponent pravice	right-winger
souhlasit/ nebo nesouhlasit se vstupem	expressing agreement/disagreement with the intervention by
(spřátelených) vojsk	(friendly) troops
Chartra 77	*officially described as a pamphlet*
signatáři	*signatories*
samozvanci, zaprodanci rozvratníci	*pretenders/usurpers, traitors, disruptive elements*
samizdat	*samizdat*
edice Petlice	*"Petlice" edition*
pokrývač	*"roofer"*
jít do stoupy	*"to be sent to the shredder"*
trezorový film	*"a film to be kept in the safe"*

Together with scans of 91 propaganda publications of varying lengths.

2.2 "Language of the ruled" – material has been gathered from

Extracts from literary sources – novels etc. Personal experience – existing only in spoken form, these are expressions used among people who felt they could trust each other
Language of the "ruled" – unofficial language

2.2.1 1950s

kdy se to (v)obrátí	when will it turn round
kdy to rupne/ praskne	when will it crack/burst
je načichlej	he's "impregnated"
kopečkář, utýct (za kopečky)	runaway, running away "over the hills"
partajník, fabrika, fárplán	party man, factory, plan
1960s, 1970s and 1980s	
pravý džíny	real jeans

2.2.2 1970s and 80s

byl odejít ze strany	*to be made to leave the party*
Husákovo ticho	*Husák's silence*
Vokovická Sorbona	*The Vokovice Sorbonne*
RSDr. (ironicky Rodné	*(an academic title, referred to ironically as “Doctor by the De-*
cision of the Party”	*nebo Rozhodnutím strany doktor) strany doktor*
rychlokvaška	*upstart, fast-track expert*

2.2.3 Used throughout the communist period

aparátčík	*apparatchik*
papaláš	*bigwig*
Dederon (dederonský), dederon	*slang for someone from East Germany*

3 Description of A small dictionary of life under communist totalitarian rule (Czechoslovakia 1948-1989)[1]

Includes more than 1,400 entries, drawn from a number of fields.

Includes:

1. Language of propaganda – drawn from the “Totalitarian Corpus”

2. Everyday language, capturing how people respond to propaganda, gathered through extracts from texts, through surveys, on the basis of personal experience and knowledge;

a) Language which captures the life of the time, through surveys and on the basis of personal experience and knowledge;

b) The entries also include very specific uses of language (e.g. the language of the secret police, of dissidents, prisoners.) It only includes words that came into common parlance

3.1 Some types of entries in A small dictionary of life under communist totalitarian rule: encyclopedia-type entries

The entry is made up of an encyclopedia-type explanation, taken from an example of the word used in context and stating the source of the example.

Action Kulak was the code name for a secret police operation between 1951-1954, under which awkward peasant families were forced to move and their property confiscated, they were tried on false

[1] Schmiedtová (2012) Malý slovník reálií komunistické totality, Nakladatelství Lidové noviny, Praha ISBN 978-80-7422-192-7

pretences, imprisoned and discriminated: *Exactly fifty years have gone by since the beginning of Action Kulak, which the communist regime directed against peasants throughout the country in 1952* / internet.

censorship */occurring only in texts from the 60s and 70s/* a central pillar of the regime; its discontinuation was one of the main prerequisites of the Prague Spring; it functioned under the auspices of the Federal Press and Information Department (FÚTI), up to 1968 under the Press Monitoring dept.: *It was far worse previously, when real censorship was exercised in newspapers and periodicals, cleverly managed and concealed as "journalistic solidarity"; following the badly organized, politically ill-prepared and ill-considered cancellation of censorship, the press came under the control and decisive influence of rightwing, opportunistic groups* / Corpus of Totalitarianism

agitation centre these were centres established by the Communist Party in villages, town districts and later in workplaces. Political agitation was carried out here, party education, information was published on noticeboards, instant messages and notices were put together and radio broadcasts were prepared, which were broadcast to people living nearby or to people at the workplace: *Under the principles approved by the secretariat of the Central Committee of the Communist Party agitation centres have been established in various places and socialist organizations* / Corpus of Totalitarianism

3.2 Some examples of words and phrases typical for the dictionary

agent */occurring predominantly in the 50s/* **= diversionist, spy** the high occurrence is the result of a phobia, seeking out people perceived as trying to subvert the new regime; people working for enemy intelligence organizations, trying to damage the communist order: *an agent of the American intelligence service; agents of American imperialism; agents of western imperialists; an agent of the bourgeoisie and an enemy of the Communist Party; CIA agents; with the help of a treacherous gang of agents* / Corpus of Totalitarianism

gang */the word occurs frequently primarily in the 50s/: What was this* ***gang of conspirators*** *Slanský and his accomplices aiming at?; Slanský and his* ***criminal gang***; ***a gang of Tito supporters***; *smashing the* ***treacherous and marauding gang*** *of Clementis and co.*/ Corpus of Totalitarianism

not one grain should go to waste! a popular slogan, primarily during the period of collectivization; the slogan also came to be parodied: *so that there will be enough bread in our republic, so that not a single grain of our rich harvest goes to waste* / Corpus of Totalitarianism

facing the masses a communist slogan: *Each communist is committed to the words of comrade Gottwald "Facing the masses"; during the continuous work to win the masses for the political work of the party – fulfilling the principle "facing the masses"*/ Corpus of Totalitarianism

3.3 Words which reflect the real life of the time

"androš" */the word does not occur a single time in the Corpus of Totalitarianism/* **1** independent musical style, underground: *The only real underground Czech music is that of the Plastics and DG 307* / internet **2** an underground musician: *Brabenec's journey from the underground to exile is a clear example of how the regime dealt with those it couldn't control/* internet **3** a person with the outward appearance and lifestyle of the musical underground (long hair, shapeless sweater, scruffy jeans, avoiding regular work, hanging around in pubs, a kind of Czech "hippy"): *it's true that for many years I haven't given a shit about your average citizen, I'm more interested in non-average citizens – I mean guys with long hair, hippies, underground people [androše] or punks* /SYN [The word derives from the English word "underground"]

bon */the word does not occur a single time in the Corpus of Totalitarianism/* a token which could be obtained in exchange for hard western currency, and through which it was possible to buy goods in "Tuzex" shops. These were special shops where primarily western goods were sold. People without access to western currency could only buy these tokens on the black market from illegal currency traders. Officially one token was worth one Czechoslovak crown. On the black market in the 1980s the price for a token was around five crowns: *a whole hierarchy of illegal traders came into being, through whom even "ordinary" citizens could obtain tokens.* / SYN

4 Software used

We use Bonito (created by Pavel Rychlý, 2004) and TchwaneLex TLex Suit, version 7.1.0.726.

5 Conclusion

As we would expect under any political system, the language of totalitarianism in the former Czechoslovakia works within the semantic structure of Czech. However, it uses this structure for propaganda purposes, so words from the usual vocabulary are often abused to propaganda ends. The language is aggressive and monotonous, it frequently repeats certain associations, phrases and slogans. To certain words it adds its own evaluating positive or negative gloss. For example, the word *American* always has a negative semantic connotation, even though it is referring to a geographical concept; the word *Soviet* is always positive. Totalitarianism often abuses, to its ideological ends, words with a positive semantic connotation. It creates new meanings for words by expanding their polysemy, for example *western = capitalist.* It is fond of certain semantic connections, such as *building a better future; the struggle against enemies of the new order; "democratization of culture and education"*, which is a veiled reference to censorship in these fields. With the aim of concealment it often uses euphemisms *(struggle for liberation).* This language is not creative, it draws from automatized components of the language. It often uses set phrases. To this day users often apply these phrases as ironic quotes, referring to the period.

The various tools of propaganda – techniques of persuasion, brainwashing, euphemisms – separate people into those who are with us and those who are against us, into the good and the bad, words take on new meanings, which have a political sense, linguistic stereotypes are used, which are repeated again and again, the propaganda works on the emotions, it is directed at ordinary people, which it perceives as a mass and a collective group, it tries to build its legitimacy on science, it speaks out strongly against the church.

The language of the ruled is spoken language, reacting to the pressure of propaganda. It is highly creative. It often parodies official language, it very often uses humour (e.g. *the "Vokovice Sorbonne", "to be made to leave the party of your own free will"*). It also captures the atmosphere of the time, which was influenced by the way the communist regime functioned (e.g. *real jeans, Tuzex token, Lenon Wall).* ¨

6 References

Čermák, F., Cvrček, V., Schmiedtová, V. (2010) *Slovník kmunistické totality.* Prague: Lidové noviny*.*

Schmiedtová, V. (2006) What did the totalitarian language in the former socialist Czechoslovakia look like? *The First Conference of The Slavic Linguistics Society* - http://www.indiana.edu/~sls2006/page2/page2.htm

Schmiedtová, V. (2007) Totalitní jazyk v bývalém Československu. Koncept slova práce. In: *Totalitarismus 3, sborník z konference, katedra antropologie, FF ZU, s.110 – 116*

Schmiedová, V. (2008) Hodnotící prostředky v totalitním jazyce 1948-1989 v bývalém Československu. In: *Totalitarismus 4, sborník z konference, katedra antropologie* FF ZU. Plzeň, s. 186 – 196

Schmiedtová, V. (2011) Die Sprache der Propaganda in der Tschechoslowakei 1948-1989 In: *Brücken, Germanistisches Jahrbuch Tschechien-Slowakei,* Nakladatelství Lidové noviny ISBN 1803-456X, s. 93-115

Rychlý, P.: Bonito - graphical user interface to the system Manatee, version 1.80

Кобрин, К. (2012) Vita Sovietica. Неакадемический словарь-инвентарь советской цивилизации. *Издательство Август,* Пермь Россия

Mokijenko, V. M., Nikitina T. G. (1998) Toľkovyj slovar jazyka Sovdepii, *Sankt Peterburgskij gosudarstvennyj universitet* Possija

Mokijenko, V. M. (2003) *Novaja russkaja frazeologija,* Opole, Polsko

Aknowledgements

This study was written within the Programme for the Development of Fields of Study at Charles University, No. P11 Czech national corpus

A Frequency Dictionary of Dutch

Carole Tiberius, Tanneke Schoonheim, Adam Kilgarriff
Institute of Dutch Lexicology, Lexical Computing Ltd
{carole.tiberius,tanneke.schoonheim}@inl.nl, adam@lexmasterclass.com

Abstract

In this paper, we present a corpus-based frequency dictionary of Dutch containing the 5000 most frequent words of Dutch. The dictionary has been published at the beginning of 2014 as part of the Routledge Frequency Dictionaries series, a well-established series with titles available for 11 languages at the time of writing. Novel in the Dutch frequency dictionary is that genre has been foregrounded. The dictionary does not contain one single frequency list, but multiple lists are presented, of which four are genre specific covering fiction, newspaper, spoken and web. Throughout the dictionary there are also thematically organised lists featuring the top words from a variety of key topics such as animals, food and other areas of daily and cultural life. Words specific to Dutch in Belgium are also included. The dictionary is based on a 290-million-word corpus which includes both written and spoken material from a wide range of sources.

Keywords: frequency dictionary; genre; Dutch.

1 Introduction

The *Frequency Dictionary of Dutch* provides the 5000 most frequently used words in contemporary Dutch and is specifically targeted at the beginning and intermediate language learner. This is not the first and only frequency list for Dutch, but there was certainly a need for an update. The best-known reference for word frequencies in Dutch is *Woordfrequenties in geschreven en gesproken Nederlands* by P.C. Uit den Boogaart from 1975. Another much used resource, the CELEX database, is more recent (the second release dates from 1996), but it is still over 15 years old and not widely distributed amongst language learners. The current frequency dictionary is contemporary. It is based on a large corpus of Dutch, spanning the past forty years and concentrating on the last twenty.

In Section 2, we briefly summarise the methodology used to compile the frequency dictionary. Section 3 presents the dictionary and discusses a number of issues we encountered while compiling the dictionary and how we have dealt with them. Section 4 concludes the paper.

2 Methodology

The dictionary is based on a 290 million word corpus of contemporary Dutch divided between four genres: fiction, newspaper, spoken and web.[1] This corpus is the result of a compilation of existing Dutch corpus material (Corpus Spoken Dutch (CGN), fiction from INL corpora and newspaper and web material from the SoNaR corpus).

A central problem in preparing frequency lists on the basis of corpora is the '*whelks*' problem: if there is a text about whelks (a variety of mollusc) then the word *whelk* will probably occur many times in this text but not in the other texts of the corpus. If all occurrences of the word *whelk* are given equal weight, the resulting word frequency list will be skewed as this one text about whelks will push up the count of this otherwise rare word. To deal with this problem, we used a fixed-sample-size corpus (cf. the Brown corpus). We first truncated very long texts at 40,000 words, so that we did not have too many samples from any single text, and then we simply concatenated all the texts of each genre and cut into samples of 2000-words each.

Once the corpus had been assembled, it was lemmatised and tagged using the Frog software (van den Bosch et al. 2007).

Some manual checks were carried out (see Section 2.1), and then we calculated, for each genre, for each word, what proportion of samples it occurred in and normalised these figures to give percentages.[2] We then defined an algorithm for determining which words go into which list(s). See Kilgarriff and Tiberius (2013) for a detailed description. As some words occur in more than one of the four genre lists (e.g. *aankomst* $_{\text{fiction(885) | newspapers(499)}}$ 'arrival' occurs in fiction and newspaper), the sum is slightly higher than 5000. The words are distributed across the lists, as follows:

LIST	**Core**	**Fiction**	**Newspaper**	**Spoken**	**Web**	**General**
WORDS	943	1084	1129	155	523	2004
CORPUS SIZE (millions of words)		23	167	9	91	

Table 1: Number of words in the different lists and subcorpora.

