PAPERS
IN COMPUTATIONAL LEXICOGRAPHY
COMPLEX 2005

Edited by
Ferenc Kiefer, Gábor Kiss and Júlia Pajzs

Linguistics Institute
Hungarian Academy of Sciences, Budapest

PAPERS IN COMPUTATIONAL LEXICOGRAPHY
COMPLEX 2005

PAPERS
IN COMPUTATIONAL LEXICOGRAPHY
COMPLEX 2005

Edited by
Ferenc Kiefer, Gábor Kiss and Júlia Pajzs

LINGUISTICS INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST
2005

Proceedings of the 8th International Conference on
Computational Lexicography, Complex 2005
Budapest, Hungary

All correspondence should be sent to
Júlia Pajzs
pajzs@nytud.hu
Linguistics Institute
Hungarian Academy of Sciences
Department of Lexicography and Lexicology
Hungary, Budapest
H-1063 Benczúr utca 33.

Cover design by Gábor Kiss

ISBN 963 9074 35 7

CONTENTS

PREFACE

The present volume contains the papers to be presented at the 8th, and most likely the last COMPLEX meeting to be held in Budapest in June 17-18, 2005. At the beginning the COMPLEX conferences were organized jointly by the Research Institute of Linguistics of the Hungarian Academy of Sciences and by the Laboratoire d'automatique documentaire et linguistique (LADL) of the CNRS located at the Université de Paris 7. Maurice Gross, the director of LADL and myself, we felt at that time that computational lexicography was still a relatively neglected field and it needed some advertizing. This feeling was shared by many other researchers of the field. The first conference was held in 1990 in Balatonfüred (at the lake Balaton) with mainly French and Hungarian participants. Very soon, however, the COMPLEX conferences became more and more international and the number of participants was steadily growing (especially in 1992, 1994, 1996). But then, towards the end of the nineties, the interest to have a special meeting devoted to computational lexicography started to diminish. The reason was certainly not the lack of interest in the field itself, but was rather due to the fact that the original function of COMPLEX was taken over in part by the EURALEX conferences and in part by the COLING meetings.

The COMPLEX conferences certainly played an important role in the past, and have contributed to the development of the field in essential ways. The quality of the papers in the present volume, too, testify that the possibilities are still not completely exploited. The papers offer new methodological insights as well as new possibilities for applications, an adequate round-off of our conference series.

Budapest, May 26, 2005

Ferenc Kiefer

Development of a New Implementation of the Unification-Based Morphological Analyzer HUMOR for the Croatian Language

MELITA ALEKSA

FACULTY OF PHILOSOPHY, L. JÄGERA 9, 31 000 OSIJEK, maleksa@ffos.hr

ROBERT WOLOSZ

FACULTY OF PHILOSOPHY, IFJÚSÁG U. 6, PÉCS, robert.wolosz@gmail.com

HUMOR, the morphological parser, has been successfully implemented among others to an agglutinative language, the Hungarian and an inflectional language, the Polish. Developing it for another highly inflectional language, Croatian, has been a great challenge, since the language itself struggles with certain problems concerning language policy, but linguistics as well. The present paper discusses the problems that arise when trying to implement HUMOR to Croatian language, concentrating mainly on verbs. Presenting the present stage of the project and its difficulties will hopefully provide additional help with the implementation of the programme to other South Slavic languages.

Introduction

The present paper discusses the linguistic problems that arise when trying to develop a new application of the existing morphological parser HUMOR. HUMOR, standing for High-speed Unification MORphology, developed by MorphoLogic, has already been successfully implemented among others to a highly agglutinative language – the Hungarian, and an inflectional language – the Polish (Prószéky, Kis 1999). Since the creators of MorphoLogic have argued that the analyzer had been suitable for all kinds of languages and all kinds of operating systems due to its unique system of operation, a new application of HUMOR has undergone further development, namely its implementation to another highly inflectional language – the Croatian. The programme itself has been used as a basis for the translational supporting systems, MoBiMouse and MoBiDic, as well as the MoBiCAT, a sentence analyzer, which is now able to translate sentences from English into Hungarian. Up to the present day, HUMOR has been developed to successfully cope not only with an agglutinative language – Hungarian, but also with Czech, English, French, German, Polish and Romanian. Developing HUMOR for the Croatian language has been a great challenge, since the language itself, belonging to the Slavic group of languages, successfully copes with problems concerning the language politics, but with other linguistic problems, which include the codification of Croatian and its linguistic division form Serbian, i.e. former Serbo-Croatian. The project itself is still under development, and the present paper presents the linguistic problems connected with the implementation of a successful morphological parser for agglutinative languages to a highly inflectional Slavic language, concentrating mainly on verbs. The goal of the paper is to present the present problems that have been encountered so far, but that are present in other South Slavic languages so that the problem of implementing HUMOR to other Slavic languages is clarified, and the process of an implementation of the programme itself to other languages becomes more successful.

Implementation of HUMOR to the Croatian language

HUMOR itself, as the authors (Prószéky, Kis 1999) have already pointed out, has several applications. The main goal is not the development of industrial spelling checkers, hyphenators and thesauri, since these modules have been on the market for several years, but the linguistic parsing of lemmas for searching purposes, as well as the shallow or full parsing in translational supporting systems. The Croatian version of HUMOR, however, will be put to another use, namely the categorization of verbal and nominal inflections, which, when summarized, will provide an additional help for the learning and teaching of Croatian as a second or foreign language. It is important to notice that there are no such works in Croatian, that would clarify and precisely determine verbal and nominal inflections in the Croatian language, whereas the Hungarians are able to rely on the *A magyar nyelv szóvégmutató szótára* (Papp 1969) and the *Magyar ragozási szótára* (Elekfi 1994). These linguistic works provide additional help for the learners of Hungarian, by providing an insight into the morphological system of the language itself. The Croatian implementation of HUMOR will hopefully provide an important basis for developing similar works in Croatian language.

As the basis of the lexical part of HUMOR, the latest Anić's *Rječnik hrvatskoga jezika* (The Dictionary of the Croatian Language) has been used with 60,000 lexical entries. Furthermore, the categorization of the grammatical and inflectional entries has been made upon the Težak – Babić *Gramatika hrvatskoga jezika. Priručnik za osnovno jezično obrazovanje* (The Croatian Standard Language Grammar), Barić (1995) *Hrvatska Gramatika* , Raguž (1997) *Praktična hrvatska gramatika* and the spelling rules upon the Babić – Finka – Moguš's (1996) *Hrvatski Pravopis*, as well as Težak´s works (1991, 1995, 1999, 2000).

Linguistic Problems

The Croatian language, belonging to the Slavic group of languages, encounters, naturally, differences in the whole language system, when compared to some other agglutinative or inflectional languages. Whereas in the Hungarian version of HUMOR the linguists had to solve problems typical for the agglutinative languages, the Croatian version of HUMOR has encountered several difficulties concerning the morphological, syntactic and the semantic domains. Since HUMOR has been used for the analysis of written language only, the phonetical and phonological spheres have not been taken into consideration. When trying to implement HUMOR to the Croatian language, it has been essential to use the existing engine, namely the programme itself, and develop a new database, consisting of several parts. After defining the characters belonging to the Croatian alphabet, it has been essential to make up a lexicon, consisting of a minimum number of entries, which encountered a number of problems. The Croatian language, belonging to the Slavic language group, has most of the time in history been paired with Serbian and was therefore, until 1991, categorized only as the Serbo-Croatian language (Težak 2004). Until that time, there has not been a contemporary Croatian dictionary or a contemporary codification. Since the 19th ct and earlier codifications of the Croatian language have nowadays been considered archaic, the question is whether all the lexical entries in this project should belong to the contemporary standard language. The problem arises when trying to select a valid number of lexical entries for the linguistic corpus to be analyzed. The main problem concerning the authenticity of the further language corpus used is the presence of Serbian lexemes and Serbian versions of words in texts written in Croatian. The contemporary Croatian language is partly an artificial language made up after the disintegration of the Yugoslavia and forced upon the Croats through the media. The language policy nowadays, considers the Croatian and Serbian languages as separate languages and not dialects. However, there have been a number of linguists, who claim that these two languages are merely dialects of the same language, namely Serbo-Croatian (Wardhaugh 1991). In his work, Wardhaugh (1991: 29) argues that the main differences between the Croatian and the Serbian lie mainly in word preferences, and that there are no grammatical or phonetical differences between these two languages. This opinion has been criticized by Croatian linguists, who argue that the differences between the Croatian and the Serbian lie in every language sphere, and include differences in morphology, syntax, semantics, as well as phonetics (Težak 2004). Concerning the contemporary language politics, one can conclude that the Serbo-Croatian was developed either by the process of synthesis or the analysis from these two languages. Since 1991, many works have been written to codify the contemporary Croatian language, which proved the differences between certain variants of words like Brodnjak`s (1991) *Razlikovni rječnik srpskog i hrvatskog jezika*. Words that have been included in this dictionary have been categorized as belonging unexceptionally to the Croatian or the Serbian language. The problem lies in the fact that the oral language, as well as some of the texts written in the Croatian language, still contains words that, according to this dictionary, belong to the Serbian language. If these are left out of the lexical part of the programme, the programme itself will not be able to use most of the corpus written up to 1995. In addition to that, the texts published on the internet will also have to be left out, because of the presence of Serbian words. For example, according to Brodnjak (1991: 411) the Serbian word *ponekad* has its counterpart in the Croatian language, namely *katkad*. Nevertheless, when analyzing Croatian internet pages, the word *ponekad* occurs approximately on 110 000 pages (Google 2005). If the "purist" version of HUMOR is chosen, the language corpus to be analyzed cannot include 20th century texts written before 1991. Since the Hungarian and the Polish versions of HUMOR made it possible to analyze 19th century texts, the question is whether the programme should be implemented in a way that it is possible to do that in the Croatian language too. This would then mean developing a morphological analyzer, which would recognize a great number of lexical entries in texts with a questionable Croatian / Serbian origin or implementing it for the Serbian language as well. The question is whether it can be labelled then as a Croatian version of HUMOR only.

The Basics of HUMOR

After creating the lexical database, there had to be some actions done in order to obtain the linguistic categories, which will later be used by the parser itself. The lexical basis, therefore, consists of a range of specially handled and categorized roots – *stems* and affixes – *terms* (Prószéky, Kis 1999). The traditional expressions root *and affixes* are deliberately not used, since their definitions do not comply with the traditional categories. In our case the *stem* is not the linguistically considered root of a certain word, to which e.g. suffixes can be added, but the part of the word which remains unchanged during the inflections. E.g., the Croatian noun *noga* /leg/ linguistically consists of a ROOT (nog-) + SUFFIX (-a). Nevertheless, in HUMOR, the stem of this word is unanimously *no-*, whereas the term of the noun *noga* is unexceptionally *-ga*, only because the *no-* part of the word remains unchanged during the inflections of the mentioned noun. (The dative case of *noga*, e.g. is *nozi*). The categorization of lexemes and their division into stems and terms only is naturally insufficient There is also a system of codification present which clearly determines all the possible terms that can be added to a certain stem in order to get a meaningful word. Therefore, the last part of the database is actually the linguistic categorization of stems and terms, where each stem is given a certain grammatical category. (In the above mentioned case it would be *no-* → noun, feminine and *-ga* → Nominative, Sg.) It is also essential to mention that the pronunciation rules of Croatian words have not been taken into consideration, although sometimes they have a distinctive function. The reason for that lies in the application of HUMOR to written texts only. The programme itself has not been used for the semantic purposes, i.e. the semantical analyses of words, but for syntactical parsing only. The polysemic and homonymic words are considered as separate lexical entries only when they have alternate inflections (Prószéky, Kis 1999).