2.1 Manual checking and correction

While the Frog tagging and lemmatisation software is good, it does produce occasional unwanted results. For instance, inflected forms were sometimes analysed as separate lemmas. This occurred in particular with singular and plural forms of certain nouns (e.g. *belasting* 'tax.SG' and *belastingen* 'tax.PL', *maand* 'month.SG' and *maanden* 'month.PL'), as well as with masculine and feminine forms of cer-

[1] We use genre in the general sense referring to broad text types.

[2] Thus frequency in the dictionary is always the percentage of documents that a word occurs in.

tain nouns (e.g. *advocaat* 'laywer.MASC' and *advocate* 'laywer.FEM') and with diminutive forms (e.g. *lied* 'song.SG' and *liedje* 'song.DIM'). Inflected forms of some pronouns, adjectives and verbs also produced double lemmas (e.g. *elk, elke* 'each'*; raar, rare* 'strange' and *herkend, herkennen* 'to recognise'). We have corrected the most frequent and evident errors manually, producing a list of lemmas of which the frequencies had to be counted together. In addition, we decided to count abbreviations together with their corresponding full forms, e.g. *kilometer, km.* This was also a manual task.

One of the characteristics of Dutch is that it is possible to separate parts of compound verbs like *uitleggen* 'to explain' , *vasthouden* 'to hold' etc. in the sentence allowing others parts of the sentence to occur in between them (e.g. *hij legt het probleem duidelijk uit* 'he explains the problem clearly' and *hij hield het meisje stevig vast* 'he held on to the girl firmly'). Automatic recognition of such separable verbs is error-prone and there were many instances where they were tagged as separate lemmas. Unwanted particles resulting from these split separable verbs have been manually filtered out of the resulting lemma list.

3 The dictionary

The main part of the dictionary is formed by the six frequency lists. These are:

- **Core:** words occurring with high frequency in all four genres
- **Fiction:** high-frequency fiction words
- **Newspaper:** high-frequency newspaper words
- **Spoken:** high-frequency words in spoken Dutch
- **Web:** high-frequency words on the Dutch web
- **General:** the next band of words which have high frequencies across at least three of the genres.

The words in the lists are sorted by frequency. In the core and the general list sorting is done on the basis of the overall frequency of the words in all four genres. In the genre-specific lists, the ordering is based on the frequency within that genre, rather than the overall frequency. Each entry in these lists contains the headword, its part of speech, an English translation of the commonest sense, and an example sentence showing how the word is used as is illustrated below:

core(509) televisie noun de(f) television

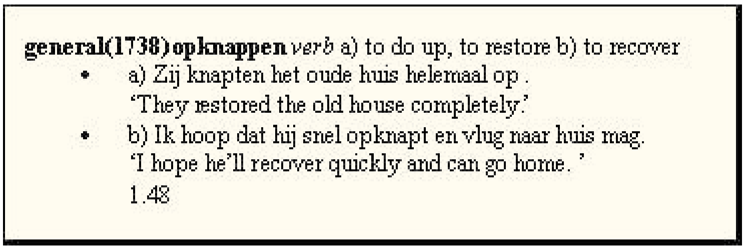
- Hij zette de televisie aan om naar het nieuws te kijken.

  'He switched the television on to watch the news.'

  16.55

This entry shows that word number 509 in the rank order list for the core vocabulary is the noun *televisie* 'television'. It is a feminine noun, which takes the article *de* in Dutch and has an overall frequency of 16.55 per 100 documents. The example sentence is taken from the corpus and shows the word as much as possible in a representative natural context.[3]

Normally, only an example of the commonest sense is given. If however a word has two meanings which are both equally common, two example sentences are given so both meanings can be illustrated.



general(1738)opknappen *verb* a) to do up, to restore b) to recover

- a) Zij knapten het oude huis helemaal op .
  'They restored the old house completely.'
- b) Ik hoop dat hij snel opknapt en vlug naar huis mag.
  'I hope he'll recover quickly and can go home. '
  1.48

In addition to the six frequency lists, the dictionary contains:

- an alphabetically-sorted index;
- an index of the commonest words by part of speech (nouns, verbs, adjectives, adverbs, prepositions, conjunctions and interjections).

Furthermore, there are boxes throughout the book which contain smaller lists of thematically related words, e.g. body, food, materials or grammatical information, e.g. paradigms of auxiliary verbs or lists of pronouns.

In the remainder of this section, we discuss a number of difficulties that we encountered whilst compiling the dictionary and how we solved them.

3.1 Example sentences

For each entry in the dictionary, an example sentence is given. The example sentences were supplied semi-automatically using the GDEX tool (Kilgarriff et al. 2008) from the Sketch Engine. GDEX *(Good Dictionary Examples)* is a tool which automatically sorts the sentences in a concordance according to how likely they are to be good dictionary examples. That is, the best examples are sorted to the top of the list and they are the ones the lexicographer sees first. GDEX was designed for English, so the heuristics that are used are specific to English or they were set with a particular group of users in mind. The tool had not been used on a large scale for Dutch before this project.

[3] Examples are not translated in the dictionary, but a translation has been added here for clarity.

For the frequency dictionary GDEX automatically provided six candidate sentences from the corpus (or from the relevant subcorpus for the genre lists) for each headword which were put in an EXCEL spreadsheet. From these six examples, the best one was chosen manually, marking it with a Y.

	Zij was dan iemand net als zijzelf en niet als de mooie dames op de televisie.
	Hij was iets kleiner dan ik had gedacht op grond van zijn optreden op de televisie.
	De televisie staat op sneeuw.
	Op een avond heb ik haar betrapt terwijl ze huilend voor de televisie zat.
Y	Hij zette de televisie aan om naar het nieuws te kijken.
	Ik ga soms pontificaal voor de televisie staan als ik iets wil zeggen.

Figure 1: Automatically generated example sentences for the noun televisie 'television'.

This worked surprisingly well considering that the tool has not been customised to Dutch. In many cases we shortened or simplified the original corpus sentences to make them more suitable for the language learner. For instance, referential pronouns and personal names have been replaced by personal pronouns.

If none of the automatically selected example sentences were good enough, an alternative example was selected and prepared after examining more corpus examples. This applied to words, like the noun *gek,* which also occur frequently as part of a phrase (i.e. *voor de gek houden* 'to pull someone's leg'*, voor gek staan* 'to look like a fool'*)* or as another part of speech (i.e. the adjective *gek*).

Montaigne schrijft ergens dat hij niet weet wie wie voor de gek houdt als hij met zijn kat speelt.
Of gekken als geheime agenten.
Die bol draait als een gek in de rondte en slaat zonder onderscheid van alles bij je bewustzijn naar binnen.
Ze staat hier voor gek.
Zij acht aan artiesten als aan gekken die elk ogenblik gevaarlijk konden worden.
Maar het is gek dat je bij die dingen nooit denkt dat het ook zo dicht bij je gebeuren kan.

Figure 2: Automatically generated example sentences for the noun gek 'fool, idiot'.

3.2 Translations

The dictionary contains the 5000 most frequent words of Dutch. For each, an English translation of the commonest sense is given. High frequency words are often polysemous and it has not always been straightforward to determine what the commonest meaning of a word is or whether there are different meanings which are all equally common. An example is the verb *optreden* $_{\text{core(727)}}$ which has been translated as ‘to appear’, but can also mean ‘to perform’. As the corpus is not sense-tagged, this is a grey area and decisions on what the commonest meaning is have been made after manually inspecting the corpus data and relying on other resources (ANW, Van Dale).

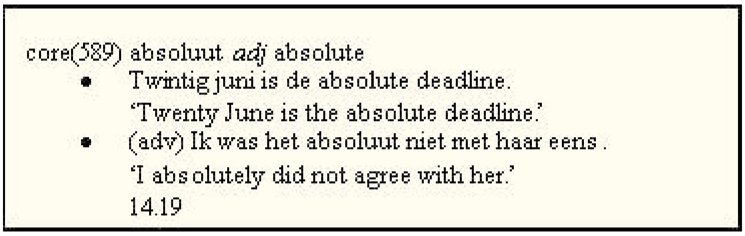
There are also cases where a different translation is more appropriate depending on the genre in which the word is used. For instance, the verb *besturen* $_{\text{newspaper(681) | web(429)}}$ has been translated as ‘to govern’ in the newspaper list and as ‘to drive’ in the web list.

As the dictionary is targeted at language learners we have tried to assure as much as possible that the translations used belong to the core vocabulary of English. This has not always been possible. We have had long discussions about the appropriate translation for *wijf* $_{\text{fiction(552)}}$ in English. In unmarked cases it can be translated as ‘woman’. For the marked case we have ultimately settled for the word ‘broad’ which is neither core, nor general vocabulary (Van Dale marks it as American-English), but seems to express the Dutch connotations of the word best. Another example of a problematic translation was the opposite *gelovig* $_{\text{general(1893)}}$ – *ongelovig* $_{\text{fiction(889) | web(512)}}$ which we have translated as ‘faithful’ and ‘faithless’ in the thematic box of opposites. The adjective *gelovig* is mostly used in a religious context, whereas the opposite *ongelovig* also has a broader sense namely of expressing disbelief which seems to be more common in fiction texts.

Translation of specific terms related to local politics such as *gemeente* $_{\text{newspaper(16) | web(7)}}$ ‘municipality’, *schepen* $_{\text{newspaper(153)}}$ ‘local councillor’ also proved difficult, because these do not match exactly the words that look like their English counterparts.

3.3 Syntactic category

As a rule of thumb, we used the part of speech assigned by the Frog tagging and lemmatisation software in the dictionary. However, there are a few cases where we have diverted from this strategy. This is the case for the adverbial use of adjectives, where the adverbial use of the word is considered secondary to the adjectival use. In the dictionary, these words have been labelled as adjectives, even if the adverbial use was more common in the corpus. An example sentence of both uses is given as is illustrated in the entry for *absoluut*:



core(589) absoluut *adj* absolute

- Twintig juni is de absolute deadline.
  ‘Twenty June is the absolute deadline.’
- (adv) Ik was het absoluut niet met haar eens .
  ‘I abs olutely did not agree with her.’
  14.19

Note that the English translation for the adjectival and adverbial use are not the same.

There were also a few lemmas where it was difficult to assign a single and consistent part of speech, for example, the lemmas *meer* ‘more, *meest* ‘most’ and *minder* ‘less’, *minst* ‘least’. Existing resources for

Dutch (e.g. WNT, Van Dale, the official Dutch spelling guide *Woordenlijst Nederlandse Taal*) do not agree here on the part of speech, indicating the words as adverbs, adjectives and numerals in various combinations (see Table 2):

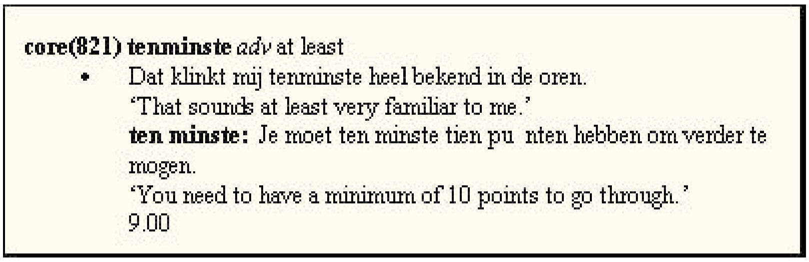
Lemma	WNT	Woordenlijst	Van Dale GW	Van Dale Hedendaags
meer	adv;num	adj	adv;num	adv;num
meest	adj;adv;num	adj	adj;adv	adv

Table 2: Comparison of the part of speech attributed to meer and meest.

The most frequent use of these words in the corpus appeared to be as an indication of a certain amount. This use is considered to be typical for numerals and so in the *Frequency Dictionary of Dutch* these words are labelled as numerals.

3.4 Other cases

In some cases, an entry headword has been given a subentry. This has been done with headwords which are known to cause spelling errors, such as *ten minste* and *tenminste.* Both lemmas exist in Dutch, but they have a different meaning. The word *tenminste* means 'at least' while *ten minste* written in two separate words means 'with a minimum of'
as is illustrated by the two example sentences in the entry below:



core(821) tenminste *adv* at least
- Dat klinkt mij tenminste heel bekend in de oren.
'That sounds at least very familiar to me.'
ten minste: Je moet ten minste tien pu nten hebben om verder te mogen.
'You need to have a minimum of 10 points to go through.'
9.00

It is very likely that in the corpus the appropriate form has not always been used in the appropriate context and thus counts will be skewed anyway. Our approach has been to list them in a combined entry.

Subentries have also been used for multi word expressions as for example in the case of the noun *beslag* $_{\text{general(444)}}$ which means both 'batter' and 'fittings', but also occurs as part of the phrase *in beslag nemen* 'to confiscate'.

Reflexive verbs have been marked by including the reflexive pronoun *zich* behind the verb entry. For example *beklagen (zich)*$_{\text{fiction(1070) | newspapers(847)}}$. The verb *beklagen* is not obligatory reflexive. In a sentence like *Zij kijkt hem vol medelijden aan en beklaagt hem* it means 'to pity'. When used with the reflexive pronoun *zich*, the verb *beklagen* means 'to complain': *Hij beklaagde zich erover dat hij zijn kantoor niet kon bereiken.* 'He complained that he could not reach his office.' An example of each is included in the dictionary.

4 Conclusion

In this paper we have discussed the *Frequency Dictionary of Dutch* that has just appeared as part of the Routledge Frequency Dictionary series. It provides first and foremost a valuable resource for learners of Dutch, but it is fascinating for anyone interested in the Dutch language. The web material (never used before in a Dutch frequency dictionary) appears to be an interesting mixture of informational language like the language used in the newspaper genre, and a written form of spoken language, as used in blogs and discussion groups. Newspaper material shows a focus on economy and sports, whereas the material taken from fiction tends to be rather conservative.