Linguistic Dilemmas

When trying to implement HUMOR for the morphological parsing of the Croatian language, several linguistic dilemmas have occurred as well. Whereas the grammar of the Hungarian language describes only three tenses – the present, past and future, the Croatian language operates (although in written form only) with six tenses, three of which describing the Simple Past Tense – *perfekt, aorist* and *imperfekt*, and one for the Past Perfect, Simple Present Tense and the Simple Future Tense, respectively. Apart from that, there have been six conjugational groups of verbs in the Croatian language, which include more than 100 conjugational types. Unlike the Hungarian language, where there are two types of conjugations, namely the subjective and the transitive conjugations, the Croatian language preserves differences according to genders. For example, the Past Tense of the verb *to eat* , according to the gender types in the Croatian is inflected in the following way:

		masculine	feminine	neuter
Sg.	1.	*jeo sam*	*jela sam*	-
	2.	*jeo si*	*jela si*	-
	3.	*jeo je*	*jela je*	*jelo je*
Pl.	1.	*jeli smo*	*jele smo*	-
	2.	*jeli ste*	*jele ste*	-
	3.	*jeli su*	*jele su*	*jela su*

An additional problem lies in the fact that the past in the Croatian language is expressed by two lexemes, separated by the empty character, namely the space, which HUMOR interprets as an analysis boundary.

Apart from the past tense, used in the Croatian language for expressing past actions, there have been additional two tenses mentioned in the grammar books, namely the *aorist* and *imperfekt*. Another problem when considering only the past tenses lies in the fact that *aorist* and *imperfekt* have nowadays been considered archaic, and therefore out-of-use, but are still implemented in the grammatical descriptions of the Croatian language, although with inadequate and insufficient information. The problem still lies in the contemporary use of these forms. Namely, Barić in his book mentions verbs like *peći, sjeći* and *strići*, which leads us to certain ambiguous imperfekt forms, namely *pecijah / pečah, sjecijah / sječah* and *strigah / strizijah.*

The archaic forms of the verbs occur not only in the past tense, but when expressing the present tense as well. The inflected forms, e.g. of the verb *gnjiti* in the present tense are: *gnjijem, gnjiješ, gnjije, gnjijemo, gnjijete, gnjiju* (Barić 1995). Unlike Barić, Anić also mentions an additional form, *gnjim*, which then leads us to the questionable 3rd p. Pl. form of *gnju* or *gnjiju.*

Apart from verbal, there have also been some distinctions in the nominal forms of the words when considering the Croatian and Hungarian language. A major distinction between the Hungarian and the Croatian lies in the agglutination, i.e. inflections. There are seven cases in the Croatian language, but unlike in Hungarian, an agglutinative language with the defined position of affixes, the inflected nouns in the Croatian language make up different cases with the help of prepositions as well. The analysis of a syntactical concordance of e.g. a PREP + NOUN, however, belongs to further applications of HUMOR.

The problem that arises, when trying to define the cases and the nominal inflections are the double/ triple forms in some cases. The best way to illustrate that is to take a look at the inflections of numbers, that have to be in concordance with the inflected nouns that follow. E.g., in the sentences *Ovdje je jedan pas* (Here is a dog), *Ovdje je jedna mačka* (Here is a cat), *Ovdje je jedno dijete (*Here is a child) *jedan* (m), *jedna*(f) and *jedno* (n) are inflected in the following way, depending on the gender of the noun.

masculine	feminine	neuter
N: *jedan,*	*jedna*	*jedno*
G: *jednog/jednoga,*	*jedne*	*jednog /jednoga*
D:*jednom/jednomu/jednome*	*jednoj*	*jednom/jednomu /jednome*
A: *jednog/jednoga*	*jednu*	*jedno*
V: *jedan*	*jedna*	*jedno*
L: *jednom/jednome*	*jednoj*	*jednom/ jednome*
I: *jednim*	*jednom*	*jednim*

The problem lies in the fact, that the dative forms are interchangeable. Another problem, when implementing HUMOR to the Croatian language lies in the definition of abbreviations, full-stops, spaces, name strings, gerunds, as well as already mentioned homonymic pairs of words, all of which will be handled according to the frequency of their use in the contemporary Croatian texts.

Conclusion

All the issues described in this paper have been encountered until the present stage of the whole project, namely in the course of the morphological analysis of verbs. Nevertheless, although encountering several linguistic dilemmas when implementing HUMOR to the Croatian language, one should bear in mind the benefits of such a morphological analyzer and the linguistic uses not only for parsing, but for the development of translational systems for Croatian and other minor languages as well.

References

Anić, V. 2000. *Rječnik hrvatskoga jezika*. Zagreb: Novi Liber

Babić et al. 1996. *Hrvatski pravopis*. Zagreb: Školska knjiga

Barić E. et. al. 1995. *Hrvatska Gramatika*. Zagreb: Školska Knjiga

Brodnjak, V. 1991. *Razlikovni rječnik srpskog i hrvatskog jezika*, Zagreb: Školske novine

Elekfi L. 1994. *Magyar ragozási szótár*. Budapest: MTA Nyelvtudományi Intézete

Papp F. 1969. *A magyar nyelv szóvégmutató szótára*. Budapest: Akadémiai Kiadó

Prószeky, G., Kis, B. 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland, USA: College Park. 261-268.

Raguž, D. 1997. *Praktična hrvatska gramatika*, Zagreb: Medicinska naklada

Težak, S. 1991. *Hrvatski naš svagda(š)nji*. Zagreb: Školske novine

Težak, S. 1995. *Hrvatski naš osebujni*. Zagreb: Školske novine

Težak, S. 1999. *Hrvatski naš (ne)zaboravljeni*. Zagreb: Tipex

Težak, S. 2004. *Hrvatski naš (ne)podobni*, Zagreb: Školske novine

Wardhaugh, R.1995: *Szociolingviszika*. Budapest: Osiris–Századvég

Semantic Description of Collocations in a Lexical Database

MARGARITA ALONSO RAMOS
Universidade da Coruña
Campus da Zapateira s/n, 15071 CORUÑA (SPAIN)
lxalonso@udc.es

Abstract

The aim of this paper is to draw attention to the need for the lexicographical description of the meaning of collocations. In existing English collocation dictionaries, collocations are semantically grouped, but no explicit semantic labels are available that can help the user to find the adequate collocate to express a given meaning. With regard to Spanish, existing general dictionaries fail to provide an accurate description of collocations. Here we briefly present the structure of the collocation database we are compiling, *Diccionario de colocaciones del español* (DiCE), following the guidelines of the Explanatory and Combinatorial Lexicology (Mel'čuk et al. 1995). Within this framework, collocations are described in terms of *lexical functions*. We have also included the translation of the lexical functions into a natural meta-language, which we call *gloss*. In this paper we focus on the notion of *gloss* or paraphrase that represents the meaning of collocate and on the didactical application of glosses. We intend to highlight the benefits of providing a semantic description of collocations in a lexical data base, which include the creation of an on-line language learning environment, integrating a dictionary and an exercise module.

1. Introduction

Most existing collocation dictionaries – either paper dictionaries or dictionary databases – suffer from two major shortcomings: (a) they fail to provide sufficient information, and (b) they do not offer selective access to the information given. By (a) we mean that they lack a semantic description of the collocations as well as sentential examples which illustrate their contextual use. By (b) we mean that they do not offer the option to access collocation information either via the base, via collocate, or via the type of collocation – thereby exploiting the flexibility offered by the electronic medium. In this paper, we present a lexical database of Spanish which takes both (a) and (b) into account.

In English collocation dictionaries such as BBI (Benson *et al.* 1986), LTP (Hill and Lewis 1997) and OCD (Crowther *et al.* 2002), collocations are semantically grouped. However, no explicit semantic labels are available that would help the user to find the adequate collocate to express a given meaning. Nor is the grouping sufficiently detailed to ensure that the members of one group are (quasi-) synonymous. Thus, in the OCD, under the headword *anger*, we find the following collocates in the same group: *be filled with, feel, shake with, tremble with.* However, a learner of English cannot possibly know whether these four collocates are quasi-synonyms or not: is there any difference in intensity between *be filled with* and *feel*? or in other words, can both collocates serve to express the same degree of anger? In the same way, a user may wonder whether the verbs that designate the symptom of an emotion can be considered equivalent to the verb which designates the experience of an emotion. From the example *His eyes were filled with anger*, the non-native user cannot infer that in English not only *eyes* can be *filled with anger*, but also the experiencer, that is, the person who feels the anger, such as in *My brother was filled with anger*. The meaning is not exactly the same: in the first case, the eyes are the means of the manifestation of the emotion and the collocation could be paraphrased as 'to manifest anger', whereas in the second case, the adequate paraphrase would be 'to feel intense anger'. As we see, existing collocation dictionaries have been designed to allow the user to infer the difference in meaning between collocates with the same base; but as shown above, they cannot succeed unless they explicitly specify the semantics, and unless they take advantage of the electronic medium, free from space limitations and which does not enforce a sequential search.

To date, general Spanish dictionaries have failed to provide an accurate description for collocations. At times, they are treated as idioms; at others, they appear between the examples of a given sense (Bargalló et al. 1997-1998). For this reason, we have undertaken the task of compiling a collocation database: the *Diccionario de colocaciones del español* (hereafter referred to as DiCE, Alonso Ramos 2001, 2002, 2003). For the time being, we have opted to focus on the field of emotion nouns. Unlike the English paper dictionaries mentioned above, DiCE has been conceived as an electronic lexical database. This allows us to provide more information and to implement a flexible means of access. A demonstration of DiCE is available at http://www.dicesp.com and http://www.colocacionesp.com.

Our framework is the Explanatory and Combinatorial Lexicology (ECL, Mel'čuk *et al.* 1995). In ECL, collocations are described in terms of *lexical functions* (LFs, Wanner 1996). An LF encodes the relation between two lexical units among which one of them (the *base* of the collocation) controls the lexical choice of the other one (the *collocate*). For instance, the LF `Magn` encodes the relation between the following adjective-noun pairs: *honda pena* 'intense pain', *terrible vergüenza* 'deep shame', and *ferviente admiración* 'great admiration'. Each of the three adjectives (= the collocates) is selected to express, in combination with the corresponding noun, the same meaning –'intense'. LFs have been used in the four volumes of the French *Dictionnaire explicatif et combinatoire* (Mel'čuk *et* *al.* 1984/1999), in the ongoing project *Lexique actif du français* (LAF, Polguère 2000), and now also for Spanish in DiCE.

As has already been pointed out by several authors, LFs are THE MEANS to describe collocations because they satisfy three indispensable requirements for a useful collocational resource: 1) they represent the meaning of the collocation; 2) they describe the syntax and the actantial structure of the collocation, and 3) they encode the functional dependency of the collocate in relation to the base. However, LFs are usually criticized for not being user-friendly (see, e.g., van der Wouden 1992): a user of an ECL dictionary is obliged to handle more 50 LFs and their combinations, which often hinders the effective usage of such dictionaries. As Polguère (2003) pointed out, the notion of LF has been hidden by a formal "wrapping", which is not necessarily the best one. Therefore, in LAF and in DiCE, we have opted to use natural language *glosses* to encode the meaning of the collocations. As LFs, glosses (or *paraphrases*) that represent the meaning of collocates are part of a meta-language. Thus, even if we describe a value of the LF `Magn` by the gloss `intenso` 'intense', the gloss is not a meaning of Spanish, but rather of "meta-Spanish". The gloss can be considered the translation of an LF in a natural meta-language.

2. Presentation of the DiCE in the web

DiCE is oriented towards language production. This implies that the primary access to the collocation information is via the base of the collocation. As Lea and Runcie (2002: 826) say, "you might be looking for the verb for what you do in response to a 'challenge'. But you would not choose 'meet' and then decide what to meet (a challenge, an acquaintance, your death, the expense)"[1]. Thus, if a user wants to know what other verb he can use to say *tener esperanzas* ('to have hope'), he will find in the entry for ESPERANZA, among others, the verb *abrigar* 'to cherish'. However, an access via the collocate side is also ensured. Thus, by clicking on a collocate, the inverse information search is launched: the user obtains the list of bases which co-occur with the collocate in question. In the case of *abrigar* 'to cherish', the user will be shown the other bases that share the collocate *abrigar*: *confianza* 'trust', *ilusión* 'hope', *miedo* 'fear', *rencor* 'grudge', *sospecha* 'suspicion', etc.

Our lexicographic unit is the lexical unit (LU) – rather than the word. For instance, for the noun *vergüenza* 'shame', several entries are available: for *shame I.1a* 'shame', *vergüenza I.2* 'shyness', *vergüenza I.3* 'decency'. For each LU, the entry provides the following information: a) the *semantic tag* that represents the

[1] That is precisely the procedure followed by the recently published combinatory dictionary of Spanish, named REDES and directed by Bosque (2004). Here headwords are collocates and the bases are ordered by lexical classes.

generic meaning of the LU in question; b) the *actantial structure* that represents the participants of the situation designated by the noun; c) corpus examples, most often from the Corpus of the *Real Academia Española* (CREA) available on the web; and d) the quasi-synonyms and the quasi-antonyms of the LU to help the user select the proper LU. Taking a specific LU as the starting point, the user can choose between five types of information search:

(1) **Attributes of the participants**. Under this heading, we have grouped those attributes or nouns that refer to the participants of the situation designated by the LU. For example, in the entry for ADMIRACIÓN 'admiration', the user finds *digno de admiración* ('worthy of admiration') or *admirable*, both referring to the participant that can compel admiration.

(2) **LU+ Adjective**. Here, the user finds adjectives that co-occur with the LU (either in an attributive or predicative position);

(3) **Verb +LU**: In this section, we have grouped the verbs that take the LU as a direct complement or as a prepositional complement; as, e.g., *despertar antipatía* '[to] arouse dislike' or *gozar de respeto* lit. '[to]enjoy of respect'.

(4) **LU + Verb**: This section contains verbs that take the LU as the grammatical subject; as, e.g. *el enfado se le pasó* 'his anger subsided'.

(5) **Noun *de* LU**: In this section, we include collocate nouns that precede the LU introduced by the preposition *de* 'of'; cf. *atisbo de esperanza* 'a glimmer of hope'.

Once the user has entered one of the above sections, he will find a list of collocates preceded by a gloss. Thus, if a user wishes to choose an adjective that combines with the LU ALEGRÍA 'joy', he must click on the heading (2). There, he will find a list of adjectives with the corresponding gloss. Among others, these are (the glosses are given in the Courier font; the corresponding adjectives in parentheses): `intensa` 'intense' (*desbordante*, *gran*, *impagable*, *indecible*), `más intensa de lo conveniente` 'more intense than is appropriate (*descontrolada*, *desmesurada*), `compartida por muchos` 'shared by many' (*generalizada*), `que no dura` 'which does not last' (*efímera*, *pasajera*), `causada por un buen motivo` 'caused by a good reason' (*sana*), `causada por el mal ajeno` 'caused by other people's evil'(*maligna*), and so on. Furthermore, users who are familiar with the notion of LFs can click on an LF-icon to see the LF that describes each collocation.

In order to give the reader an insight into the way collocations are described in DiCE, we have included an extract from the entry for VERGÜENZA 'shame'. First, we list the different LUs and secondly, we will focus on some collocations of the first UL. Here we offer the glosses in English and the LF as a subscript.

VERGÜENZA

I.1a. SENTIMIENTO 'feeling' [*la vergüenza de no haber tenido el valor de quitarnos la vida*]

I.1b. HECHO/ENTIDAD 'fact/entity' [*Eres la vergüenza de la familia; Es una vergüenza para nuestra familia que te hayas presentado como candidato para ese partido*]

I.2. SENTIMIENTO 'feeling' [*Me da vergüenza hablar ante tanta gente*]

I.3. CUALIDAD 'quality' [*Si tiene vergüenza, te devolverá ese dinero*]

II. plural only; colloq. PARTE DEL CUERPO 'body part' [*Tápate las vergüenzas*]

I.1a. (...)

VERGÜENZA+ADJ:

INTENSE $_{\text{Magn}}$ **terrible** me veo confesándome ante ti, con la terrible vergüenza de tener que admitir que mi jefe me gusta, **profunda** Por un momento sentí una profunda vergüenza de mí mismo, **enorme** siente una enorme vergüenza de pertenecer a la especie humana, **intensa** sentía una vergüenza intensa; **irreprimible** sus amigos lo saben todo. Una irreprimible vergüenza le abruma

VERBO + VERGÜENZA :

TO FEEL ~ $_{\text{Oper1}}$ **sentir** [~] los anuncios en la televisión, programas en los que uno siente vergüenza de ser mujer, **tener** [~] no por eso quiero que nunca piensen que tengan que tener vergüenza de su madre, **pasar** [(por) ART~]Podrían pillarte en una mentira y no querrás pasar por esa vergüenza, ¿verdad?; **sufrir** [ART~] Coco sufrió la vergüenza de saber que su nombre había sido deliberadamente omitido en la lista de invitados**, soportar** [ART~] la vergüenza que he tenido que soportar durante muchos años al verme señalada siempre como la hermana de la que se escapó ;

TO FEEL AN INTENSE ~ $_{\text{Magn+Oper1}}$ **morirse** [de ~] Me moriría de vergüenza; **caerse la cara** [de ~] Se me caería la cara de vergüenza;

TO CAUSE SOMEONE TO EXPERIENCE ~ $_{\text{CausFunc1}}$ **dar** [~ a X] Sí, claro que me dio vergüenza, **provocar** [~ en X] ese viejo síndrome de vergüenza que suele provocar en los hijos el modo de comportarse de los padres;

TO CAUSE SOMEONE TO EXPERIENCE INTENSE ~ $_{\text{Magn+CausOper1}}$ **llenar** [a X de ~] Estos chicos van a costarle a España bastante más que Lemóniz. Y, además, nos llenan de vergüenza

TO CAUSE HIMSELF TO CEASE EXPERIENCING ~ $_{\text{Liqu1Func0}}$ **sobreponerse** [a ART ~] El pobre no ha podido sobreponerse a la vergüenza y desde entonces vive recluido en esa especie de sanatorio

NOMBRE *de* **VERGÜENZA :**

LIGHT SIGN OF ~ $_{\text{Sing}}$ **asomo** [de ~] el derecho a un Estado laico con separación de poderes entre el civil y el religioso se incumple descaradamente y sin el menor asomo de vergüenza; **acceso** [de ~] he sentido, aparte de la inevitable tristeza, un acceso de vergüenza ajena; **ataque** [de ~] Ella sigue jugando al despiste mientras su hija Thais, presa de un ataque de vergüenza ajena, huye de la prensa al galope

As we have shown, even if LFs are not really visible in DiCE, they serve as the means for the description of collocational information. Therefore, we consider that DiCE is proof that the notion of LF can be employed without the technical peculiarities that dissuaded many users from using the ELC-dictionaries.

3. Considerations on the notion of gloss

The purpose of a gloss is to describe the contribution of the collocate to the global meaning of the collocation. By the term *gloss* we mean a brief indication of the meaning of the collocate in connection with the base. Thus, as we have seen, the gloss `intenso` serves to group various adjectives such as *terrible*, *profunda*, *enorme*, etc. which, in combination with VERGÜENZA 'shame' fulfil the same role, although they do not have strictly the same meaning.

Depending on our objectives, we can establish different degrees of granularity for the glosses. Sometimes, a vague `intense` can be enough to paraphrase the LF `Magn`, although at other times the linguistic glosses of different values of `Magn` can vary considerably. For example, although the three adjectives of the following collocations are represented by `Magn`: *honda pena* 'deep sorrow', *fumador empedernido* 'heavy smoker' and *sospecha vehemente* 'strong suspicion', their glosses are very different: `intense` for the first case; `[which smokes] a lot` for the second one and `such as X is very sure that Y is true` for the third one. Therefore, depending on what we wish to do with the glosses and depending on the nature of the collocate, we can be interested in a finer paraphrase or we can prefer a more approximate gloss such as `intense` for all collocate represented by the LF `Magn`.

The term *gloss* encompasses various notions, all of which are related to the semantic description of collocations. We can distinguish three different notions of gloss. The first consists of paraphrasing the meaning of the LF in natural meta-language. We will call this *LF formula paraphrasing*. Thus, for the first interpretation, a possible gloss for the complex LF `IncepPredMinus` can be `empezar a ser menos` 'to begin to be less', which is the translation of the LF in natural meta-language. A second notion of gloss is more closely related to the application of an LF. In that sense, the gloss could be the *default value* of the LF. For example, the default value for `IncepPredMinus` is `disminuir` 'to diminish'. However, in most cases, the default value of an LF depends on the semantic nature of the keyword. Thus, the gloss cannot be the same for *miedo atroz* 'terrible fear' and for *crimen atroz* 'terrible crime' even if both are encoded by the same LF. For the first collocation, the default value for `Magn` applied to the emotion nouns will be taken as the gloss (`intenso`); but for the second collocation, there is no a default value. In such cases, a third notion of gloss may be useful: a *short definition* of the collocate in the context of the collocation. For instance, *atroz* in *crimen atroz* is paraphrased by `que impresiona mucho` 'that causes a strong impact'.

All interpretations of the gloss have advantages and disadvantages. Thus, *LF formula paraphrasing* is closer to the semantic representation of the collocation and therefore, more suited to a possible system of automatic generation. Yet paraphrasing is less understandable for the human user, unlike the default value that is stylistically more elegant. The short definitions of collocate have the advantage of semantic precision but also the inconvenience of loss of generalization.

As stated above, the formulation of glosses must follow certain criteria in keeping with our objectives. If our aim is didactical and oriented towards students of Spanish as a foreign language, glosses have to be formulated using clear, albeit not very precise language. If our aim is scientific, glosses may be less elegant from a stylistic point of view, but considerably more precise. In any case, a gloss cannot violate the language. Even if the gloss is considered as a meta-language, it is always formulated in meta-Spanish or meta-English or any meta-language[2].

Formulating a gloss is no easy task. However, we wish to draw attention to the need for the lexicographical description of the meaning of collocations. In the following section, we will highlight the problems facing a student of English as a foreign language when attempting to answer the exercises included in the OCD.