Besides this, it is material which provides lots of new research questions for (socio-)linguists and lexicographers. As van Oostendorp (2014) points out, while reading the dictionary, you can't help wondering why *schouders* 'shoulders' are so popular in fiction and why *januari* 'January' is the most frequently mentioned month on the web followed by *juni* 'June', *mei* 'May', *december* 'December', *oktober* 'October' and *maart* 'March'. The frequency dictionary itself does not provide the answers, but these are intriguing observations about our use of the Dutch language.

5 References

Bosch, van den A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium. 99-114.

Kilgarriff, A. and Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In: Elisenda Berndal and Janet De Cesaris (eds), *Proceedings of the XIII EURALEX International Congress*, Barcelona, Spain. 561-569.

Killgarriff, A. and C. Tiberius (2013). Genre in a Frequency Dictionary. In: Andrew Hardie and Robbie Love (eds.) *Corpus Linguistics 2013 Abstract Book.* Lancaster. UCREL, 142-144.

Oostendorp, van M. (2014). Het karakollenprobleem. Accessed at: http://nederl.blogspot.nl/2014/03/het-karakollenprobleem.html. [04/04/2014].

Uit den Boogaart, P.C. (1975). *Woordfrequenties: in Geschreven en Gesproken Nederlands.* Utrecht: Oosthoek, Scheltema & Holkema.

Resources:

Algemeen Nederlands Woordenboek (ANW) Accessed at: http://anw.inl.nl/ [20/08/2013]

The CELEX Lexical Database (1995), R.H. Baayen, R. Piepenbrock and L. Gulikers, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Corpus Gesproken Nederlands (CGN), (2004), Nederlandse Taalunie, Den Haag.

INL-Corpora, Instituut voor Nederlandse Lexicologie, Leiden. Accessed at http://chn.inl.nl [01/02/2014]

SoNaR (2010), Nederlandse Taalunie, Den Haag.

Woordenboek der Nederlandsche Taal (WNT) Accessed at: http://gtb.inl.nl/ [04/04/2014]

Woordenlijst Nederlandse Taal, Nederlandse Taalunie, Den Haag. Accessed at: http://woordenlijst.org/ [04/04/2014]

Van Dale *Groot Woordenboek van de Nederlandse Taal*, Van Dale Lexicografie, Utrecht. Online version [04/04/2014]

Van Dale *Groot Woordenboek Hedendaags Nederlands*, Van Dale Lexicografie, Utrecht. Online version [04/04/2014]

Acknowledgements

We gratefully acknowledge our colleagues at INL who have helped us in the process of compiling the frequency dictionary. We also like to thank the Sketch Engine team for their support and Karen Newton and James McAllister for their help with the English.

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

Vida Vukoja
Old Church Slavonic Institute, Zagreb
vidal@stin.hr

Abstract

Croatian Church Slavonic is a literary, bookish idiom used in Croatia from the XI/XII until the XVII c. based on Old Church Slavonic, an idiom created by Sts. Cyril and Methodius, and shaped by the Croatian vernacular.

The Croatian Church Slavonic corpus consists of the texts excerpted from 11 breviaries, 4 missals, 3 psalters, 3 rituals, 15 miscellanies. It also incorporates all the 26 fragments dated from the period up to and including the XIII c. and several auxiliary sources. The corpus was created over a period of thirty years (from 1959 to the early 1990s) at the Old Church Slavonic Institute in Zagreb. It is a historical, referential, representative paper card-file of excerpts. It is also a parallel corpus, as it contains Latin and Greek parallel texts, those that were identified as closest to the actual source texts for the translational Croatian Church Slavonic texts.

The *Dictionary of the Croatian Redaction of Church Slavonic* has been compiled on the basis of the Croatian Church Slavonic corpus. The fascicles of the *Dictionary* have been published since 1991. So far, 1 (1991)–19 (2012), with the dictionary articles A–ŽRЬTVA (according to the Old Cyrillic alphabet) have been printed.

Keywords: corpus of the Croatian Church Slavonic; Dictionary of the Croatian Redaction of Church Slavonic; (Paleo)slavistic lexicography

1 Basic Information on the Croatian Church Slavonic Language (acr. CCS)

The Cyrillo-Methodian landmark mission, with its far-reaching influence on Medieval European culture and history, felt on a broad scale even today, commenced in 863, when the Thessalonian brothers arrived in Great Morava. Cyril and Methodius created a new language, now known as Old Church Slavonic (acr. OCS), for the purpose of translating Greek biblical and liturgical texts, literary texts as well as those concerning administration and law. That language did not match any particular Slavic vernacular, as it was constructed to potentially serve all the Slavic peoples ready to be evangelized in the

Slavic language (i.e. OCS), and willing to embrace literacy in a newly-invented script for its notification – Glagolitza or the Glagolitic script, created probably by Cyril.

After the end of the mission, a certain number of Cyril and Methodius' disciples arrived in the territory populated by the Croats. Under the influence of the Croatian vernacular, a new Church Slavonic language system came to be. That literary, bookish idiom used in Croatia from the XI/XII until the XVII c. is known as the Croatian Church Slavonic. It had a privileged status throughout the Croatian Middle Ages within the Croatian diasystem,[1] characterized by the Croatian/CCS diglossia[2], as it was a liturgical language, whose usage regularly marked a high literary style.

Two things should be mentioned in order to signal the importance of CCS. The first one considers European culture and history, and the even broader context of the Catholic Church history. CCS was the only close-vernacular idiom which gained and retained the explicit permission of the Pope to be used for liturgical purposes (besides Latin, Greek and Hebrew). Therefore, ahead of the decision of the Second Vatican Council allowing Catholic liturgy in vernacular, a CCS mass was served in the Roman St. Peter's basilica. The second thing to mention considers Croatian literacy. Namely, CCS is the first Croatian literary language, used from the end of XI c. until 1561. Its significance is reflected in the fact that it is the language of the *Baška tablet* (Cro. Bašćanska ploča, dated in 1100), one of the first and most important Croatian written monuments, but also the language of the first Croatian incunabula, *Missale Romanum Glagolitice,* which is the first missal in Europe not published in the Latin script and Latin language. Six out of nine Croatian incunabula were printed in CCS. The preserved Croatian texts in the Cyrillic script date from later periods, as is the case with the texts in Latin script.

2 The Corpus of the Croatian Church Slavonic texts (abbr. the CCS Corpus)

2.1 Basic Information on the Corpus of the Croatian Church Slavonic texts (abbr. the CCS Corpus)

Here, the term "corpus" is understood as a language material, a cluster of texts, purposefully collected to testify choices and combinations of choices made by users of a particular language (Sinclair 2003:167; cf. Svensén 2009: 43). The CCS corpus is primarily prepared for the compiling of the *Dictionary of the Croatian Redaction of Church Slavonic* (acr. DCRCS), but due to its features, it serves as a prime source for various linguistic investigations and other types of research.

[1] For the meaning of the term diasystem v. Weinreich (1954), cf. Brozović (1970 [1967]): 14. For the CCS and Croatian vernacular and literary idioms as constituents of one common diasystem v. Katičić (1992).

[2] For the character of the Croatian/CCS diglossia v. Mihaljević (2010).

The history of the CCS corpus started with the Fourth International Slavistic Congress, which took place in Moscow in 1958[3], where the suggestion to compile a thesaurus of the Church Slavonic language was embraced by the leading (paleo)slavists. The planned thesaurus was supposed to incorporate all the national Church Slavonic versions (or redactions): Bulgarian, Macedonian, Czech, Russian, Romanian, Bosnian, Serbian. It is important to mention that during that very time, the first fascicles of the landmark *Slovník jazyka staroslověnskeho* (abbr. *Slovník*; v. Štefanić 1962; Nazor 1991: I), based on the corpus of the canonical OCS texts, started to appear. The form and principles applied to the corpus prepared for the *Slovník* compiling, and the *Slovník* compiling itself, decisively influenced the CCS corpus and the DCRCS compiling respectively. The decision of the Croatian lexicographers to follow their Czech colleagues was not due to any lack of ingenuity, but was incited by the long-sighted priorities to shape the CCS corpus as similarly as possible to the corpus for *Slovník* compiling and to shape the DCRCS as similarly as possible to the *Slovník*. The aim of those decisions was to enhance the possibilities for comparative research ultimately aimed at attaining the structural knowledge of: (a) particular national versions of Church Slavonic, (b) the (Old) Church Slavonic language system, taken apart from any (national) vernacular's influence.

For that reason, the arrival of the reputable and experienced Czech paleoslavistic lexicographer, František Václav Mareš, one of the principle collaborators at the compiling of *Slovník*, was very much welcomed, as was his cooperation with the then-young Croatian paleoslavists, in the task of his laying-out of the main principles for the CCS corpus and the compiling of the DCRCS drafted in Mareš (2007[1962]), as well as setting up the long-lasting work of the DCRCS compiling. It took about thirty years to complete the CCS corpus (from 1959 until the beginning of the 1990s). Despite its apparent excessive duration, it was actually the expected length for such a demanding task.

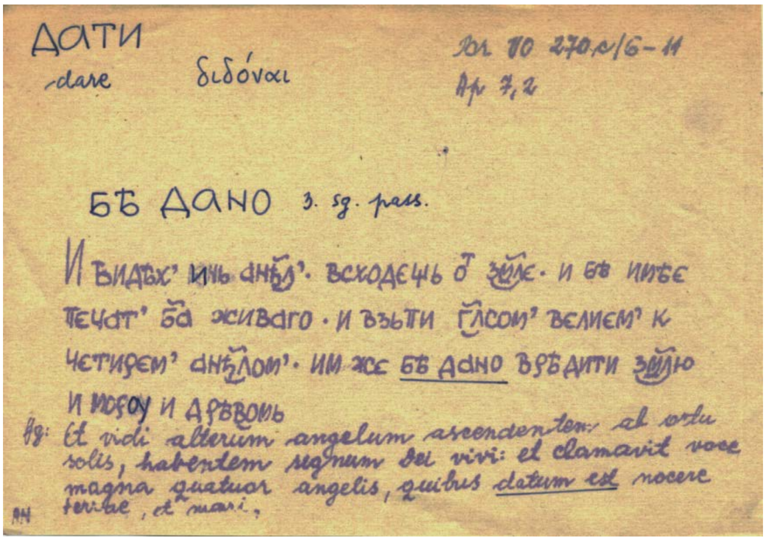
2.2 The Constituents, Card-files and Features of the CCS Corpus

The CCS corpus consists of selected CCS sources, manuscripts and incunabula dating from (XI/)XII to mid-XVI c., with the priority given to earlier and integral versions of particular texts. Its constituents are as follows: 11 breviaries, 4 missals, 3 psalters (1 with Psalter commentary), 3 rituals, selected texts from 15 Croatian Glagolitic miscellanies, all the fragments dating from the period up to and including the XIII c. (altogether 26 pieces), auxiliary sources are excerpted in the cases of lexicographically interesting lemma occurrences: another 2 missals and 2 breviaries.[4] A vast range of text genres found their place within the CCS corpus: liturgical texts (including biblical passages), biblical and apocryphal texts, sermons and homilies, moral and didactical texts, legal texts, legends and visions, hagiographies, disputations and other literary texts.

[3] For the information on the prehistory of the CCS corpus v. Nazor (2008).

[4] For the exhaustive list of the sources, constituents of the CCS corpus v. Nazor (1991).

The exact number of the CCS corpus tokens is not known, but is should be somewhere between 1 400 000 and 2 100 000. The excerpts are of various lengths and multiplied so that they can be organized within two card-files (an example of excerpt cards can be seen in the Figure 1).



ДАТИ
dare διδόναι Br VO 270c/6-11
Ap 7,2

БѢ ДАНО 3. sg. pass.

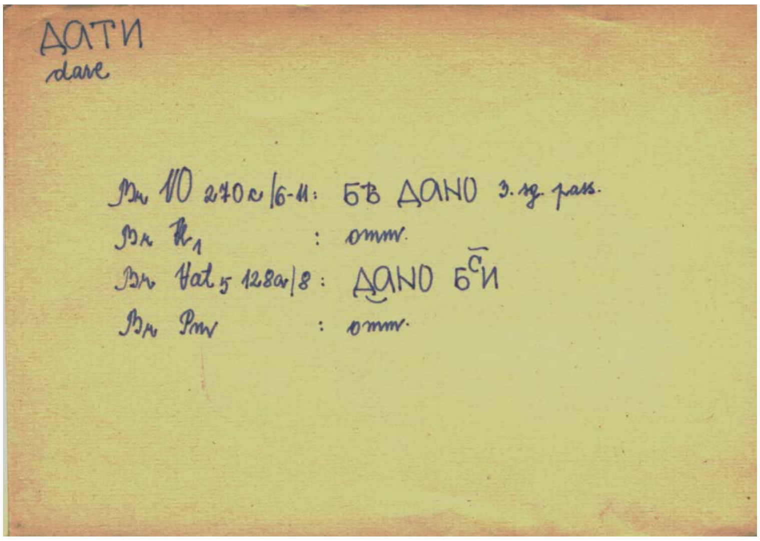
И видѣх' инь анћл'. всходещь ѿ зап(а)де. и бѣ имѣе печат' б(ог)а живаго. и възъпи гласом' велием' к четирем' анћлом'. им же бѣ дано врѣдити землю и морꙋ и древомь

fg: Et vidi alterum angelum ascendentem ab ortu solis, habentem signum Dei vivi: et clamavit voce magna quatuor angelis, quibus datum est nocere terrae, et mari,

AH

Figure 1: An excerpt card (used in the sources card-file and in the azbuka card-file).

Along the excerpt cards run the so-called parallel cards with the variations of the lexical constituents of the excerpt as found in other sources (an example of a parallel card is given in Figure 2.).