4. Didactical Applications of Glosses

The authors of collocation dictionaries are aware of the importance of collocations in the learning/teaching of a foreign language. One of the authors of the LTP has published a book whose title shows clearly the relevance of this subject, *Teaching collocations* (Lewis 2000). Another English collocation dictionary such as the OCD comes complete with a wide range of exercises designed to put the information included in the dictionary to practical use. However, learners of English will be unable to do these exercises without a monolingual English dictionary as they require the ability to make fine distinctions between the meaning of the collocates. For instance, in a gap-fill exercise, which provides the initial letter for the collocate, we are asked to supply the quantifier for the noun *depression*. The compiler offers "frequent b.... of depression". When

[2] For instance, we do not accept `ejecutar` 'to execute' as a gloss of the Spanish collocation *dar un golpe* 'to give a blow', because **ejecutar un golpe* is ungrammatical in Spanish.

answering, the learner has to look up the entry *depression* and under the heading "quantifiers", he will find *bout, fit, period*. But, how can this information be interpreted? Can both *fit* and *period* also be used? A collocation dictionary without semantic description cannot help the learner to answer this question. Similarly, another common exercise is based on completing a story. In this case the learner has to guess the required meaning, which is not always obvious. In a short story about a party, we find the following text: "The **wine** had flowed freely [...] and now my **head** was t... and my **stomach** was c... ".The learner is supposed to look up the entries for *head* and for *stomach* to find the good collocates. However, using the text the learner is not provided with sufficient indications as to how to find the answer, as there are various possible collocates from different parts of speech. Thus, if the learner looks for an adjective beginning with "t" which combines with *head*, he will be unable to find it because the correct answer is a verb, *to throb*; the same is true in the case of *stomach*, where the correct collocate is *churning*.

Unlike these dictionaries, DiCE provides the semantic description of collocations, which allows for the creation of an on-line language learning environment, integrating both a dictionary and exercise module (see Selva *et al.* 2002): the user is asked to choose or to find the correct collocate starting from a gloss of the meaning. We will now offer a brief presentation of the DiCE exercise module. We have divided it into two sections: one for production and another for comprehension. In both sections, we have included several types of exercises: some consist of choosing the correct answer and others of filling in the correct answer. All are automatically evaluated by the system. The sub-module includes various exercises devised to identify the correct collocate using their semantic description. For instance:

Si Juan empieza a tener ganas de ir al cine, entonces a Juan... [If Juan begins to feel like going to the cinema, then Juan...]

1. *le aparecen las ganas*
2. *le surgen las ganas*
3. *le entran ganas*
4. *le duran las ganas*

In the statement in this exercise, the gloss of the required collocate is provided: 'empezar a tener ganas' ('to begin to feel like') is supposedly the paraphrase for the collocate *entrar ganas*. Another type of exercise consists of filling in the correct collocates. For example:

Busca un verbo que exprese el sentido 'empezar a sentir' en la secuencia *Te ha ... cariño y sólo aspira a ser tu amigo* [Look for a verb which expresses the meaning 'to begin to feel' in the sequence *He has...affection for you and he only wants to be your friend.*

Escriba la respuesta: Write the answer here
¿Has acertado? Were you right?

In the comprehension sub-module, the exercises are also based on semantic description. In this case, the learner has to choose the appropriate gloss for a collocate. For instance:

En la secuencia *Un niño despierta ternura*, di cuál es el significado de *despertar* [In the sequence *A child arouses tenderness*, identify the meaning of *arouse*]:

1. sentir
2. causar
3. manifestar
4. tener

All the exercises must be classified according to learner level, but we believe that much of this didactic material can be taken from the DiCE.

5. Conclusions and Future Work

We have attempted to show the advantages of a semantic description of collocations in a database in terms of LFs with their translation in glosses. First, LFs offer a grid that helps the lexicographer to detect collocations in the corpus. Secondly, given that LFs constitute a formal language, they systematise the collocational information. Thirdly, structured information makes it easier to integrate the dictionary and exercise module into an on-line language learning environment. Future work will address several tasks: From a lexicographical approach, we must look more closely into the possibilities of generalizing glosses in accordance with the meaning of bases, whilst from a didactical perspective, we plan to study the possibility of generating and automatically correcting exercises based on the data included in the dictionary.

Acknowledgments

I would like to take advantage of this opportunity to thank Leo Wanner for reading the draft version of this text and for his encouraging comments. My thanks also go to Igor Mel'čuk for his consistent habit of covering the texts in red, which I trust goes a long way to improving the end results. The English version of this text was proofread by Sarah Moss. This paper was written within the framework of the following research projects: BFF2002-04226-C03-01 (MCyT and partially funded by FEDER), PGDIT02PXIB30501PR and PGIDITPXIC101401PN (Xunta de Galicia).

References

Alonso Ramos, M. (2001), "Construction d'une base de données des collocations bilingue français-espagnol". *Langages*, 143, p. 5-27.

____ (2002), "Un vacío en la enseñanza del léxico del español como lengua extranjera", in A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, volume II*, p. 551-561, Copenhague, CST.

____ (2003), "La nature des collocatifs: leur statut en tant qu'unités lexicales", in F. Grossmann et A. Tutin (eds.), *Les collocations : analyse et traitement, Travaux et Recherches en Linguistique appliquée*, Amsterdam: Editions De Werelt, p. 45-60.

Bargalló, M. et al. (1997-1998), "El tratamiento de los elementos lexicalizados en la lexicografía española monolingüe", *Revista de lexicografía*, 4, p. 49-65.

Benson, M., E. Benson and R. Ilson (1986), *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam/ Philadelphia: John Benjamins.

Bosque, I. (dir.) (2004), *Redes. Diccionario combinatorio del español contemporáneo*, Madrid: SM.

Crowther, J., S. Dignen and D. Lea (eds.) (2002), *Oxford Collocations Dictionary for Students of English.* Oxford: Oxford University Press.

Hill, J. and M. Lewis (eds.) (1997), *LTP Dictionary of Selected Collocations*. London: LTP.

Lea, D and M. Runcie (2002), "Blunt Instruments and Fine Distinctions: a Collocations Dictionary for Students of English", in A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, volume II*, p. 819-829, Copenhague, CST

Lewis, M. (ed.) (2000), *Teaching Collocation. Further Developments in the Lexical Approach*. London: LTP.

Mel'čuk, I. et al. (1984-1999), *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques* I-IV, Les Presses de l'Université de Montréal: Montréal.

Mel'čuk, I., A. Clas and A. Polguère (1995), *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Polguère, A. (2000), "Towards a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French", *Proceedings of the Ninth EURALEX International Congress, volume II*. Stuttgart: Universität Stuttgart, p. 517-527.

____ (2003), "Collocations et fonctions lexicales: pour un modèle d'apprentissage", in F. Grossmann et A. Tutin (eds.), *Les collocations : analyse et traitement, Travaux et Recherches en Linguistique appliquée*, Amsterdam, Editions De Werelt, p. 117-133.

Selva, T., S. Verlinde and J. Binon (2002), "Le DAFLES, un nouveau dictionnaire électronique pour apprenants du français", in A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, volume I*, p. 199-208, Copenhague, CST.

van der Wouden, Ton (1992), "Prolegomena for a multilingual description of collocations", in *EURALEX '92 Proceedings I-II*, University of Tampere, p. 449-456.

Wanner, L. (ed.). 1996. *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/ Philadelphia: John Benjamins.

Valency information for dictionaries and NLP lexicons: Adapting valency frames from The Danish Dictionary to an LFG lexicon

JØRG ASMUSSEN
Dept. for Digital Dictionaries and Text Corpora
Society for Danish Language and Literature, DSL
Copenhagen, Denmark
ja@dsl.dk

BJARNE ØRSNES
Dept. of Computational Linguistics
Copenhagen Business School, CBS
Copenhagen, Denmark
boe.id@cbs.dk

Abstract

This paper investigates how the valency information of The Danish Dictionary – a dictionary intended for human users – can be transformed to a more generalized notation which can serve as an NLP resource, e.g. as input for an LFG lexicon. Based on the requirements for a generalized representation of syntactic information, this paper also proposes a coding scheme for valency codings in dictionaries that can easily be translated into other formats, including the user-friendly, 'human' notation given in the printed version of The Danish Dictionary.

1 Introduction

The Danish Dictionary (DDO: Hjorth et al. (2003), cf. Lorentzen (2004)) is a corpus-based dictionary of modern Danish published 2003-2005 in six volumes by the Society for Danish Language and Literature, DSL. Even though the DDO has been edited according to explicit rules as a highly structured document and consequently appears quite consistent, it is conceptually designed as a printed dictionary for humans. The general editorial strategy is thus to avoid redundant or over-explicit information in favour of condensed and implicit information wherever possible.

This concept often runs counter to the desire to use the dictionary in a digital context with more elaborate query facilities, cf. Asmussen (2004), or as a general lexical resource for NLP purposes. In order to meet these sometimes conflicting demands on printed vs. digital dictionaries and dictionaries vs. NLP lexicons, DSL's new web-based *Ordnet.dk* project is currently developing a more general dictionary design that can serve all these demands.

This paper shows how the valency information given in the DDO can be transformed to a more generalized notation which easily can be converted to e.g. an LFG lexicon. It also proposes a coding scheme for future valency codings in *Ordnet.dk*. This scheme has the advantage that it can readily be translated into other formats, including the more user-friendly notation in DDO.

2 Valency information in the DDO

The DDO is the first dictionary of Danish offering semi-formal valency frames for verbs. Surprisingly, this feature and its potential for spin-off products such as NLP lexicons has never been conveyed to a broader community of lexicographers, with the exception of Nimb (1996).

The following examples illustrate the basic structure of the valency frames in the DDO:

```
(1)  [NGN/NGT specificerer NGT]          : [SBD/STH specifies STH]
(2)  [NGN spadserer (+STED/+RETNING)]    : [SBD walks (+PLACE/+DIRECTION)]
(3)  [NGN teoretiserer (over NGT)]       : [SBD theoretisizes (over STH)]
(4a) [NGN barberer sig/NGN/NGT]          : [SBD shaves oneself/SBD/STH]
(4b) [NGN/NGT barberer HÅR af/væk/bort]  : [SBD/STH shaves HAIR off/away/away]
(4c) [NGN barberer NGT ned/væk/bort]     : [SBD shaves STH down/away/away]
(5a) [NGN diskuterer (NGT) (med NGN)]    : [SBD discusses (STH) (with SBD)]
(5b) [NGL diskuterer (NGT) med hinanden] : [SBD_PLUR discuss (STH) with each_other]
(5c) [NGN diskuterer om+SÆTN/hv+SÆTN]    : [SBD discusses if+CLAUSE/wh+CLAUSE]
(6)  [NGN planlægger (NGT/at..)          : [SBD plans (STH/that+CLAUSE/to+INF)]
```

Capitalized strings refer to complements (including the subject) (ex. 1–6), non-capitalized strings are the verb proper (ex. 1–6) and other literal words, e.g. particles, prepositions heading valency-bound PPs (ex. 3, 5a–5b), reflexive pronouns (ex. 4a), or a somehow fixed or restricted vocabulary (ex. 5b). NGN, NGL, and NGT principally represent NP arguments, adverbials are prefixed by + (ex. 2), whereas x+SÆTN (ex. 5c) refers to a subordinate clause; at.. (ex. 6) denotes a constituent containing the complementizer *at*, 'that', i.e. a clausal complement, or a constituent containing the infinitive marker *at*, 'to', i.e. an infinitival complement. NGN and NGT contain the semantic restrictions '+human' and '–human', NGL both gives '+human' and the morphosyntactic restriction '+plural'. Other capitalized strings are complements (ex. 4b), or, if prefixed by a +, adverbials (ex. 2), restricted to a certain semantic field. Hence, HÅR denotes an NP restricted to the semantic field HAIR and +STED an adverbial semantically restricted to PLACE. Brackets enclose optional complements (ex. 2–3, 5a–5b, 6), slashes indicate alternation between constituents or literal words (ex. 1–2, 4a–4c, 5c, 6). The order of the alternating elements is principally determined by corpus frequency. As the notation reflects canonical constituent order (basically S-V-O) followed by more oblique functions, it is in principle quite straightforward to derive the corresponding grammatical functions without having them explicitly represented.

3 Requirements for a generalized representation of syntactic information

The DDO representational scheme conflates different dimensions of information into a single compact representation intended for human users. In spite of the fixed format of the representation, it is not be optimal for NLP purposes:

- NLP systems represent valency information in different ways according to the underlying linguistic theory. LFG represents valency in terms of syntactic *functions*, HPSG in terms of syntactic *categories* on valency lists, etc. The DDO frames do not separate categories and functions, but by explicitly giving both kinds of information the representation can fulfil both needs.
- In the DDO representation syntactic function is read off the relative position of the item in the representation. This is straightforward, but cumbersome for use in NLP applications.
- The conflation of syntactic, semantic, and grammatical information leads to ambiguous representations of e.g. NP-complements since NPs may be either `NGN`, `NGT` og `NGL`. Any reference to an NP-complement thus involves a three-way disjunctive statement.
- Valency bound PPs are represented by the preposition in question followed by the category of the regimen. Thus, a PP has to be identified by the occurrence of a preposition in the frame. This is not adequate for NLP applications given that it requires access to a list of the possible prepositions.

Thus a generalized notation must provide for an explicit representation of all the information present in the DDO frames so that the individual pieces of information are readily available. This approach will facilitate searches based on different kinds of lexical properties, and it will make it possible for NLP applications to extract the information in a format anticipated by these applications.

In the present conception of the generalized notation, alternation and optionality does not occur. This means that all frames involving alternations and optional complements are expanded into individual frames. Thus, a frame such as `[NGN <verb> NGN/NGT (+ADVL)]` gives rise to four individual frames. Nothing hinges on this strategy, but it is has been adopted primarily for practical reasons. The current experiment is tested on an LFG grammar for Danish implemented in the Xerox Linguistic Environment (XLE) where alternations and optionality are given as disjunctive valency frames. Possibly this kind of representation may also be useful for more systematic investigations of co-occurring valency frames in the spirit of Levin (1993).

As the generalized notation aims at a “multi-purpose” representation of valency information, we adopt a fairly traditional inventory of syntactic functions and categories. The representation distinguishes: subject (SUBJ), object (OBJ), oblique (OBL): typically subcategorized prepositional phrases, indirect object (IOBJ), particle (PART), and ADJUNCT for subcategorized adjuncts. Predicative complements are represented by the function XCOMP (inspired by LFG).

However, we also employ a function for parts of a complex predicate (COMPPRED). In some cases the DDO frames contain strings of individual words representing collocationally restricted complements such as

```
[NGN sætter NGN/NGT i relation til NGT] : [SBD puts SBD/STH into relation to STH]
[NGN skyder NGT i sænk]                 : [SBD shoots STH down]
```

In these cases the strings *i relation til* and *i sænk* represent PPs predicated of the object, thus suggesting that they could be treated as parts of complex predicates. This means that the verbs are equipped with a specific function COMPPRED whose values are the lexical items in question. Alternatively these strings could be represented as oblique PPs restricted to specific lexical items by a separate feature, cf. below. However, this would require for us to parse these strings of lexical items in order to determine their syntactic category. The adequacy of representing these strings as parts of a complex predicate remains to be tested.

The inventory of syntactic categories includes the traditional categories S, NP, VPinf, PP, etc. instead of e.g. CP, IP, or DP. We believe that the traditional categories can be mapped to the categories foreseen by the adopted linguistic framework. Valency-bound PPs are represented as PPs, but in addition, the category of the regimen is provided as it may be idiosyncratically selected by the verb: PP-NP for a PP with an NP regimen.

The other information types are illustrated below.

4 Valency information made explicit

As can been seen from the examples in section 2, the valency frames in DDO are a condensed representation of the necessary morpho-syntactic and to some extent semantic information pertaining to constructional information in the printed dictionary. In section 3 it has been shown that this condensed notational style can be difficult to use directly for NLP purposes, hence our approach to establish a generalized notation which (1) does not allow *alternation* nor *optionality* within a frame, and which (2) clearly distinguishes different types of linguistic content from the valency information of the DDO. Our notation thus gives explicit information on

a. syntactic *function*: subject, object, etc. (mandatory)
b. restrictions on syntactic *category*: NP, PP, ADVP, etc.
c. *morphology*, i.e. morpho-syntactic restrictions: plural complements, reflexive pronoun, etc.
d. *selectional restrictions*: human, non-human, place, direction, etc.
e. restrictions on specific *lexical items*, e.g. prepositions heading prepositional objects, or collocationally restricted vocabulary: verbal particles, adverbs, parts of complex predicates, etc.

In addition, our notation also could give information on

- frequency rank where two or more constituent types alternate, as the order of alternating elements given in the DDO valency notation aims to reflect the frequency observed in the DDO Corpus[1]
- canonical constituent order, as the constituent order given in the DDO notation reflects this order

The frequency information is left out because the corpus size of 40 million tokens is too small to draw reliable quantitative conclusions on syntax, cf. Asmussen (2005); canonical constituent order is implicitly reflected in our notation – however, constituent order should be treated as part of the grammar, not as part of the lexicon.

Let us look at ex. 6 from section 2 to see how the conversion from the DDO valency notation into our notation works. The first step is to make explicit the syntactic information given by `at..` by expanding it to `at+SÆTN/at+INF` (`that+CLAUSE/to+INF`). Secondly, *alternation* and *optionality* are removed which results in the following four frames:

```
(6.1) NGN planlægger
(6.2) NGN planlægger NGT
(6.3) NGN planlægger at+SÆTN
(6.4) NGN planlægger at+INF
```

Hence, in our notation, the entry for *planlægge* can be represented as four matrices (6.1–6.4):

[1] A detailed description of the design of the DDO Corpus can be found in Norling-Christensen and Asmussen 1998.

		6.1	6.2		6.3		6.4	
a.	function	SUBJ	SUBJ	OBJ	SUBJ	OBJ	SUBJ	OBJ
b.	category	NP	NP	NP	NP	S	NP	VPinf
c.	morphology	-	-	-	-	at	-	at
d.	sel. restrict.	+hum	+hum	-hum	+hum	-	+hum	-
e.	lexical item	-	-	-	-	-	-	-

The matrix notation is mapped into an attribute-value pair notation where irrelevant/underspecified information, marked with dashes in the matrices, is left out.

```
(6.1) SUBJ(cat='NP', sel='+hum');
(6.2) SUBJ(cat='NP', sel='+hum'); OBJ(cat='NP', sel='-hum')
(6.3) SUBJ(cat='NP', sel='+hum'); OBJ(cat='S', mor='at')
(6.4) SUBJ(cat='NP', sel='+hum'); OBJ(cat='VPinf', mor='at')
```

This notation, partly inspired by the TSNLP-based IMSLex syntax representation (cf. Lezius et al. (2000)), is our generalized representation mediating between dictionaries for humans and NLP lexicons. The conversion process from the condensed DDO notation into the generalized notation is done by a Perl script applying simple string substitutions based on regular expressions. A more elaborate approach would have been to design a context-free grammar-based parser for the DDO notation, but as we are aiming to replace the DDO notation by one close to our generalized representation anyway, the conversion task only has to be run once, therefore our simple (but unflexible) approach seemed to be the cheapest one.

5 An LFG lexicon for Danish

LFG (cf. Kaplan and Bresnan (1982), Bresnan (2001)) is a non-derivational constraint-based theory based on an architecture of parallel correspondences between different dimensions of linguistic representation. Generally LFG posits two dimensions of syntactic representation: a c(onsituent)-structure depicting linear precedence as well as hierarchical relationships among the constituents, and a f (unctional)-structure representing the basic syntactic relations between the constituents. Additional projections such as a semantic projection may be defined. The mappings are given as functional annotations in lexical entries and on context-free phrase-structure rules. LFG is the framework of the international Pargram-project aiming to produce parallel LFG-grammars for a number of languages on the XLE platform. Our experiment is tested on the Danish grammar, cf. Ørsnes and Wedekind (2003).

In LFG the valency of predicators is given as a list of syntactic functions required in the f-structure of the predicator. However, it is also possible to impose categorial constraints on these functions by defining restrictions on the c-structure nodes in the inverse correspondence of a certain piece of f-structure. In LFG, valence alternations such as passivization are treated by lexical rules, and in addition it is possible to generalize over lexical descriptions by means of templates which may in turn call other templates, thus giving the effect of a lexical inheritance network. Here we will confine ourselves to a presentation of the lexical entries resulting from an automatic conversion of the generalized representation presented in section 4.

6 Converting the generalized representation into LFG

The following examples are all based on valency frames for the verb *købe* ('to buy').

The DDO frame [NGN køber NGT] : [SBD buys STH] is mapped to the generalized representation SUBJ(cat='NP', sel='+hum'); OBJ (sel= '-hum'). This frame in turn maps to the LFG-representation below:

```
(^PRED)='%stem<(^SUBJ)(^OBJ)>'
(s::(^SUBJ) SEL) = +HUM
(s::(^OBJ) SEL) = -HUM.
```

The PRED value is given by the stem-value of the verb. The lexical form lists a SUBJ and an OBJ that must obligatorily occur in a well-formed f-structure with the verb as the main predicator. Selectional restrictions are treated as belonging to a separate s-structure projected off the f-structure. The equation reads: the semantic structure of the value of the SUBJ attribute contains the specification +HUM in its selection path (SEL). Selectional restrictions are not currently used in the Danish grammar but can be used to filter out unwanted or less likely readings. A more complex example involving prepositional complements and optionality is given below.

In the DDO-frame [NGN køber NGN (til NGT)] : [SBD buys SBD (for STH)], the optionality of the oblique complement resolves into two generalized representations:

1. SUBJ(cat='NP', sel='+hum'); OBJ(cat='NP', sel='+hum'); OBL(cat='PP-NP', sel='-hum', lex='til')
2. SUBJ(cat='NP', sel='+hum'); OBJ(cat='NP', sel='+hum')

The second frame was already dealt with above while the first frame maps into the following LFG-representation:

```
(^PRED)='%stem<(^SUBJ)(^OBJ)(^OBL-til))>'
@CAT((^OBL-til){NP})
(s::(^SUBJ) SEL) = +HUM
```

```
(s::(^OBL-til) SEL) = -HUM
(s::(^OBJ) SEL) = +HUM.
```

The OBL function is individuated by the stem-value of the preposition heading the PP. By using the preposition as part of the name (instead of a semantic characterization such as OBL-theme) we can dispense with additional constraints on the actual PFORM value of the OBL functions. Note that the oblique prepositional phrase must contain an NP in the c-structure, due to the CAT specification.