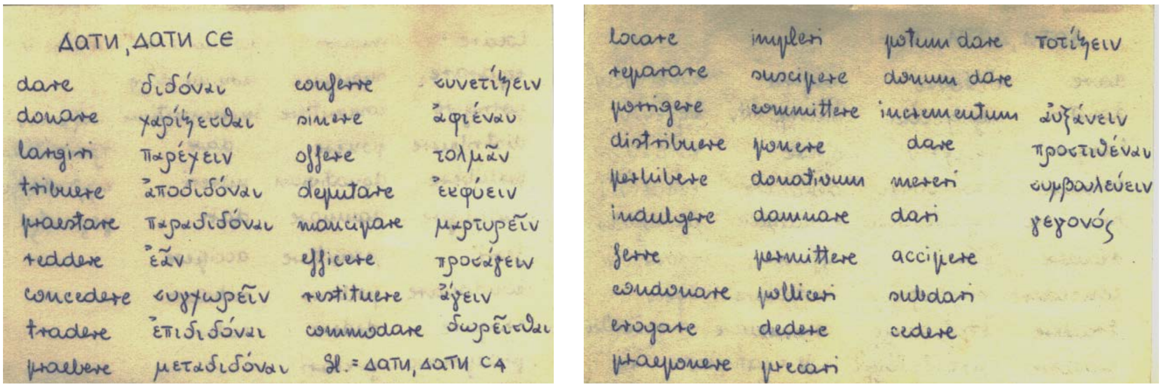
ДАТИ
dare

Br VO 270c/6-11: БѢ ДАНО 3. sg. pass.
Br N₁ : omm.
Br Vat₅ 128a/8 : ДАНО БСИ
Br Pm : omm.

Figure 2: A parallel card (used in the sources card-file and in the azbuka card-file).

The first card-file is established according to the sources (informally, the sources card-file) and it contains approximately 420 000 cards. The second card-file is established according to the azbuka sequence of the lemmas (informally, the azbuka card-file), with more than 400 000 cards. The azbuka card-file contains fewer cards than the sources card-file because not every token is taken into consideration for the compiling of the entry of the DCRCS, and consequently, its card doesn't appear in the latter card-file.

There are three auxiliary card-files. The first one to be mentioned is the card-file of the CCS lemmas (systematized according to the Old Cyrillic azbuka) with approximately 18 100 cards (of which approximately 8500 nouns, 5400 verbs, 2500 adjectives; an example of a card is shown in the Figure 3).



ДАТИ, ДАТИ СЕ

dare
donare
largiri
tribuere
praestare
reddere
concedere
tradere
praebere

διδόναι
χαρίζεσθαι
παρέχειν
ἀποδιδόναι
παραδιδόναι
ἐᾶν
συγχωρεῖν
ἐπιδιδόναι
μεταδιδόναι

conferre
sinere
offere
deputare
mancipare
efficere
restituere
commodare
Sl. = ДАТИ, ДАТИ СА

συνετίζειν
ἀφιέναι
τολμᾶν
ἐκφύειν
μαρτυρεῖν
προσάγειν
ἄγειν
δωρεῖσθαι

locare
reparare
porrigere
distribuere
perhibere
indulgere
ferre
condonare
erogare
praeponere

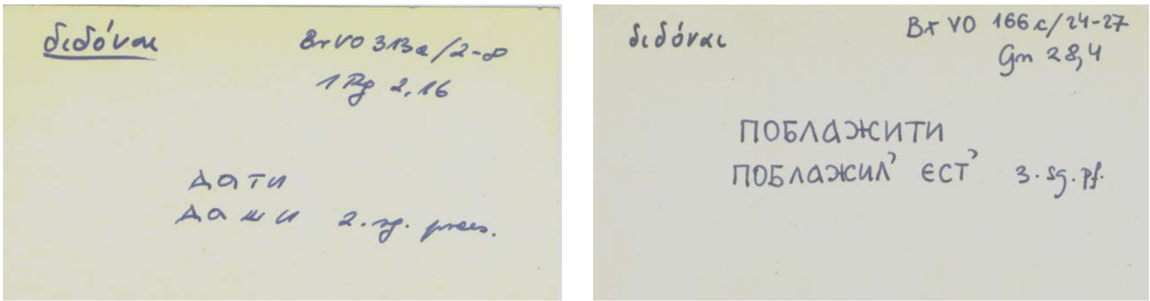
impleri
suscipere
committere
ponere
donativum
damnare
permittere
polliceri
dedere
precari

potum dare
donum dare
incrementum dare
mereri
dari
accipere
subdari
cedere

ποτίζειν
αὐξάνειν
προστιθέναι
συμβουλεύειν
γεγονός

Figure 3: A card from the card-file of CCS lemmas (*recto* and *verso*).

The second one is the card-file of the Greek parallels of the CCS lemmas (systematized according to the Greek alphabet) with approximately 60 000 cards (two examples are given in the Figure 4).



διδόναι
Br VO 313a/2-8
1 Rg 2,16
Дати
Даши 2. sg. praes.

διδόναι
Br VO 166c/24-27
Gm 28,4
ПОБЛАЖИТИ
ПОБЛАЖИЛ҆ ЕСТ҆ 3. sg. pf.

Figure 4: Two cards from the card-file of Greek parallels.

Also, there is the card-file of the Latin parallels of the CCS lemmas (systematized according to the Latin alphabet) with approximately 200 000 cards (two examples can be seen in the Figure 5).

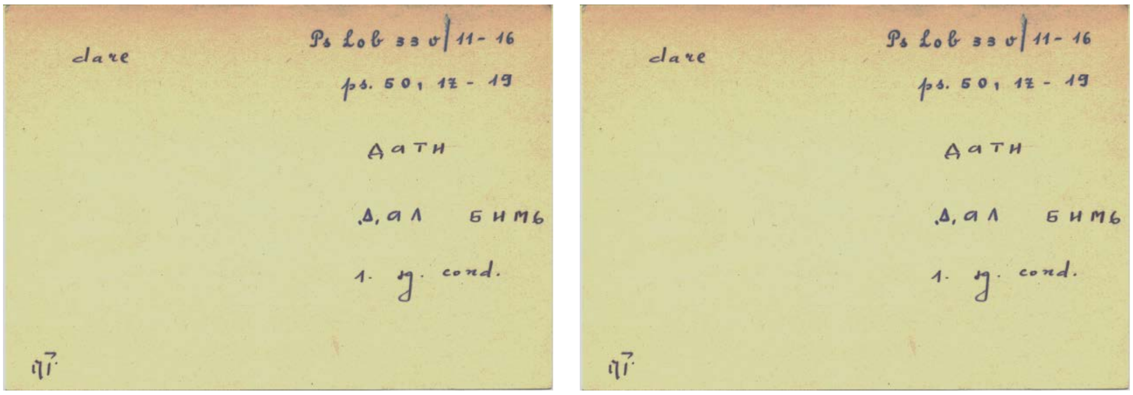


Figure 5: Two cards from the card-file of Latin parallels.

The main features of the CCS corpus are shaped by the principles articulated by Mareš (2007[1962]), and they can be listed as follows:

(1) The CCS corpus is *historical*: it contains texts originating from (XI/)XII to XVI c.

(2) The corpus is made with the aspiration to enable the production of a *thesaurus*, i.e. a dictionary containing the entire vocabulary range confirmed in the CCS texts.

(3) It is *referential*, general; i.e. it contains virtually all of the confirmed CCS lexemes in a number of instances and types of context which represent the situation in the entire (preserved) corpus of the CCS texts. Also, it contains various versions of the same lemma in terms of its graphic features (such as manners of noting /ь/, /ъ/, /m/, Latin or Cyrillic script initials in otherwise Glagolitic texts etc.) as well as its phonological, morphological, lexical, textological features. Therefore, the corpus meets the requirements for various types of research, be it linguistic or non-linguistic, such as textological, literary, liturgical, historical etc.

(4) The corpus is *representative*. It contains texts of different genres, mirroring in quantity the real share of particular genres in the entirety of CCS texts.

(5) In form, the corpus is *paper card-file* of excerpts, where excerpts are used for practical reasons, as expected considering the period it was set up. Still, it cannot be perceived as a mere collection of citations, as the sequences of the card-files contain integral texts, and even entire hundreds-of-pages long manuscripts, just as it is expected in the case of a proper corpus (and unlike expected for a collection of citation).

(6) The corpus *constituents* are *not strictly divided*. Despite being fundamentally systematized on the basis of the source type (missals, breviaries, psalters, rituals, fragments, texts selected from miscellanies), the corpus is not subdivided into compartments. But, if important versions of a particular text appear in different groups of sources, the differences are noted on a string of parallel cards.

(7) It has characteristics of *static* and *dynamic* corpora. A static corpus (also, *sample corpus*) is a structured collection of a limited number of examples of the lemmatized lexemes, which, once introduced in the collection, cannot be excluded from the corpus. And the CCS corpus is such a collec-

tion in its azbuka card-file. A dynamic corpus (also, *monitor corpus*) consists of entire texts and it is open to receiving a newly-found text or confirmation of lemmas shown to be desirable concerning the referential and representative features of the corpus, and the CCS corpus is open to the incorporation of such newly-found texts.

(8) The corpus contains *only written texts*, with a great majority of texts being manuscripts and a limited number of incunabula. As it concerns medieval language material, no spoken texts are expected.

(9) Besides the above reason, the CCS corpus can contain written texts only, as CCS itself is a *bookish* idiom (not spoken), *non-organic*, and *not* a *native language* of anyone.

(10) Texts for the corpus are *excerpted directly from the original sources or photocopies of the original sources* (not from any secondary editions of the selected texts).

(11) The corpus is *parallel*, not monolingual (on parallel corpora v. Svensén 2009:55-57; also Borin 2002). Whenever an excerpted CCS text is actually a translation from Greek or Latin (and most often this indeed is the case), if the parallel Greek or Latin text has been determined, the cards contain that parallel text below the CCS texts, as strictly aligned as possible.

(12) The corpus is formed in such a manner that the *basic version of the text is determined and differentiated* from the secondary versions of the same text found in various sources.

(13) The *critical, phototypical and facsimile editions* of the texts included in the corpus are *published*, the most ambitious undertakings being the capital editions of Hrvoje's Missal (Grabar et al. 1973) and the Second Novi Breviary (Pantelić & Nazor 1977).

(14) Approximately *18,100 lemmas* are *expected* in the dictionary once it is finished.

(15) Without exception, the texts incorporated in the CCS corpus are written in the *Glagolitic script*, but on the card-files of the CCS corpus they appear transliterated in the Old Cyrillic script, according to the decision made at the Fourth International Slavistic Congress. The principles of the transliteration from Glagolitic into Old Cyrillic script are appropriated from Jagić (the table of Glagolitic and Old Cyrillic correspondents can be found in Jagić 1879: XXXVII; cf. Bratulić 1981: 145-146). The lemmas of the DCRCS are also written in the Old Cyrillic script. Of course, Greek parallel texts are written in the Greek alphabet and Latin parallel texts are written in the Latin script.

(16) The *tokens* of the CCS corpus, appearing in the sources card-file, are *grammatically tagged* (parsed) and inserted into the azbuka card-file in all the cases except those of extremely frequent types, such as the forms of the verb *biti* ('to be') or the noun *bogь* ('god').

(17) All the types are *lemmatized* in a *normalized form*. The principles of normalization are given in Grabar et al. 1991: VIII-XIII, XIX-XXV.

2.3 The European and Croatian Context of the CCS Corpus and the Compiling *of the DCRCS*

2.3.1 The Situation in the (European) Paleoslavistic Lexicography

Currently, after the completion of the landmark four-volumes OCS dictionary (i.e. *Slovník*), the Czech paleoslavistic lexicography has commenced five major lexicographic works based on several paleoslavistic (meta-)corpora: *Etymologický*, *Slovník* V., revision of *Старославянский*[5], *Srovnavaci index k slovnikům zpracovavanym v ramci Komisie pro cirkevněslovanske slovniky*[6], *Řeckostaroslověnsky*. Among these, special attention should be drawn to *Srovnavaci*, as the CCS corpus is a component of its meta-corpus, and the DCRCS' lemmas are taken into consideration during the process of determining similarities and differences of various Church Slavonic idioms.

After publishing major Old and Middle Bulgarian[7] dictionaries (*Старославянский; Бончев* 2002-2012), Bulgarian paleoslavists took up the compiling of a voluminous digitized corpus *Компютърни и интерактивни средства за исторически езиковедски изследвания*[8], which is to become a permanent basis for the comprehensive diachronic description of Bulgarian (including its Church Slavonic constituent) from X to XVIII c. Also, Bulgarian paleoslavistic production abounds with dictionaries and indices of particular texts and sources (e.g. *Тасева* 2010, esp. pp. 533-818; *Димитров* 2010, 2013; *Илиева 2013a, 2013b).* Macedonian paleoslavists are working on the dictionary of the Macedonian Church Slavonic, based on the corpus comprising documents from the XII to XVI c. *(Речник)*. Once the corpus of the Serbian Church Slavonic texts have been created (and it is currently in preparation), the Serbian Church Slavonic dictionary compiling can be expected (*Српскословенски* serves as its introductory fascicle).

Important state-financed projects were recently run or are being run at present in non-Slavic countries, among which two will be mentioned: *SlaVaComp. COMPutergestützte Untersuchung von VAriabilität im KirchenSLAvischen* in Germany (Freiburg, 2013-)[9] and *Die kirchenslavische Übersetzung der Werke von Gregorios Palamas und Barlaam von Kalabrien* in Switzerland (Bern, 2010-2013)[10].