The frame [NGN køber NGT/NGN (af NGN) (for NGT)] : [SBD buys STH/SBD (from SBD) (for STH)] involves one alternation and two optional constituents, i.e. the object may be +hum or -hum and both obliques are optional. This gives rise to eight frames in the generalized representation in total and consequently to eight separate LFG-entries – of course, substructures are shared among these entries thus introducing some redundancy as mentioned above:

1. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’-hum’)
2. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’+hum’)
3. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’-hum’); OBL(cat=’PP-NP’, sel=’-hum’, lex=’for’)
4. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’+hum’); OBL(cat=’PP-NP’, sel=’-hum’, lex=’for’)
5. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’-hum’); OBL(cat=’PP-NP’, sel=’+hum’, lex=’af’); OBL(cat=’PP-NP’, sel=’-hum’ lex=’for’)
6. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’+hum’); OBL(cat=’PP-NP’, sel=’+hum’, lex=’af’); OBL(cat=’PP-NP’, sel=’-hum’ lex=’for’)
7. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’-hum’); OBL(cat=’PP-NP’, sel=’+hum’, lex=’af’)
8. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, sel=’+hum’); OBL(cat=’PP-NP’, sel=’+hum’, lex=’af’)

The DDO frame [NGN køber NGN/sig +ADJ/+ADVL] : [SBD buys SBD/ oneself +ADJ/+ADVL] is an example of a frame that must be simplified in the corresponding LFG-representation. The frame represents the extremely common resultative construction where a complement describing the result of the action denoted by the verb is predicated of the object. In LFG, the predicative complemented is treated as the open function XCOMP where the unexpressed subject is identified with object of the matrix verb by a functional control equation. XCOMP is the function of both ADJ and ADVL in the DDO-frame, and the reflexive object covers the cases where the controller is identical to the matrix subject. Thus the four generalized representations:

1. SUBJ(cat=’NP’, sel=’+hum’); OBJ(cat=’NP’, mor=’refl’); XCOMP(cat=’AP’)

2. `SUBJ(cat='NP', sel='+hum'); OBJ(cat='NP', mor='refl'); XCOMP()`
3. `SUBJ(cat='NP', sel='+hum'); OBJ(cat='NP', sel='+hum'); XCOMP(cat='AP')`
4. `SUBJ(cat='NP', sel='+hum'); OBJ(cat='NP', sel='+hum'); XCOMP()`

are covered by the following LFG-entry:

```
(^PRED)='%stem<(^SUBJ)(^OBJ)(^XCOMP))>'
(^OBJ)=(^XCOMP SUBJ)
(s::(^SUBJ) SEL) = +HUM
(s::(^OBJ) SEL) = +HUM.
```

7 Translating the frames into LFG

Basically, the frames of the generalized representation can be linked to the LFG-grammar in two ways. Either each frame is associated with a name of a tempate defined in the LFG-grammar, or the generalized represention is translated into LFG piece-wise. The first approach has the disadvantage of hampering updating. If the dictionary is updated with new frames, or if information in the generalized representation is changed, the templates have to be revised as well. For that reason we rely on a piece-wise translation of the information in the frames even though it may introduce redundancy in the lexical entries as we will see.

The automatic conversion of the frames of the generalized representation into LFG involves inserting the variable information into predefined functional equations of LFG.

For example, in the frame `[SUBJ(cat='NP', mor='plur'); OBJ (cat= 'NP', sel='-hum')]` for the verb *afregne* ('to settle one's account'), the function names `SUBJ` and `OBJ` are collected and inserted into the skeletal LFG equation `(^PRED) = '%stem<list_of_function_names>'`.

The information of category selection may be inserted into the predicate `@CAT ((<function>){<set_of_categories>})`. The predicate `CAT` picks out the c-structure nodes corresponding to a certain piece of f-structure. Thus the predicate `@CAT((^SUBJ){NP})` succeeds if the set of c-structure nodes mapping to `SUBJ` contains an `NP`. However, we adopt the general strategy of LFG to avoid specifications of category unless it is completely idiosyncratic. Currently only the regimen of an oblique phrase is given a category specification.

Information on selectional restrictions is inserted in the frame `(s:: (^<function>) SEL) = <value>`. The prefix `s::` identifies the separate semantic projection of the function, e.g. subject or object. This projection contains the attribute `SEL` with the specified selectional restriction (basically `+HUM` or `-HUM`).

As it stands, morpho-syntactic information includes information on number, complementizer type of embedded clauses, and pronominal subcategories. Specification on number is inserted into the frame `(^<function> NUM) =c <value>`. A functional equation arising from the schema could be `(^SUBJ NUM) =c PLUR`. Here we use a constraining equation requiring the presence of the specified attribute with the specified value.

If the frame contains an `XCOMP` the functional controller of the unexpressed subject of the `XCOMP` has to be identified. Here we rely on LFG's lexical rule of functional control: if an object is present in the frame, it is identified as the controller, otherwise the subject is represented as the controller.

8 Improvements of the DDO representational scheme

Apart from facilitating use in NLP application based on different linguistic assumptions, the explicit differentiation of function and category has some conceptual advantages.

As it stands, the DDO-representation seems to give rise to some coding inconsistencies. The term `NGT` semantically covers both non-animate NPs and propositional complements. Some coders make `NGT` cover propositional complements, while others indicate propositional complements with category `at+SÆTN/that+CLAUSE`. Such inconsistencies may be avoided if the coders are urged to provide information about function as well as category.

In the existing frames, `PP` ranges over a host of very different syntactic patterns. Thus, the DDO frame `[NGN <verb> NGN til NGT]` covers both the resultative construction where the `PP` is predicated of the object, as in *de udnævnte ham til formand* ('they announced him chairman'), and constructions with place adverbials such as *de sendte ham til London* ('they sent him to London'). This difference can be captured by assuming that the PP has different syntactic functions (`XCOMP` and `OBL` respectively) while the syntactic category remains the same.

The DDO is currently represented as an SGML document where the valency information is given as literal strings (`#PCDATA`) within a `<valency>` element. The structure within this element is exclusively determined by the editorial guidelines which describe the condensed DDO valency notation, not by the DTD proper. It has been shown that the condensed and relatively free style applied in the DDO blurs a clear distinction between the different types of linguistic information involved in the notation with some inconsistencies in the valency description as a severe disadvantage of this approach. In order to make it more clear to the editors which information is obligatory (syntactic function and syntactic category) and what can be left out (e.g. over-explicit or too vague semantic information) we have designed an SGML/XML structure based on our intermediate representation augmented by explicit information on canoncial constituent order, determined by the order and notation of constituent alternation and optionality, cf. the DTD below.

```
<! ELEMENT syntax (subj, iobj? , obj? , comppred? , part? , obl*, xcomp? , adjunct? )>
<! ELEMENT subj (attribs+)>
<! ELEMENT iobj (attribs+)>
<! ELEMENT obj (attribs+)>
<! ELEMENT comppred (attribs+)>
<! ELEMENT part (attribs+)>
<! ELEMENT obl (attribs+)>
<! ELEMENT xcomp (attribs+)>
<! ELEMENT attribs (cat, mor? , sem? , lex? )>
<! ELEMENT cat (#PCDATA)>
<! ELEMENT mor (#PCDATA)>
<! ELEMENT sel (#PCDATA)>
<! ELEMENT lex (#PCDATA)>
```

This DTD reflects the structure of our proposed generalized notation for syntactic information on verbs much better than the original one used in the DDO. Therefore we propose this structure for future editorial work on the DDO.

9 Conclusion and future work

We have investigated the prospects of using a dictionary intended for human users as a resource for computational applications. Specifically, we have proposed a generalized representation of the valency information where the separate pieces of information are automatically factored out and given an unambiguous representation without loss of information which opens up for adding more information if necessary. We have shown how the DDO valency notation can be converted into the generalized representation and how this format can be converted into an LFG-based computational lexicon for Danish exploiting the parallel architecture of LFG. Finally, we gave some suggestions on improving the syntactic information given in the DDO by explicitly keeping functional and categorial information apart, and we proposed an enhanced DTD for future valency notation in the DDO.

Future work will include a conversion from the generalized notation to the enhanced DTD and an evaluation of the generalized notational approach by the editors of the DDO. Future work should also deal with the possibility of identifying verb classes on the background of the generalized representation, i.e. classes of verbs sharing some salient valency alternations such as dative-shift, and should deal with the identification of valency alternations that may best be treated as lexical rules in LFG.

References

Asmussen 2004 Asmussen, J. (2004). Feature Detection - A Tool for

Unifying Dictionary Definitions. In *Proceedings of the 11th EURALEX International Congress*, volume 1, pages 63–69, Lorient. Euralex.

Asmussen 2005 Asmussen, J. (2005). Towards a methodology for corpus-based studies of linguistic change. Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 Corpora of Danish. In Archer, D., Rayson, P., and Wilson, editors, *Corpus Linguistics Around the World*. Rodopi, Amsterdam.

Bresnan 2001 Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell, Oxford.

DDO: Hjorth et al. 2003 DDO: Hjorth, E., Kristensen, K., Lorentzen, H., Trap-Jensen, L., Asmussen, J., et al., editors (2003). *Den Danske Ordbog 1-6*. DSL & Gyldendal, København/Copenhagen.

Kaplan and Bresnan 1982 Kaplan, R. M. and Bresnan, J. (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. The MIT Press, Cambridge, MA.

Levin 1993 Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Lezius et al. 2000 Lezius, W., Dipper, S., and Fitschen, A. (2000). IMSLex – Representing Morphological and Syntactical Information in a Relational Database. In *Proceedings of the 9th EURALEX International Congress*, pages 133–139, Stuttgart. Euralex.

Lorentzen 2004 Lorentzen, H. (2004). The Danish Dictionary at large: presentation, problems and perspectives. In *Proceedings of the 11th EURALEX International Congress*, volume 1, pages 285–294, Lorient. Euralex.

Nimb 1996 Nimb, S. (1996). Collocations of Nouns: How to Present Verb-noun Collocations in a Monolingual Dictionary. In Gellerstam, M., Järborg, J., Malmgren, S.-G., Norén, K., Rogström, L., and Papmehl, C. R., editors, *Proceedings of the 7th EURALEX International Congress*, Göteborg. Euralex.

Norling-Christensen and Asmussen 1998 Norling-Christensen, O. and Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8:223–242.

Ørsnes and Wedekind 2003 Ørsnes, B. and Wedekind, J. (2003). Parallelle datamatiske grammatikker for norsk og dansk. In Holmboe, H., editor, *Nordisk Sprogteknologi 2002*. Museum Tusculanums Forlag, København/Copenhagen.

What is *compatible* with what? Or, Reducing the collocational chaos in the predicate-argument structure, with a little help from metonymy

Mario Brdar
University of Osijek
L. Jägera 9, HR-31000 Osijek
mbrdar@ffos.hr

The present paper argues that metonymy as a basic cognitive process provides us with an extremely useful tool in keeping a balance between the richness of data input, on the one hand, and the optimal degree of generality in their interpretation and presentation, on the other. It is argued that introducing metonymy (as it is understood in cognitive linguistics) into the conceptual apparatus of lexico-grammatical enterprises, in this specific case, into the account of the predicate-argument structure, as well as into the whole lexicon structure, may result in making lexicographic handbooks more functional and streamlined, individual entries becoming more compact. In fact, it is argued that it in fact may result in an increase of the generative power of lexicographic handbooks in the sense of enabling their users to make relatively safe guesses about the acceptability of some novel grammatical combinations not actually recorded in the lexicon or corpora.