2.3.2 The Situation in the Croatian Historical Lexicography

Currently, in Croatia, besides the DCRCS, there are three historical lexicographical projects in different phases of progress: *Dictionary of Croatian Kajkavian literary language* (Cro*. Rječnik hrvatskoga kajkavs-*

[5] Emilie Bláhová is the redactor of the revisited and supplemented edition of С**тарославянский**.

[6] The first volume should be published in forseeable future.

[7] Bulgarian slavists use the term "Middle Bulgarian" for the idiom other slavists usually name "Bulgarian Church Slavonic", and "Old Bulgarian" for the idiom the other slavists usually name "Old Church Slavonic".

[8] The corpus has been created at the Софийски университет „Св. Климент Охридски" under redaction of Dora Ivanova-Mirčeva. For more information on that corpus v. Totomanova (2012).

[9] The project is led by J. Besters-Dilger and G. Schneider (University of Freiburg). More information can be found at the web-site: http://www.slavacomp.uni-freiburg.de/ [15/10/2013].

[10] The project was led by Y. Kakridis (University of Bern) and financed by *Schweizerischer Nationalfond*.

koga književnog jezika), *Dictionary of the Croatian literary language from the National Revival to I. G. Kovačić* (Cro. *Rječnik hrvatskoga književnoga jezika od preporoda do I. G. Kovačića*; compiling of which was resumed in 2008, after 18 years of hiatus), *Old Croatian dictionary* (Cro. *Rječnik starohrvatskoga jezika*, in the middle of the elaborate preparation of its corpus).

Here, the last of the three lexicographical projects is the most important one, because the corpus which is being prepared for the *Old Croatian dictionary* compiling shares two very important features with the CCS corpus: (a) both are referential which also makes the dictionaries based on them referential; (b) both corpora consist of integral texts found in manuscripts and incunabula originated from the same period, namely (XI/)XII—XV/XVI c. But, the fact that the two corpora and the two dictionaries are compatible is even more important than their aforementioned overlapping features. Indeed, the two corpora and the two dictionaries taken together cover two major components of the medieval Croatian diasystem: CCS (in the case of the DCRCS) and literary medieval Croatian idioms (in the case of the *Old Croatian dictionary*).[11] Thus, to once have both corpora and both dictionaries at disposal is critical for the trustworthy diachronic investigation of the Croatian language. It. cannot be emphasized enough that it is virtually impossible to conclude the investigation of the medieval Croatian diasystem without both of the corpora and respective dictionaries published.

The research of the CCS component of the Croatian diasystem appears to be more demanding for a contemporary linguist then the research of the literary medieval Croatian idioms due to the fact that the CCS corpus of texts is much less familiar, even obscure, to the majority of linguists dealing with the diachrony of the Croatian language, which is not much of a surprise if the following is kept in mind:

(1) it is written in an unusual script (i.e. Glagolitic);

(2) it is written in an idiom (i.e. CCS) not entirely comprehensible if relying solely on the good command of the Croatian vernacular diasystem;

(3) in numerous cases, the linguistic features of a given CCS text cannot be interpreted correctly if its Greek or Latin parallel text is unfamiliar to the investigator;

(4) CCS documents are scattered not only across Croatia, but also across the world (including countries such as Russia, Austria, Slovenia, United Kingdom, USA and so forth);

(5) in many cases, the CCS sources are at least partially damaged due to age and mal-manipulation, sometimes even to the point of being barely readable.

It can be added that the above-mentioned features of the CCS corpus give a clue to a range of competences CCS lexicographers have to have at their disposal, as well as to the indispensability of their contribution to the research of the overall diachrony of the Croatian language.

[11] For the comparison of the two corpora v. Vukoja (2012).

3 The Contribution of the DCRCS Compiling to Croatian Studies and to the Historical Lexicography in General

3.1 The Contribution of the DCRCS Compiling to Croatian Studies

Being referential and representative, besides its other features, the CCS corpus provides the most effective basis available for the relevant research of the CCS idiom. If conducted in a methodologically correct manner, the conclusions drawn from the analyses of the corpus can be taken as trustworthy for the whole of the CCS idiom.

In general, every research of the CCS has to take into consideration three layers of factors: (i) the inherited Old Slavonic state, with its characteristic Greek and Latin influences; (ii) the Latin influences that CCS acquired through the adaptation to the Western (Church) tradition, which should be taken separately from the Romance influences acquired through spoken language; (iii) the influences of Croatian vernacular idioms. At present, systematic research of CCS is conducted on the basis of the CCS corpus in several fields: grammar (phonology, morphology, syntax, semantics), lexicology, textology, translation theory and practice, research on the medieval conceptualization of feelings.

With the DCRCS, Croatian lexicography would vastly improve the starting point for all the research on the Croatian diachronic (medieval) diasystem, as it implies diglossia consisting of the Croatian vernacular and CCS. The compiling of the DCRCS adds up critically to another enterprise of the Croatian lexicography, namely the creation of the corpus for the *Old Croatian dictionary* and its compiling. Without the DCRCS, the work of the lexicographers engaged in the project of *Old Croatian dictionary* would be considerably more difficult, even hardly accomplishable in numerous cases, as the CCS features in prevailingly vernacular texts would easily remain unnoticed, which, in turn, would result with a poor diachronical description of the Croatian language.[12] The DCRCS and its corpus offer help to all the scholars who for various reasons need to understand CCS texts. Due to the vast range of the CCS text genres, such scholars may be hagiographers, liturgists, general historians, historians of specific fields (feelings, medicine etc.) ethnologists, and so forth.

3.2 The Contribution of the DCRCS Compiling to the Historical Lexicography in General

In the context of the (international) historical lexicography, the corpus for the DCRCS and the DCRCS itself are one of the pivots of the international (paleo)slavistic lexicography. At the moment, the results achieved within the management of the CCS corpus and the compiling of the DCRCS are the constitutive elements of the work on *Srovnavaci*, but the CCS corpus and the DCRCS are also at the disposal of all the lexicographers working on or with various Church Slavonic corpora. They are both also

[12] Cf. e.g. the difficulties in determining the origin of certain features of language forms used by Bartol Kašić, v. Vrtič (2009:117-118.218-219.291-293).

valuable help for all the linguists engaged with the national Church Slavonic idioms (Bulgarian, Macedonian, Czech, Russian, Romanian, Bosnian, Serbian). As national Church Slavonic idioms have been rightly recognized to be integral components of their respective national standard languages, the CCS corpus and the DCRCS support, within their capacities, the research of various Slavic languages. The presentation of the CCS material in both formats, corpus and dictionary, is most-welcome because CCS in its original text, written in Glagolitic and Old Cyrillic scripts usually proves to be a tough nut for slavists, especially those not versed in dealing with the Slavonic idioms.

4 The DCRCS Dictionary Article

The DCRCS is compiled in accordance with the general lexicographical theoretical knowledge (Zgusta 1971, Sinclair 2003; Svensén 2009), fashioned in accordance with the practical knowledge of the most experienced paleoslavistic lexicographers, those from the Czech paleoslavistic lexicographical tradition (v. Mareš 2007[1962]), but finally formed so that it recognizes and appropriately displays a specific range of the CCS features in semantics, grammatical forms, textological and translational traits (Nazor 1991; Grabar et al. 1991; Vukoja 2012).
The methodology of the DCRCS compiling is based on the compiling methodology of the earlier Old Slavonic dictionary, *Slovník*, in order to achieve the formal lexicographical concordance needed to enable, help and enhance various paleoslavic research.[13] However, the particularities of the CCS material has asked for extensive adaptations, e.g. in the areas of normalisation, differentiation of semantic variants etc. By all means, the DCRCS is made according to the highest standards of the relevant historical dictionaries.
The head of the dictionary article is written in the Glagolitic and Old Cyrillic alphabets, but its body is in the Latin script, except for the Greek parallels of the lemma and pertaining citations. The body of the dictionary article contains Croatian and English translations of the CCS lemma, Greek and Latin parallel lexemes, as well as an encyclopaedic identification (in the Latin language), if the given lemma is an anthroponym, a toponym or a technical term (most often liturgical). Every recognised meaning is accompanied by CCS examples followed by Greek and Latin parallel phrases as well as CCS variants of the lemma, if existing. In choosing the examples, not only semantic variations are sought to be presented, but also a range of different contexts (and text genres) as well as the range of the lemma forms. Special attention is given to the phrasemes, the differences in spelling (words abbreviated under a tilde, shortened by suspension or by omission of the first syllable with which the preceding word ends) and other particularities of the lemma. At the end of the dictionary article, relevant synonyms are enlisted. Also, possible presence of the given lemma in *Slovník*, Miklosich (1862-1865) and ARj is indicated.

[13] For the basic methodological principles of the DCRCS compiling v. Mareš (2007[1962]); also Grabar et al. (1991: VIII-XXX).

5 The Current State of Affairs Regarding the CCS Corpus and the DCRCS Compiling

5.1 The Current State of the CCS Corpus Conversion into Digitized Form

At present, two main paper card-files (the sources and the azbuka card-files) are scanned and preserved in the JPEG format (digitization editor: Marica Čunčić, software: Antonio Magdić, scanned by the Croatian State Archives, ArchivePRO). This way, the safety of the data is largely improved, but very little is done in terms of the digital manageability of the corpus (digital readability and searchability, practicality of the DCRCS' compiling process). For internal use, one of the DCRCS compilers edited the JPEG-formatted sources card-file in a more user-friendly manner, and the DCRCS' compilers are using that version in their daily work. Still, paper card-files are indispensable in the majority of research cases.

At the moment, a combination of factors (among which the lack of sufficient financial support is the most notable one) contributed to the decision of putting on hold the project of digitization, which should end only when the full digital readibility and searchability of the CCS corpus is achieved. Another notable obstacle is that the cards in the card-files are written by more than two dozen different hands, which practically excludes all known Optical Character Recognition (acr. OCR) options, and which requires an exploration of the Intelligent Character Recognition (acr. ICR) possibilities. Still, the collaborators on the project of the compiling of the DCRCS, as well as the authorities of the Old Church Slavonic Institute are constantly looking for a solution that would make digitization viable.

5.2 The Current State of Affairs Concerning the DCRCS Compiling

The fascicles of the DCRCS, one per year, each containing 64 pages, have been published since 1991. So far, 1 (1991)–19 (2012), with the dictionary articles A–ŽRЬTVA (according to the Old Cyrillic alphabet) have been compiled. The first 10 fascicles are bound in Vol. 1. (DCRCS 2000), which has been peer-reviewed as an extraordinary lexicographical accomplishment on several occasions.[14] The fascicles 1–19 exist also in the PDF format, with a limited range of text-searching options. They are available on the Institute's intranet, and placed at the disposal of any interested researcher.

The fascicle 20 is only days away from publishing. Once it is printed, Vol. 2 of the DCRCS is to be bound. Despite their dedication, five lexicographers who are engaged in the compiling of the DCRCS are not able to produce the fascicles at a faster pace, but hopefully with additional collaborators the pace will be intensified in the foreseeable future.

[14] E.g. at the presentation of the DCRCS Vol. I. at the International Slavistic Congress in Ljubljana 2003, also Грковић-Мејџор (2007: 187).

6 Conclusion

The CCS corpus is an indispensable tool for the research of the CCS idiom as well as the prime-quality source for all the scholars who for various reasons need to consult the CCS texts. Its present two formats (paper card-file and JPEG) seek for a thorough digital conversion of the corpus, which is on hold at the moment due primarily to the financial reasons.

The DCRCS with its forthcoming 20th fascicle is in progress, although at a moderate pace, due to the limited number of available lexicographers. Hopefully, the compiling will be intensified in forseeable future.

Despite the difficulties just mentioned, the work on the CCS corpus and the DCRCS compiling should be continued due to its major importance in the context of Croatian as well as (paleo)slavistic studies.

7 References

ARj = *Rječnik hrvatskoga ili srpskoga jezika*, Vol. I–XXIII. *Zagreb: JAZU.* (1880-1976).

Бончев 2002-2012 = архимандрит Анастасий Бончев. *Речник на църковнославянския език*. Т. 1. (А-О; 2002), Т. 2. (П-Я; 2012). София: Народна библиотека «Св. Св. Кирил и Методий».

Borin, L. (ed.). (2002). *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999.* Amsterdam: Rodopi.

Bratulić, J. (1981). Ediciona praksa hrvatskih istraživača i izdavača srednjovjekovnih tekstova u XIX i XX stoljeću (Historijski prikaz). In Д. Богдановиђ (ур.)*Међународни научни скуп: Текстологија средњовековних јужнословенских књижевности, 14-16. новембра 1977.* Београд: САНУ, pp.137-147.

Brozović, D. (1970[1967]). *Standardni jezik.* Zagreb: Matica hrvatska.

DCRCS 2000 = *Rječnik crkvenoslavenskoga jezika hrvatske redakcije.* Vol. I. (*a-vrêdь*.). 2000. Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta.

Димитров, К. (2010). Речник - индекс на Словата на авва Доротей (По ръкопис 1054 от сбирката на М. П . Погодин). Велико Търново: Университетско издателство "Св. св. Кирил и Методий".

Димитров, К. (2013). Авва Доротей. Слова. Среднобългарски превод. Гръцко-български словоуказател. Велико Търново: Университетско издателство "Св. св. Кирил и Методий".

Etymologický = *Etymologický slovník jazyka staroslověnského.* E. Havlová, A. Erhart & I. Janyšková (red.). Praha, Brno: Akademie věd České Republiky, Academia, Tribun EU. (1989-).

Grabar, B., Mareš, F.V. & Mulc, I. (1991). Oblikovanje i sastav natuknice. In *Rječnik crkvenoslavenskoga jezika hrvatske redakcije. (sveščić 1., Uvod).* Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta. pp. VIII-XVIII. (transl. in Eng. pp. XIX-XXX.)

Grabar, B., Nazor, A. & Pantelić, M. (1973). Missale Hervoiae ducis Spalatensis croatico-glagoliticum = Hrvatskoglagoljskimisal Hrvoja Vukčića Hrvatinića = Croato-glagolitic Missal of Hrvoje, Duke of Split: transcriptio et commentarium. V. Štefanić (red.). Zagreb, Ljubljana, Graz: Staroslavenski institut, Mladinska knjiga, Akademische Druck- u. Verlagsanstalt. + Faximil.

Грковић-Мејџор, J. (2007). Rječnik crkvenoslavenskoga jezika hrvatske redakcije, I. svezak (a-vrêdь), Staroslavenski institut, Zagreb, 2000. *Прилози за књижевност, језик, историју и фолклор,* LXXII/1-4, pp.185-187.

Jagić, V. (1879). Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus. Characteribus cyrillicis transcriptum notis criticis prologomenis appendicibus auctum adiuvante summi ministerii Borussici liberalitate edidit V. Jagi. Berolini: Apud Weidmannos.

Илиева, Т. (2013a). Старобългарският превод на Стария завет, том 3: Старобългарско-гръцки словоуказател към Книгата на пророк Иезекиил. София: Кирило-Методиевският научен център, Българска академия на науките.

Илиева, Т. (2013b). *Терминологичната лексика в Йоан-Екзарховия превод на* "De Fide orthodoxa". София: Самиздател.

Katičić, R. (1992). 'Slovênski' i 'hrvatski' kao zamjenjivi nazivi jezika hrvatske književnosti. In *Novi jezikoslovni pogledi*. Zagreb: Školska knjiga, pp. 312-328.

Mareš, F.V. (2007[1962]). Návrh přípravných prací pro slovník jazyka církevněslovanského. *Církevněslovanská lexikografie 2006*. In Václav Čermák (sestavil); E. Bláhová, E. Šlaufová & V. Čermák (eds.) Praha: Slovanský ustav AV ČR, Euroslavica, 64-84. (Russian translation: Мареш, Ф.В. (1966). Проект подготовки словаря церковнославянского языка. *Вопросы языкознания XV.* Москва: Академия наук СССР, Институт языкознания, пп. 86-99.)

Mihaljević, M. (2010). Položaj crkvenoslavenskoga jezika u hrvatskoj srednjovjekovnoj kulturi. Свети Наум Охридски и словенската духовна, културна и писмена традиција (организиран по повод 1100-годишнината од смртта на св. Наум Охридски). Зборник на трудови од Меѓународниот научен собир. Охрид, 4-7 ноември. Скопје: Универзитет „Св. Кирил и Методиј", pp. 229-238.

Miklosich F. (1862-1865). *Lexicon palaeoslovenico-graeco-latinum*. Vienna: Guilelmus Braumueller.

Nazor, A. (1991). Uvod; Popis izvora; Navedena literatura. In *Rječnik crkvenoslavenskoga jezika hrvatske redakcije. (sveščić 1., Uvod).* Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta. pp. I-III (transl. in Eng.: pp. IV-VII.).XXXI-XXXVI.XXXVII-XXXIX.

Nazor, Anica. (2008). Rječnik crkvenoslavenskoga jezika iniciran na IV. međunarodnom slavističkom kongresu u Moskvi 1958. godine (u povodu 50. obljetnice inicijative). In M. Samardžija (ed.) *Vidjeti Ohrid. Referati hrvatskih sudionika i sudionika za XIV. međunarodni slavistički kongres (Ohrid, 10.-16. rujna 2008.).* Zagreb: Hrvatsko filološko društvo, Hrvatska sveučilišna naklada, pp. 65-82.

Pantelić, M., Nazor, A. (1977). Uvod; Bibliografija. In *Drugi novljanski brevijar: hrvatskoglagoljski rukopis iz 1495.* Phototypical edition. Zagreb: Staroslavenski institut "Svetozar Ritig", Turistkomerc, pp. 7-37.

Řeckostaroslověnsky = Řeckostaroslověnsky index. Index verborum graeco-palaeoslovenicus. Praha: Slovansky ustav AV ČR, Euroslavica. (2008-).

Речник = Речник на црковнословенскиот јазик од македонска редакција. Том I. Вовед. А–Б. Скопје: Институт за македонски јазик. (2006).

Rosenwein, B.H. (2006). *Emotional Communities in the Early Middle Ages*. Ithaca: Cornell University Press.

Sinclair, J. (2003). Corpora for lexicography. In P. van Sterkenburg, (ed.) *A Practical Guide to Lexicography*. Amsterdam, Philadelphia: Benjamins Publishing Company, pp. 167-178.

Словарь = Словарь древнерусского языка (XI–XIV вв.). В 10 т. Москва: Институт русского языка Российской академии наук. (1988-).

Slovník = Slovník jazyka staroslověnskeho. Lexicon Linguae Palaeoslovenicae. I.-IV. Praha: ČSAV Slovansky ustav. (1958–1997).

Slovník V. = *Slovník jazyka staroslověnského*. Sv. V. (Addenda et Corrigenda). Praha: Euroslavica. (2010-).

Српскословенски = Српскословенски речник јеванђеља. Огледна свеска. Саставио: Виктор Савић. Уредник: Гордана Јовановић. Београд: Институт за српски језик САНУ. (2007).

Старославянский = Старославянский словарь (по рукописям X–XI веков). Под редакцией Р.М.Цейтлин, Р.Вечерки и Э.Благовой. Москва: Славянский институт академии наук Чешской республики, Институт славяноведения и балканистики Российской академии наук, "Русский язык". (1994).

Svensén, B. (2009). *A Handbook of Lexicography.* Cambridge: Cambridge University Press.

Štefanić, V. (1962). Problem rječnika južnoslavenskih redakcija staroslavenskog jezika. In *Slovo*, 11-12, pp. 181-187.

Тасева, Л. (2010). Триодните синаксари в средновековната славянска книжнина. Текстологично изследване. Издание на Закхеевия превод. Словоуказатели (Monumenta linguae slavicae dialecti veteris LIV). Freiburg im Briesgau: Weiher Verlag.

Totomanova, A-M. (2012). Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Historical Dictionary. In *Studia Ceranea*, 2, pp. 219-229.

Vrtič, I. (2009). *Sintaksa Kašićeva prijevoda* Svetoga pisma. PhD. thesis. Filozofski fakultet Sveučilišta u Zagrebu, Zagreb, Croatia.

Vukoja, V. (2012). O korpusu Rječnika crkvenoslavenskoga jezika hrvatske redakcije i njegovu odnosu prema korpusima hrvatskoga jezika. In *Filologija*, 59, pp. 207-229.

Weinreich, U. (1954). Is a structural dialectology possible?. In *Word*, 10, pp. 388-400.

Zgusta, L. (1971). *Manual of Lexicography.* The Hague, Paris – Prague: Mouton – Academia.

Others

Considerations about Gender Symmetry in the Dictionary of Bavarian Dialects in Austria

Isabella Flucher, Eveline Wandl-Vogt, Thierry Declerck
Austrian Academy of Sciences, ICLTT, Austria;
DFKI GmbH, Language Technology Lab, Germany
bella_flu@yahoo.de; eveline.wandl-vogt@oeaw.ac.at; declerck@dfki.de

Abstract

This poster summarizes the first results of a study that has been pursued as part of an internship at the Austrian Academy of Sciences. The aim was to investigate cases of gender symmetry or asymmetry in a dictionary. As case study we focused on a traditional dialectal dictionary. An annotation schema has been developed and first natural language processing steps have been established. In this poster, forms of gender symmetry as well as asymmetry are presented, on the basis of analysis of the vocabulary employed in example sentences or excerpts used in the dictionary. The analysis of gender asymmetry was also based on the consideration of a selection of derogatory names. The work described in this poster provides a critical insight into lexicographical work, design and implementation from a feminist perspective and opens new perspectives for the development of gender-symmetric lexicographic works.

Keywords: Dialectal lexicography; Gender asymmetry; Gender symmetry; Austrian dialects

1 Introduction

This paper presents work achieved in the context of a practical training at the Austrian Academy of Sciences (ÖAW) in summer 2013[1], and which has been pursued afterwards as part of a university seminar. A task designed for this internship was to analyze a traditional dialectal dictionary along the lines of gender-specific criteria.

While we know that the primary goal of a dialectal dictionary is to describe the authentic language use in a certain region, we consider this investigation on gender symmetry (or asymmetry) to be well motivated since the foundations of the dictionary we are considering were laid down well before any feministic concerns in the field of lexicography have been raised. In this paper describing the poster, we give first a brief description of the dictionary we have been selecting for the investigation on gender asymmetry, before presenting a selection of results.

[1] In the context of this internship, Isabella Flucher, the main author of this poster, was collaborating with 3 other students, namely Nathan Balaz, Magdalena Schwarz and Andrea Steiner.

2 The Dictionary of Bavarian Dialects in Austria

For our work, we focused on language data contained in a traditional dialectal dictionary: The dictionary of Bavarian dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich, WBÖ). This traditional scientific territorial dictionary was developed since the early 20th century: The main part of the collection took place between 1915 and 1950; the dictionary itself is published since the early sixties. More recently, in the early nineties, a project has been established for developing and maintaining an integrated database (Database of Bavarian dialects in Austria; Datenbank der bairischen Mundarten in Österreich, DBÖ[2]). This database supports the storage, visualization und querying of a variety of dialectal language data and related information sources. Since 2004 the dictionary is build up as a digital platform and is now also available online via the Austrian Academy Press. Since 2013 we are developing a machine readable version (using SKOS[3] as the basic representation formalism), connecting also the data to the LOD.[4]

The whole project related to WBÖ gives thus an example for a transformation process of an encyclopedic dictionary type into the framework of cyberscience and digital humanities. .

2.1 The Project Framework at the Austrian Academy of Sciences

The investigation on gender (a)symmetry is embedded into the digital dictionary project, which we very briefly described above. In the framework of this project we are working on / with different methods and interdisciplinary knowledge to improve data access, enrichment and re-usability of data. In order to investigate if we could reduplicate results of the analysis on gender symmetry applied to the WBÖ, we decided to create a corpus containing annotation about the (semantic) genders and the type of vocabularies used. The development of this corpus is also aiming at supporting the development of natural language processing tools that could be trained on this set of annotations. Results of this work will be described in future publications, while we concentrate in this poster on the result of manual analysis applied to the annotated corpus.

2.2 The Gold Corpus

A basis for our work on gender symmetry is a WBÖ-XML-gold corpus, methodologically discussed and completely manually annotated: We were using 4 reference supplements of the WBÖ, namely 33-36. The first step of annotation included annotating headwords as well as reference entries that are

[2] The DBÖ collection of mushrooms and related lexicographical materials is available online; see (Wandl-Vogt 2010).

[3] SKOS stands for “Simple Knowledge Organization System” (http://www.w3.org/2004/02/skos/)

[4] See (Wandl-Vogt 2008) and(Wandl-Vogt & Declerck 2013). LOD stands for Linked Open Data (http://linkeddata.org/)

connected to the concept "person". We annotated real persons as well as figures, such as legendary creatures or fairy-tale figures. The (informal) schema for the annotation was:
<h> </h> - Mensch (*human*)[5]
<f> </f> - feminine (*feminine*)
<m> </m> - maskulin (*masculine*)
<Bsp> </Bsp> - Beispiel (immer kursiv– dialektale Ausdrücke) (*marking an example in the dictionary, which is carrying a gender information*)
<Bed> </Bed> - Bedeutung (*same as above, but for a text span dealing with a definition*)
<hist> </hist> - historisch (*same as above, but marking an historical context*)
<gauspr> </gauspr> - gaunersprachlich (*language of the „crooks"*)
<geschlT> </geschlT> - Geschlechtsteil (*words on genital parts*)

A few examples of annotation are given below:

- <Bsp> es mit <m>einem</m>/<f>einer </f> tun| </Bsp> (*to do it [namely: have sex] with someone [namely: him or her]*)
- <Bed> es treibt <f> ihr</f>(vor Scham) d. Röte ins Gesicht</Bed> (*she is blushing*)
- <Bsp>¿ <m>e </m>verdrosch seine<f>Mutter</f> </Bsp> (*he was beating his mother*)
- <m><geschlT>Penis</geschlT></m>

The manual analysis of the annotated data was done on the base of theoretical aspects described in the next section.

3 Gender Symmetry and Gender Asymmetry

3.1 Linguistic Perspective

The feminist linguistics aims to make visible an androcentric predominance, which is resulting in a critique of the language system on the one hand and of the language use on the other hand.[6]
The semantic field "human" is dominated by men. In that context, Pusch (1984) describes the woman as a "subclass of the men's class" and she goes even further when she says: "Human is the man".[7] Often, by "people" or "human" only man is meant, therefore it seems necessary to identify the woman with the female attribution. Thus, masculinity is enhanced in contrast to femininity, when the man is represented as a human being.[8] The attribution of the female will be used, because otherwise only the male group would be targeted, although it is a gender- neutral term.[9]

[5] Annotation of this concept has been performed by Natahn Balasz.
[6] See (Kollmann 2010: p. 14).
[7] See (Pusch 1984: p. 17).
[8] See (Pober 2007: pp. 408-409).
[9] See (Breiner 1996: p. 79).

In terms of the article entries of the WBÖ one notice that definitions which include the keyword "human" often contains a definition of woman. What looks at first glance like an imbalance in favor of the female sex, after closer analysis, is to be interpreted as the opposite phenomenon. The woman is called that often rather for the reason of a differentiation, whereas the man must not also be named because he embodies the prototype of human, the "universal human".

3.2 Examples for Gender Symmetry and Gender Asymmetry

Aiming at gender symmetry means not to reverse a possible androcentric language in a gynocentric one, but to reach a state in which the same criteria are applied to both sexes.