A notorious pitfall for accounts of predicate-argument structure that strive for comprehensiveness is one particular facet of its semantic aspects, viz. the account of regularities in—or restrictions on—the semantic combinability of predicates and their arguments. Two extreme pathways have been trodden in dealing with this issue. The more theoretical (and the more fragmentary) an account of the predicate-argument structure, the more likely it is to exhibit preferences for capturing the regularities in the semantics of combinability of predicates and arguments by means of systematic selectional restrictions, which ultimately turn out to grossly overplay regularity while still remaining rather vague. The more applied and corpus data-driven an approach, the more likely it becomes that it will satisfy itself with mere listings of collocates functioning as arguments, i.e. offer more or less uninterpreted or raw data, which ultimately gives the impression that chaos reigns supreme as far as an account of the referential semantics of arguments is concerned.

The present paper attempts to show that it is possible to pick the best from both worlds, i.e. it is possible to combine the advantage of using huge amounts of naturally occurring data on the one hand, while not giving up the aspiration to provide a principled account of all the regularities, on the other. In other words, it is argued that it is possible to significantly reduce the apparent chaos that can be observed when it comes to a description of the collocational range of arguments of a predicate. This is illustrated on the collocational range of the complements of the adjective *compatible*, using data from the British National Corpus World Edition (100 million words), from the 5-million-word Wordbank on CD (taken from the Bank of English corpus by Collins COBUILD), as well as from the Corpus of Spoken Professional American English (1.5 million words), plus my own collection of examples, all in all, 773 tokens of the structure in question.

Let us by way of introduction consider some examples of prepositional complements of *compatible* in the following data set. *Compatible* is occasionally found with prepositional complements introduced by *to*, as in (1), but by far the most frequent complements are prepositional phrases introduced by *with*, as in (2):

(1) a. The pot is then glaze fired; the glaze ingredients must melt and be come glasslike at a temperature that is *compatible to* that required for the clay.

(2) a. Depending on the bird species, plants can usually be chosen that are *compatible with* captive birds, the density and type of birds being critical factors.

b. These dates are *compatible with* the dates of the major upheavals in human populations of the New World,...

c. Even if he holds that an action is not free if it has causes that eventually lie outside the agent, his view will be *compatible with* the various views of action unless he holds the version of (1)...

d. The question of whether philosophy is *compatible with* religious law (the answer being sometimes negative) constituted the main theme of the foremost medieval Jewish thinkers.

e. Organic farming uses less petroleum than does conventional farming and is most *compatible with* diversified, small-scale, labour-intensive cultivation.

f. To be sure, Classical Theism holds to the freedom of man but insists that this freedom is *compatible with* a divine omniscience that includes his knowledge of the total future.

g. "Nuclear waste isn't *compatible with* tourism," insists Rep. Harry Reid.

h. Further treatment of the triacetate in solution in the presence of sulfuric acid splits off some acetic acid giving diacetate, soluble in acetone, and *compatible with* a range of plasticizers that can be incorporated in a rugged type of mixer without solvent to yield molding powders especially suited to injection molding.

i. The great depth of these submerged valleys, extending thousands of feet below sea level, is *compatible* only *with* a glacial origin.

j. Subsequent analysis of the hand bones from Swartkrans - which are presumed to be australopithecine - has demonstrated that they are *compatible with* tool use.

k. Skulls and teeth *compatible with* early bats are known from the Paleocene (about 60,000,000 years ago),...

It appears from what can be extracted from large collections of natural data that the semantic side of the set of collocates that function as complements of these predicative adjectives is less than fully tractable. Simply listing all the items that follow prepositions in complements (while excluding, of course, personal pronouns *it* and *they/them*, as well as demonstratives and reciprocals, since these pronouns, failing to indicate even basic semantic contrast such as animate/inanimate, do not seem to be informative on the issue we are concerned with here) results in a bewildering set of rather heterogeneous items. In addition to some well-behaved data that can be easily pigeonholed and described in a uniform way, there is a significant residue of obnoxious instances that do not fit with the rest, and must be apparently treated on a case by case basis as special, i.e. there is not only no common denominator but these examples do not even exhibit any family resemblance to each other in the broadest sense of the notion. The collocates of *compatible* found in the prepositional complement that were retrieved from the three corpora include among others:

(3) an ecological society, the existing systems, a GATT agreement, expected productivity, the polymer, the hypothesis, a maximal effect of gastrin, simple passive diffusion, your own interests, the rule of law, motherhood, an empirical approach, IBM personal computers, social progress, such deference to a received idea of the age, that belief, tourism, the charter, normal life expectancy, fabrics made from yarn of plant origin, the orderly arrangements of atoms found in crystals, health, the features and fossils found in deposits of this sort, modern notions of the close affinity between hominids and pongids, the above rules for applying the oxygen-atom-transfer and hydrogen-atom-transfer criteria, the environment, moral responsibility, its needs, measurements of the atmospheric compositions of the oldest stars, the organization, a solar wind source, its function, determinism, the press system, preset harmonies, the known laws of science, the standardized network, a long life span, the critical value, woodcut printing, the gay abandon of the true *pícaro,* any pollen that lands on it, cosmopolitan convictions, black-and-white transmission, good health, the tramlines, its military neutrality, economic and geographical conditions,

the new, large Trident submarines, rainfall observations, the tissue of a recipient, the conditions required for most waste landfills, French taste, that evidence, his artistic ambitions, the negligible possibility of such radical deception, current religious practices, such machinery, the conclusion, with Russian values and traditions, a clean environment, current body language, industry standards, American interests and ideals, conservative economics, more applications, all your programs, any table setting, any taste,...

The picture we get here verges on being a mind-boggling one. Starting from the meanings of the adjective, we might try to group these in a number of sets. *Compatible* is usually said to mean 'able to exist, live together, or work successfully with (something or someone else)' (Cambridge International Dictionary of English, CIDE), or 'able to exist, live together, or be used together or with (another thing)' (Longman Dictionary of Contemporary English, LDoCE). As will become clear from the argumentation that follows below, we should also pay attention to expressions functioning as subjects of clauses with *compatible* in the predicative position.

It appears from these definitions that it has two main senses: i. 'be able to exist or live together', and ii. 'to work successfully with something else'. In this second sense, it clearly collocates with a number of items listed above such as *IBM personal computers, such machinery, more applications, all your programs, the new, large Trident submarines*.

Turning to the first sense, we might establish a number of groups, such as:

(4) a. normal life expectancy, health, good health, a long life span,
b. such deference to a received idea of the age, that belief, moral responsibility, determinism, American interests and ideals

This would somewhat reduce the disorder emanating from the list in (5). Such lists could well be extended to accommodate further collocates from the list, but it is doubtful that the end result would be satisfying. On the one hand, it is possible that we would get a large number of very disparate groups of a relatively small number of collocates, so that basically the same chaotic nature is preserved. On the other hand, it is possible that we could establish a smaller number of groups, but these would then probably be internally very heterogeneous and therefore very difficult to defend and motivate. What is more, it is hardly possible that we could establish a group or groups containing items such as *tourism, a range of plasticizers, hand bones, skulls and teeth*, or *early bats*, if all the items that are listed in (3) are simply taken at their face value.

The most sensible solution is in my opinion to go the second way, i.e. establish as small number of groups as possible, but cut the Gordian knot of internal motivation by recognizing metonymy as a powerful conceptual principle uniting many seemingly disparate items.

The first couple of examples in (2) are relatively straightforward, particularly (2) a., and appear to conform to the most frequent dictionary entries for *compatible*. We have seen above what CIDE and LDoCE have to say, but the Oxford Advanced Learner's Dictionary (OALD) definition seems to be the most revealing of all. As shown above, there is again a sense that pertains to the ability to be used together. The definition indicates that the adjective is used in connection with machines, especially computers. In its second sense the adjective means 'to be able to exist or be used together without causing problems' and is used, according to the entry, in connection with ideas, methods or things. Finally, it is used in connection with people when they have a good relationship because they have similar ideas, interests, etc. I shall return to the second sense presently and elaborate its significance, i.e. what it captures, and what it fails to convey.

One important aspect of the meaning of *compatible* is described as follows by the COBUILD Dictionary: 'People who are *compatible* are able to live or work together in a friendly and peaceful way.' Notice that entities that are compatible are normally taken to be of the same rank, type, or belong to the same category, etc., and that a symmetrical relationship obtains between them. In sum, only likes, in the broadest sense, can be compatible. It is telling that the subject and the prepositional complement in (1) and (2) b. and c. contain the same head noun. It is *(temperature) that* and *that (temperature)* in (1), and in (2) b. and c. it is *dates* and *views*, respectively. Not infrequently, the subject is in the plural and the prepositional complement contains a reciprocal pronoun, an explicit indication of a symmetrical relationship, as in (5) a. and b. The complement can be left out altogether, as in (5) c.:

(5) a. These distributions are *compatible with* each other, a property that ensures that there exists some probability space and some family of random variables defined on the space that realizes the original stochastic process.

b. Individuals usually occupy several positions, which may or may not be *compatible with* one another: one person may be husband, father, artist, and patient, with each role entailing certain obligations, duties, privileges, and rights vis-à-vis other persons.

c. They are *compatible*.

The nouns in the prepositional phrase complementing *compatible* in (2) d-f. belong, broadly speaking, to the same domain or to similar and related domains: *philosophy – religion, farming – cultivation, freedom - divine omniscience*, and should therefore again be unproblematic. However, these examples illustrate another important aspect of the meaning of *compatible*, in addition to the symmetrical nature of the relationship.

This element is not sufficiently highlighted by the COBUILD Dictionary in the statement that 'Two things, systems of belief, ideas, etc. that are *compatible* can exist in the same place and at the same time without harming each other.' This statement is obviously very similar to the second subdefinition in OALD, in fact they overlap. Both mention ideas and things as entities of which compatibility could be predicated. One adds methods to this, and the other mentions beliefs. I will argue below that ideas, beliefs and methods do make sense here, but that the other point where the two dictionaries agree, i.e. concerning things, is misleading.

I would like to stipulate that *compatible* has an inherently very narrow range of collocates, i.e. imposes fairly strict selectional restrictions on their choice. The relationship of compatibility can inherently obtain only between abstract entities, and the second statement in COBUILD as well as the second sense in OALD reflect some facts of usage based on metonymic extension.

All the concrete nouns in the subjects or complements of *compatible* in examples in (2), such as *nuclear waste* and *plasticizers*, and particularly those in (2) j-k., *hand bones*, *skulls and teeth*, and *early bats*, do not make much sense in these contexts. Of course, such items can be conveniently subsumed under things and thus conform to the entries. However, this would practically devaluate entries, because it would imply that any concrete noun could in principle appear in any sentence with *compatible*, without actually telling us why.

One residual problem for these dictionary definitions is that there are many collocates attested above which simply fall somewhere between being concrete nouns and the three abstract nouns named. For example, *glacial origin* or *tourism*, are not concrete, and are therefore not things, but they are not methods, ideas or beliefs. What is more, in many cases the subject and the prepositional complement are so disparate that they can logically hardly be seen as exhibiting any compatibility in any ordinary sense of the concept.