A positive example of a WBO entry to illustrate the case of symmetry is the following: The verb "trackeln , - gg -" means being stupid. Its derivatives are symmetrical in relation to the two sexes, both in the length of the entry as well as their semantic shape: "Trackle , - gg -" : stupid woman [..] and "Trackler ,-g (g) -" : stupid man.[10]

A counterexample is "Trab" and "traben".[11] "Traben" has besides its first meaning of a horses walk two other meanings in their sample-sentences, where a clear gender asymmetry is found. Namely, when *he* goes on the "Trab", he is sent abroad, but when *she* goes on the "Trab", she works as a prostitute. This difference is confirmed by the term "Trab", as the male form "Traber" is referred to herein and the trotting horse, whereas the female "Traberin" stands for a prostitute.[12]

Gender asymmetry can emerge as well from unequal treatment of the length of masculine and feminine entries, for example when the male thief "Dieb" has six columns and the female thief "Diebin" has only a slim column.[13] Also within an entry the example sentences in relation to each other manifest a different rating of the sexes. For example "Ferdienst" has such various ratings of the sexes: three example-sentences reflect the man as an appreciative, rewarding power, whereas the fourth rating, which refers to the woman, is the earning that comes from a "dirty business", which could refer to prostitution again.[14]

Another form of asymmetry is to be shown on the basis of the lemmas "Trantsch", "Träntsche" and "Träntscher"[15] : "Trantsch" and "Träntsche" provide both in its primary meaning an insult mainly for women. The term "Trantsch" has six female classifications and four neutral, which means related on both gender. Thus, referring to this, most often the woman is meant with this insult, the man is not explicitly marked, only indirectly by writing "human". "Träntsche" has the same meaning, this time

[10] WBÖ, S. 234.
[11] WBÖ, S. 220-221.
[12] WBÖ, S. 221.
[13] WBÖ, S. 35-43.
[14] WBÖ, S. 55-56.
[15] WBÖ, S. 314-316.

with the mention of "woman" and "person".[16] Even "Träntscher", the male modification of "Trantsch", refers only to "human" and one time to a "person". These examples are unbalanced from a gender perspective, because the gender-classification of these three abusive terms, which belong together, is much more female-oriented.

3.3 Distribution of Gender (a)symmetric Cases across Topics

Current work is dedicated to the establishment of classes of topics in which a gender (a)symmetry in WBÖ can be established, looking at examples used for illustrating the meanings of entries. The topics we are studying are for now are "alcohol", "talkativeness" and "violence". The aim is to examine, if there are female- and male-specific categories of meaning in the WBÖ, and if so to analyse them in terms of gender criteria.

3.3.1 Alcohol

The word field of drunkenness is clearly dominated by men. Alcohol use of women is only manifested in the individual cases as "Trinkerin"[17] (female form of drinker), "Alkoholikerin"[18](a female alcohol-addict) and "Schnapsdorothea"[19] ("Schnaps" as a sort of strong alcohol combined with the woman's name Dorothea, means a woman, that drinks a lot). But what is significant, is that there exist as good as no female example-sentences referring to alcohol consumption.

- Der Lump, .. der sein ganz´s Geld versauft[20] (*a man, who spends all his money for alcohol*)
- ols a nüachta [nüchtern] is er eh recht söltn aonztreffn[21](he is rarely sober)
- er hat a weng z´tief ins Glasl g´schaut er hat einen Rausch[22] (*he is drunk*)

3.3.2 Talkativeness

Talkativeness is the one meaning category, which is attributed widely to women. It is also remarkable that the context of meaning differs regarding the sexes, where "Tratschweib" (a talkative woman in a negative sense) faces the "Maulheld" (a talkative man, but who is called a hero, a boaster).

- *Postentragerin:* Frau, die andere Personen ausrichtet, abschwächend für Verleumderin[23] (*a woman who speaks about a person in a defaming way*)
- die ist eine rechte/alte Tratsche[24] *(she talks a lot and she is old)*

[16] To which extent the term "person" suggests "woman" as well is debatable, but this is beyond the scope of this paper. However, "person" appears increasingly in association with "woman" and "man" in connection with "human".

[17] WBÖ, S. 520, S. 523.

[18] WBÖ, S. 523.

[19] WBÖ, S. 189.

[20] WBÖ, S. 238.

[21] WBÖ, S. 373.

[22] WBÖ, S. 45.

[23] WBÖ, S. 290.

[24] WBÖ, S. 332.

- Sie soll ná was drein rödn../ Da is má nöt bang, wir [werde] ihr´s Mudl schan tedten - / I kimm ihr schan grob gnua[25] *(she shouldn´t interfere my talk, if she does i will be aggressive against her)*

3.3.3 Violence

In the WBÖ-entries almost examples for words of violence are given by sentences that show violent-acting men. Women are more connected with a softened version of violence, as they are more often called böse (*evil*), launenhaften (*capricious*), zänkischen (quarrelsome), streitsüchtigen (contentious) women and wives.

- Töterling: grober Kerl, Raufbold[26] (*a rough man, sb.who likes to beat/fight*)
- töten ęa wiad des mādl (Mädchen) no ǫwidra´n[27] (to kill he will kill the girl)
- daß sie von ihm noch ihre Treff (Schläge) kriegen werde[28] (*that she will get beaten up by him*)
- Drache: zänkische, herrschsüchtige Frau; streitlustige Ehefrau[29] (*dragon: a quarrelsome, dominant woman; a contentious wife*)

We are currently extending the list of concepts, either adding new ones, or further specifying existing ones. For this we are consulting work by Dornseiff (2004) and lexical resources like WordNet[30] or Wiktionary[31], which are helping us in better classifying the words used in the dictionaries for describing the entries.

4 Conclusion

We presented actual work on gender (a)symmetry in the context of a dialectal dictionary. In the next future we will extend this work and also consider the underlying data of similar dictionaries, like collections of slips of papers, databases, or questionnaires to find out reasons of asymmetries. We also plan to focus more on natural language processing aspects and to be able to mark up relevant words in computational lexicons with this kind of gender interpretation, beyond the case of pure grammatical genders.

The work is to be continued, deepened, and extended as a research infrastructure for the comparison of lexicographical works as well as languages.

[25] WBÖ, S. 210.
[26] WBÖ, S. 212.
[27] WBÖ, S. 248.
[28] WBÖ, S. 369
[29] WBÖ, S. 222-223.
[30] http://www.sfs.uni-tuebingen.de/GermaNet/
[31] https://de.wiktionary.org/wiki/Wiktionary:Hauptseite

5 References

Barabas, Hareter-Kroiss, Hofstetter, Mayer, Piringer, Schwaiger. (2010). Digitalisierung handschriftlicher Mundartenbelege. Herausforderungen einer Datenbank. In: *Germanistische Linguistik 199–201,* Fokus Dialekt. Festschrift für Ingeborg Geyer zum 60. Geburtstag 2010, S. 47-64.

Bayerisch-Österreichisches Wörtberbuch, I. Österreich, Wörterbuch der bairischen Mundarten in Österreich, Institut für Dialekt- und Namenlexika DINAMLEX, Österreichische Akademie der Wissenschaften ÖAW (Hg.), Wien 1963.

Breiner, I. (1996). Die Frau im deutschen Lexikon. Eine sprachpragmatische Untersuchung, Wien 1996.

Dornseiff, F. (2004) *Der deutsche Wortschatz nach Sachgruppen.* De Gruyter, Berlin/Leipzig 1933–1940; 8. Auflage: De Gruyter, Berlin/New York.

Hufeisen, B. (editor) (1993). „Das Weib soll schweigen…“ (I.Kor.14,34). Beiträge zur linguistischen Frauenforschung, Kasseler Arbeiten zur Sprache und Literatur, Band 19, Frankfurt am Main 1993.

Eichhoff-Cyrus K.M. (2004), Adam, Eva und die Sprache. *Beiträge zur Geschlechterforschung, Thema Deutsch, Band 5,* Mannheim 2004.

Flucher, I. (2013). Gendersymmetrie im WBÖ Wörterbuch der bairischen Mundarten in Österreich, Seminararbeit Uni Wien (Pober M.: „toller hengst“ : „läufige hündin“ - Zufall oder verborgenes Genderskript?

Kollmann, S. (2010). Einstellungen zu geschlechtergerechtem Sprachgebrauch im Deutschen, Dip., Wien

Pober, M. (2004). Überlegungen zur geschlechtersymmetrischen Struktur eines Genderwörterbuchs im Deutschen, Dissertation, Wien 2004.

Pober M. (2007), Gendersymmetrie. Überlegungen zur geschlechtersymmetrischen Struktur eines Genderwörterbuches im Deutschen, Würzburg 2007.

Pusch, L.F. (1984). *Das Deutsche als Männersprache*. Frankfurt am Main.

Wandl-Vogt, E. (2008). wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX). In: Ernst, P. (ed) 2008, *Bausteine zur Wissenschaftsgeschichte von Dialektologie / germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen,* Wien: 93-112.

Wandl-Vogt, E. (2010). Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema). Accessed at: http://wboe.oeaw.ac.at [10/04/2014].

Wandl-Vogt, E., Declerck, T. (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data. In. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-). Verlag der Österreichischen Akademie der Wissenschaften. Wien. Accessed at: http://hw.oeaw.ac.at/cl?frames=yes [10/04/2014].

Acknowledgements

We would like to thank Nathan Balaz, Magdalena Schwarz and Andrea Steiner for their contribution to the internship, which constituted the basis for the work described in this poster.

Advancing Search in the Algemeen Nederlands Woordenboek

Carole Tiberius, Jan Niestadt, Lut Colman, Boudewijn van den Berg
Institute of Dutch Lexicology, belberg
{carole.tiberius,jan.niestadt,lut.colman}@inl.nl, boudewijn@belberg.eu

Abstract

In this paper, we will take a closer look at the advanced search option in online dictionaries from the perspective of the user. In the context of dictionaries, the advanced search allows users to retrieve sets of words which match a particular description, for example 'archaic compounds consisting of three syllables'. However, the possible sets of words that could be retrieved are endless, and the challenge is how to present all these options to the user in a way that he can grasp and understand. Studies on dictionary use and log file analyses suggest that existing solutions offered by online dictionaries are not very successful as users do not really seem to use this search option. We will discuss different types of advanced search that can be distinguished and we will present a new, more user-friendly approach to advanced search in dictionaries using the *Algemeen Nederlands Woordenboek* as our test case.

Keywords: advanced search; online dictionaries; user-friendly.

1 Introduction

Comprehensive scholarly dictionaries contain a wealth of information. They do not only provide information on the meaning of a word, but they also contain information on morphology, pronunciation, etymology, pragmatics etc. With the online medium it is theoretically possible to let the user search on all the information available in the dictionary database. Most online dictionaries attempt to do this by offering an advanced search option allowing users to retrieve sets of words matching a particular description, for example 'the set of Dutch nouns that can have more than one plural ending' or 'archaic compounds consisting of three syllables'. However, the sets of words that could possibly be retrieved are endless, and the challenge is how to present all these options to the user in a way that he can understand. Studies on dictionary use suggest that users do not really use the advanced search option. This is also the case for the Algemeen Nederlands Woordenboek (ANW) where log files show that the advanced search only accounts for 3% of all searches in the dictionary (Tiberius & Niestadt, To Appear). There may be different reasons for this: a) users do not use the advanced search as they are not familiar with such a search option in the context of dictionaries; b) users do not understand the interface that is used for the advanced search; or c) a mixture of these.

In this paper we will present an approach for a more user-friendly advanced search option for the ANW. First we define what advanced search is, then we will discuss different types of advanced search and finally we will present a new approach to advanced search in dictionaries.

2 Advanced search in electronic dictionaries: a definition

In order to define what is meant by advanced search and what user requirements it fulfils, we have looked at what different dictionaries have to say about their advanced search. The **Oxford English Dictionary** (OED) defines its advanced search as follows:

> Advanced search is a full search of the entire dictionary text. It finds your term wherever it occurs in the dictionary. This could be in the form of an entry name, part of another word's definition, in a quotation, etc. An advanced search also allows you to search for words that occur near one another, such as bread before butter. [1]

The German **elexiko** dictionary offers an advanced search which allows users to search for words on the basis of specific criteria such as orthography, grammar and word family.
Die erweiterte Stichwortsuche in elexiko erlaubt dem Benutzer**,** Stichwörter mit bestimmten Kriterien aus den Bereichen Orthografie, Wortartzugehörigkeit, Grammatik, Sinnverwandtschaft oder Zugehörigkeit zu einer semantischen Klasse zu suchen.[2]
The **Trésor de la Language Française Informatisé** (TLFi) offers what it calls an assisted search ('recherche assistée') and a complex search (recherche complexe'). The assisted search allows the user to search the dictionary articles on the basis of a number of criteria:

> Permet de rechercher **à travers tout le TLF** les articles correspondant à **plusieurs critères**.
> Quelques possibilités:
> Quels sont les mots d'origine espagnole ?
> Quels sont les exemples de Zola illustrant un sens ironique?
> Quels sont les verbes utilisés dans la marine pour la manoeuvre des voiles?
> Quelles sont les expressions contenant le mot singe? [3]

The complex search in the TLFi is similar to the assisted search, but is described as being even more powerful.
The **Algemeen Nederlands Woordenboek** does not employ the term advanced search, but offers four types of search of which the option to search from features to words is the most advanced.

[1] http://www.oed.com/public/advancedsearching/advanced-search/ [10/04/2014]
[2] http://www.owid.de/erweitert.jsp [10/04/2014]
[3] http://atilf.atilf.fr/dendien/scripts/tlfiv5/showp.exe?13;s=2657178285;p=aide.htm [10/04/2014]

> This search option is the most advanced way of searching the ANW. Through different kinds of information that is stored in the dictionary articles you can search for words, idioms and proverbs. The possibilities to search for information are almost infinite. You can search for information which can occur anywhere in the article or you can search for information in a specific field.[...]

This is a search option which requires a certain amount of creativity from the user and works better the more one gets familiar with the system. We suggest that you take some time to try out this search option.[4]
The last sentence shows clearly that we were rather idealistic when we first developed this search option for the ANW back in 2009.

Summarising, advanced search can be described as a powerful and complex search option that allows users to examine the dictionary using many different criteria.

3 Types of Advanced Search

Advanced search can be realised in different ways. We identify four types, i.e.

- **Classical/traditional advanced search:** where boxes and dropdown lists together form the search query;
- **Faceted search:** a step by step search where the user gradually refines the query by adding criteria (e.g. shopping websites);
- **Wizard search:** a step by step search where the user is given a sequence of questions and no intermediate results are shown (e.g. Foreign Labor Certification[5]);
- **Query language:** single search box which offers many possibilities, but which is hard to learn and remember.

A fifth type could be identified, i.e. natural language queries. However, we do not consider this option here, as we do not believe that the current state of technology is advanced enough for this to be a viable candidate in the context of online scholarly dictionaries.

From these four types, the classical advanced search is the most popular among scholarly e-dictionaries at the moment (cf. the ANW, the OED, *elexiko* and the complex search in the TLFi). The TLFi also offers a wizard-like search ('recherche assistée') by showing a sequence of questions which are to guide the user to an answer.[6]

The DWDS (das *Digitale Wörterbuch der deutschen Sprache*) is an example of a dictionary that does not offer a separate advanced search option but offers a query language to give the user more flexibility. For instance, the query "Stein with $p=NN" searches for occurrences of the lemma *Stein* 'stone' as a noun.

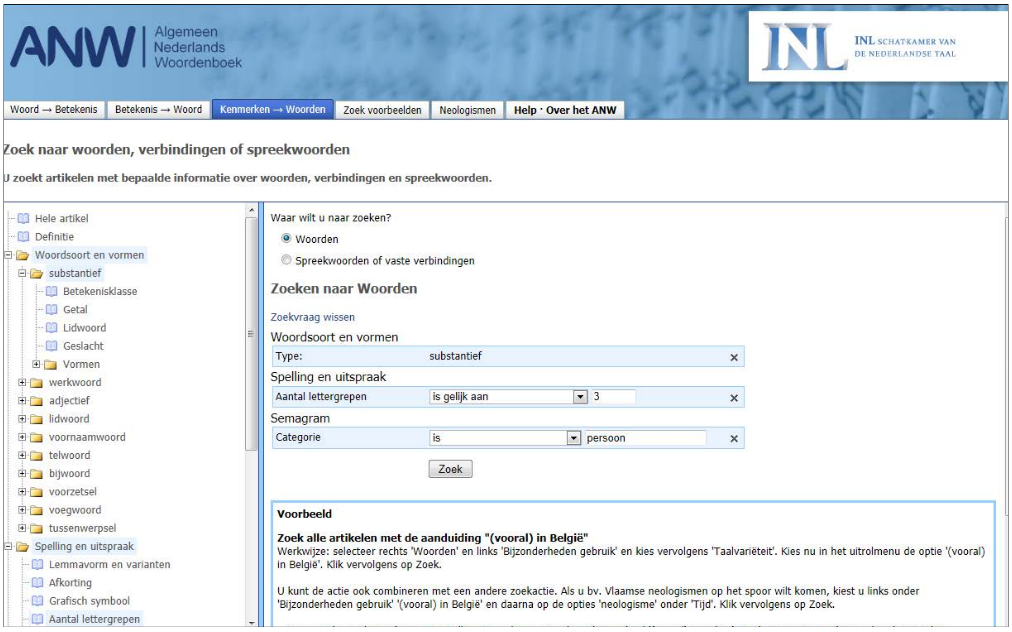
For the ANW too, a custom query language, called FunQy ('functional query language'), was developed to power its traditional advanced search option (Niestadt et al. 2009). As an undocumented feature, users can play with this query language themselves. At one time we planned for this query language to

[4] http://anw.inl.nl/show?page=help#zoek3 [10/04/2014]
[5] http://www.flcdatacenter.com/OESWizardStart.aspx [10/04/2014]
[6] http://atilf.atilf.fr/dendien/scripts/tlfiv5/showp.exe?30;s=1771472760;p=assiste.htm [10/04/2014]

be used internally at the Institute of Dutch Lexicology (INL), but this never materialised. The FunQy query language may be too complex even for language professionals to use.

4 A new way of advanced search

We now turn to our approach to realise a more user-friendly advanced search option for scholarly dictionaries. We believe that it is a mistake to think that one single advanced search option can appeal to all users equally. A full-featured search may be convenient for frequent users, but new users will most likely be put off by its complexity. However, if you do your best to capture new users with a friendly, step-by-step approach, experts will get annoyed by how much clicking is required to perform common searches. This means there is no single best approach; you have to compromise, or develop separate interfaces for different users. We decided to clearly identify the target users of the ANW to be linguists and academics more generally, who want powerful search features, but are intimidated by our current interface (see Figure 1).
The features that can be searched for are presented in a tree structure on the left of the screen. This tree structure is the same as the one used to structure the dictionary articles (as seen on the article screen). It starts with syntactic category and then spelling and pronunciation, etc. The user starts with an empty query screen and is asked to select criteria from this tree structure on the left. As soon as the user selects a criterion, a query appears on the right-hand side of the screen. By default, the user searches for words, but it is also possible to search for proverbs or idioms. This will result in a tree structure with different criteria as only a subset of the criteria that can occur in a query for words apply to idioms and proverbs.



Zoek naar woorden, verbindingen of spreekwoorden

U zoekt artikelen met bepaalde informatie over woorden, verbindingen en spreekwoorden.

Waar wilt u naar zoeken?

● Woorden

○ Spreekwoorden of vaste verbindingen

Zoeken naar Woorden

Zoekvraag wissen

Woordsoort en vormen

Type:	substantief	×

Spelling en uitspraak

Aantal lettergrepen	is gelijk aan	3	×

Semagram

Categorie	is	persoon	×

Voorbeeld

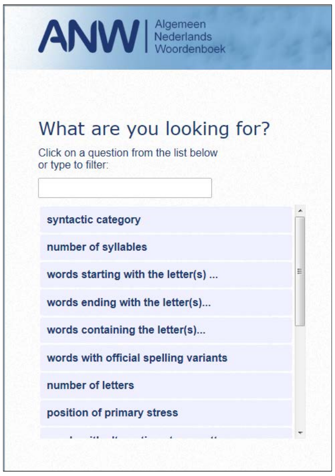
Zoek alle artikelen met de aanduiding "(vooral) in België"
Werkwijze: selecteer rechts 'Woorden' en links 'Bijzonderheden gebruik' en kies vervolgens 'Taalvariëteit'. Kies nu in het uitrolmenu de optie '(vooral) in België'. Klik vervolgens op Zoek.

U kunt de actie ook combineren met een andere zoekactie. Als u bv. Vlaamse neologismen op het spoor wilt komen, kiest u links onder 'Bijzonderheden gebruik' '(vooral) in België' en daarna op de opties 'neologisme' onder 'Tijd'. Klik vervolgens op Zoek.

Figure 1: Screendump Features → Words in the ANW.

Our aim was to combine the best parts of the different types of advanced search discussed in Section 3 in our solution. A very preliminary version of our approach has been presented at the eLex 2013 conference.[7] We are now in the process of developing a full-working prototype which will be accessible through the ANW website.

The opening page of the ANW remains as it is with a simple search box in the center of the screen for searching a word (or multi word expression). However, below the search box a link will be added for users who are looking for something else. After clicking on this link, users are asked what they would like to search for. They can choose a question or criterion from a scrollable list, or they can type in a word to filter the list. The current list covers spelling, pronunciation, combinations and pragmatics, and has been defined on the basis of the criteria offered in the current search interface of the ANW web application. In future, this list will be further expanded and refined.


ANW
Algemeen
Nederlands
Woordenboek

What are you looking for?

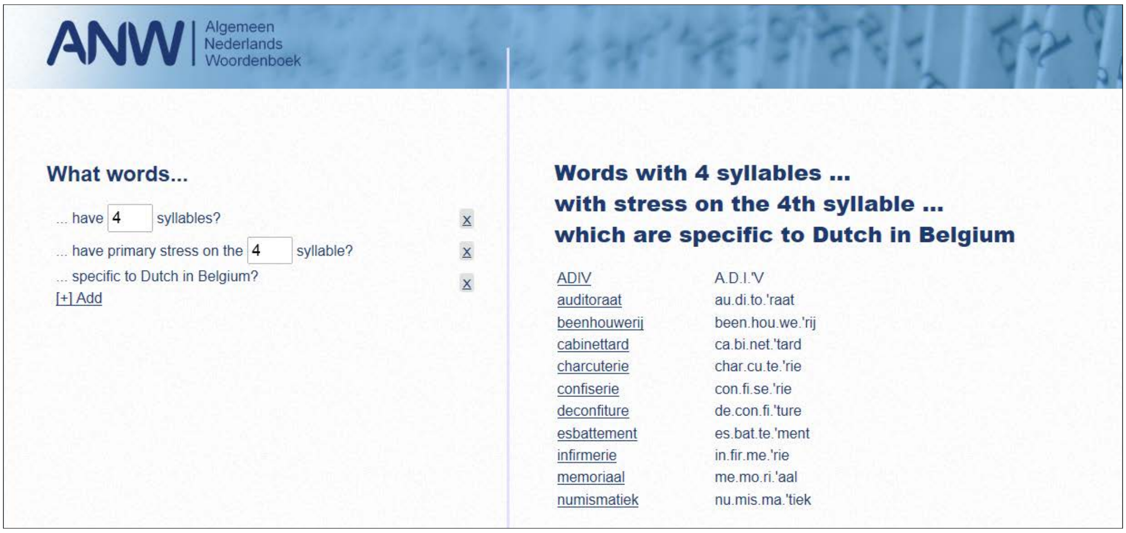
Click on a question from the list below or type to filter:

syntactic category
number of syllables
words starting with the letter(s) ...
words ending with the letter(s)...
words containing the letter(s)...
words with official spelling variants
number of letters
position of primary stress

Figure 2: Prototype of list of questions users may like to ask.

For convenience, the most frequently used questions will appear at the top of the list so that users see these first. To make the search as effective and user-friendly as possible (Lew 2012) functionalities such as autosuggest, fuzzy matching and smart filtering will be integrated. Thus, the filter box will also respond to terms that do not literally occur in the visible descriptions because hidden tags specifying the category to which questions belong have been added (e.g. typing 'morphology' also shows questions such as 'words derived from ...'). Clicking on a question opens up the search form, with a single input box for that question. As soon as the user types something in this search box, the results are shown on the right-hand side of the screen together with the relevant information from the dictionary entry: for instance, if the user was searching for words consisting of four syllables, the syllable structure of the resulting words is shown. This direct feedback makes this 'advanced' search faster and less intimidating as the user has direct access to the information he is interested in.

[7] http://eki.ee/elex2013/ [10/04/2014]



What words...

... have 4 syllables? x

... have primary stress on the 4 syllable? x

... specific to Dutch in Belgium? x

[+] Add

Words with 4 syllables ...
with stress on the 4th syllable ...
which are specific to Dutch in Belgium

ADIV	A.D.I.'V
auditoraat	au.di.to.'raat
beenhouwerij	been.hou.we.'rij
cabinettard	ca.bi.net.'tard
charcuterie	char.cu.te.'rie
confiserie	con.fi.se.'rie
deconfiture	de.con.fi.'ture
esbattement	es.bat.te.'ment
infirmerie	in.fir.me.'rie
memoriaal	me.mo.ri.'aal
numismatiek	nu.mis.ma.'tiek

Figure 3: Prototype of what a search query plus results may look like.

If the user wants to refine his query, he simply clicks 'Add' and goes again to the list of search questions where he can select another question to add to his query. As a bonus, an expert who regularly uses the same criteria with different values can bookmark the page in its current state, so he can avoid the ten clicks required to get here.

5 Conclusion

We have presented an approach for a more user-friendly advanced search option in the ANW. It combines the best aspects of the advanced search types we have discussed in Section 3. It is like a wizard, because the user is guided through constructing a query. It also includes the advantages of a faceted search because there is direct feedback when you add or change something. The possibility to bookmark a search query makes it again very similar to the classical advanced search.
So far user research has not been carried out, but this is planned for the near future when the prototype is up and running.

6 References

Algemeen Nederlands Woordenboek. Accessed at: http://anw.inl.nl [10/04/2014].
das Digitale Wörterbuch der deutschen Sprache. Accessed at: http://www.dwds.de [10/04/2014].
elexiko. Accessed at: http://www.owid.de/wb/elexiko/start.html [10/04/2014].
Lew, R. (2012) How can we make electronic dictionaries more effective? In Sylviane Granger and Magali Paquot (eds.) *Electronic lexicography*. Oxford. 343-361.

Niestadt, Jan, Carole Tiberius & Fons Moerdijk (2009). Searching the ANW dictionary. Poster presented at eLexicography in the 21st century. Louvain-la-Neuve.

Oxford English Dictionary. Accessed at: http://oed.com [10/04/2014].

Tiberius, C and J. Niestadt (To Appear) Dictionary Use: A Case Study of the ANW Dictionary. In: Tiberius, C. and C. Müller-Spitzer (eds.) *Wörterbuchbenutzungsforschung* 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik).

le Trésor de la langue française informatisé. Accessed at: http://atilf.atilf.fr/ [10/04/2014].