But if we assume that there is some sort of hidden logical compatibility, not expressed explicitly in these words but which speakers apparently establish very quickly and largely subconsciously by means of metonymic mapping or interpretation, all these data make perfect sense. Before I demonstrate how metonymy can be employed to reduce the apparent chaos, let me briefly sum up a couple of relevant points concerning the role and place of metonymy in the cognitive linguistic framework.

Adopting a cognitive linguistic approach means that all linguistic cases of the phenomenon traditionally called metonymy are reflexes of deeper running conceptual metonymies. Conceptual metonymy, just like metaphor, is one of the most basic and ubiquitous cognitive processes that closely link all our thinking, speaking and acting. It is traditionally approached as a stand-for relationship that is, unlike metaphor, not based on similarity but on contiguity or proximity. Contiguity is taken in its broader sense to cover all associative relations except similarity. This means that metonyms are expressions that are used instead of some other expressions because the latter are associated with or suggested by the former:

(6) a. *The White House* declined to comment on the issue.

b. Keep your *eye* on the ball!

In the two examples above, the expressions *the White House* and *your eye* are metonyms used for *the U.S.President and his advisers* and *your gaze*, respectively. The standard view is that a metonymic mapping occurs within a single domain, while metaphoric mapping takes place across two discrete domains. It is also possible for metonymic mapping to occur within a single domain matrix which involves a number of subdomains (cf. Croft 1993: 348). In other words, metonymic mapping across different domains within a single domain matrix, involving the conceptual effect of domain highlighting, is also possible.

As for the nature of the metonymic mapping, Kövecses and Radden (1998: 39) aptly note that it is "a cognitive process in which one conceptual entity, the vehicle, provides mental access to another conceptual entity, the target, within the same domain, or ICM [Idealized Cognitive Model]". One of the most important aspects of this definition is that metonymy provides mental access to a conceptual entity that need not be otherwise readily and easily accessible. Figuratively speaking, metonymy is an efficient mental shortcut making it possible for us to refer to entities for which there are no current or convenient (in the sense of being short and compact) linguistic expressions.

Now that we have highlighted the basics of metonymic mappings, we may return to the problem of collocational range of *compatible*. The metonymic shortcut in question works as follows: all the concrete nouns explicitly named are elements of various IDEALIZED COGNITIVE MODELs or cognitive domains which also contain by definition one other prominent abstract element that is not named but is actually the target of the metonymic mapping.

This seems to explain why some apparently subject-complement pairings sound much better than some others. We might assume that a crucial requirement for such pairings is that they should belong to the same cognitive domain or IDEALIZED COGNITIVE MODEL, or, if not to the same, than at least to two sufficiently close domains or models, possibly contained within a single matrix domain or superordinate model.

This targeted but not explicitly named element is in fact partly hinted at in the definitions by COBUILD and OALD – systems of belief such as theories, scientific or folk ones, religious systems, and also single ideas and facts which are ingredients of such systems, but also other aspects of belief systems such as (scientific) methods, and quite generally activities related to the concrete objects named. *Skulls and teeth* and *early bats* in (2) k. are shortcuts for something more general, like *the results of an analysis*, *assumption about*, *hypothesis*, *idea* or *theory on/about*. In (2) i. it is *the fact* of exhibiting great depth that is compatible with the *hypothesis/assumption* that the valleys in question are of glacial origin. This line of analysis may now also extend to collocates separately listed in (4) a., such as *expec-*

tancy, because what lies behind all these expressions is in fact some sort of idea or assumption.

In all these cases we might assume that physical objects metonymically stand for mental objects. So one part of the collocates of *compatible* may more generally be described as referring directly or metonymically to mental objects, the latter reference being achieved by explicit reference to salient elements of a cognitive domain or ICM involving these mental objects. Saliency may arise due to conventionalized links, or may arise more locally, when the relevant portion of discourse activates the domain or model in question.

This PHYSICAL-OBJECT-FOR-MENTAL-OBJECT metonymy is a specific instantiation of a more general type of PART-FOR-WHOLE metonymy, since the physical object in question is just one, though crucial, element out of several possible ones that may constitute the mental object that represents it. However, it will be seen that mental object is in turn part of a more general domain evoked by the utterance or a larger portion of discourse.

The remaining, smaller portion of collocates listed in (3) lends itself to a similar generalization, again based on a specific instance of PART-FOR-WHOLE metonymy. *Nuclear waste* and *tourism* in (2) g. stand for more general activity concepts, such as the *disposal or treatment of waste* and *engagement in the industrial activity of tourism*, respectively. It is of course possible that *nuclear waste* may merely be present, so that there is no activity involved here. In that case we would have something like *the idea/fact of the presence of nuclear waste*, i.e. again the PHYSICAL-OBJECT-FOR-MENTAL-OBJECT type metonymy. The context, however, makes it more likely that an activity is meant, more precisely, proposed activity of disposing of nuclear waste, because it is obvious from the whole magazine article that the Republican Senator is trying to protect the tourist industry in part of his constituency against a proposal to dispose nuclear waste in the region. This PHYSICAL-OBJECT-FOR-ACTIVITY type of metonymy is a special case of the PARTICIPANT-FOR-ACTIVITY metonymy, which is independently needed to account for apparently anomalous collocational ranges of adjectives such as *fond, sorry,* or *happy*. Cf. the following examples:

(7) a. Garter snakes live chiefly on insects, earthworms, and amphibians; the ribbon snake is especially *fond of* frogs.

b. Bustards are *fond of* grasshoppers, and their varied diet also includes dung beetles, termites, centipedes, grass, clover, vegetable crops, and even, in Africa, the gum from the trunks of *Acacia* trees.

c. Several examples of his work have survived, and they are sufficient to establish him as a painter of great ability, *fond of* rich, sensuous colour and softly modeled forms.

(8) a. Poor souls, I'm heartily *sorry for* them.

b. Long after he ceased to love her and was irked by her presence, he remained *sorry* for her.

(9) Bobbi and Kenny McCaughey seem thrilled to have seven new babies in tow, and last week's headlines show the world is *happy for* them.

The first two examples need not make it necessarily clear that garter snakes and bustards are fond of frogs and grasshoppers, respectively, not as such, i.e. as species of animals, or as companions, etc., but only as food, i.e. they are fond of eating these animals. However, the other two examples in (7) are compatible only with an underlying activity interpretation, not with one on the consumption end but rather on the production end. Saying that certain paintings are sufficient to establish someone as a painter of great ability, and then bringing in his fondness for certain colours, can only mean that he was fond of using these colours. Monks may be fond of acrostics, i.e. of reading them, but the second clause identifies them indirectly as writers, i.e. as poets, a fact otherwise in keeping with our encyclopaedic knowledge about the Middle Ages, and if they are now seen in the role of poets, then the only sensible interpretation of being fond of acrostics is of being fond of writing acrostics. Similarly, in (8) and (9) *sorry* and *happy* are followed by *for* and NPs denoting people, but the nouns actually refer to states of affairs, i.e. to what happened to these people. In other words, with all these adjectives we have PARTICIPANT-FOR-ACTIVITY metonymies.

Returning to *compatible*, we may now even entertain the possibility of assimilating many of the collocates associated with the first sense of *compatible*, being 'able to work or be used together with sth.'

At an even more general level, we may consider the fact that it is sometimes difficult to pinpoint a single most appropriate targeted expression as an indirect piece of evidence that the collocates function as metonymies. As pointed out above, metonymies provide more direct access to concepts that otherwise might be difficult to think and talk about.

Another relevant fact supporting the assumption about metonymic mappings is the observation that some targets, or near-targets are explicitly named in the broader context, and thus invite appropriate inferences, e.g., the words *treatment* and *soluble* in (2) h., may evoke the idea of use or processing of chemicals, while the expression *subsequent analysis* in (2) j. prompts us to activate our encyclopaedic knowledge telling us that analyses normally produce some findings or results and that these results provide the rationale for performing these analyses in the first place, and further to infer that findings of analyses enable the formation of certain beliefs, which are either compatible or incompatible with the hypothesis about tool use.

In short, we observe here a metonymic mapping of a referential type from a part of an ICM to the whole ICM. It has already become clear that this sort of analysis is not peculiar to *compatible*, it is far more general and could easily be performed not only on a number of more or less synonymous predicative adjectives expressing the idea of symmetrical relationship, such as *comparable with*, *consistent with*, etc., but on whole sections of the adjective stock of English (cf. Brdar 2000).

The nature of the mapping is quite in keeping with the findings of Kövecses and Radden (1998), who note that mappings from concrete to abstract are more natural than the other way round. In the specific case of *compatible*, we see how metonymy effectively broadens the range of possible collocates, and thus significantly determines some semantic aspects of its valency frame.

In any case, it appears now that this large initial mass of unordered and disparate collocates of *compatible* has been reduced in such a way that it becomes manageable, descriptively and lexicographically. As for the latter, I suggest that in the relevant sense, the only data needed at the highest level in the part of entry dealing with collocational potential are the specifications that they can be mental objects or activities, which should be accompanied not only by some prototypical or default examples but also by a statement as to possible metonymic extensions, of course in a didactically adequate wording which need not actually use the term metonymy at all but rather practically evoke the idea of domains or models and their constituent parts.

The case study seems to indicate that metonymy provides us with an extremely useful tool in keeping a balance between the richness of data input, on the one hand, and the optimal degree of generality in their interpretation and presentation, i.e. saving the selectional restrictions idea, on the other. It has been shown that introducing metonymy (as it is understood in cognitive linguistics) into the conceptual apparatus of lexico-grammatical enterprises, in this specific case, into the account of the predicate-argument structure, as well as into the whole lexicon structure, results in making the above and various types of lexicographic handbooks carrying them, more functional and streamlined, making individual entries more compact. In fact, it could be argued that it in fact may result in an increase of the generative power of lexicographic handbooks in the sense of enabling their users to make relatively safe guesses about the acceptability of some novel grammatical combinations not actually recorded in the lexicon or corpora.

References:

Brdar, Mario (2000). "Metonymy as a motivating factor in the system of adjective complementation in English." Suvremena lingvistika 25.1-2 (49-50): 41-55.

Croft, William (1993). "The role of domains in the interpretation of metaphors and metonymies." Cognitive Linguistics 4.4: 335-370.

Kövecses, Zoltán, Günter Radden (1998). "Metonymy: Developing a cognitive linguistic view." Cognitive Linguistics 9.1: 37-77.

Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns: Describing Event Structure in Support Verb Constructions

Silvie Cinková, Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
{cinkova,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We have recently launched a project of an XML-based bilingual lexicon of predicate nouns. Besides itemizing the commonest support verb constructions and their Czech translation equivalents, the lexicon of predicate nouns aims at providing the user with relevant construction rules. It is mainly meant to help advanced Czech learners of Swedish to master delexicalized uses of the commonest lexical verbs in SVCs. It provides a systematic description of the valency of the nouns. Apart from that, it provides their support verb collocates, sorted by the Mel'čukian Lexical Functions.
A cross-linguistic application of the Transitivity Hypothesis is used in attempt to illustrate the lexical way of rendering event structure in Swedish, which poses problems for speakers of Czech, a language with morphological aspect. We believe that the morphosyntactic behavior of the noun together with telicity conditions affect the event structure of the entire SVC in context. The structure of the lexicon is theoretically based on the Functional Generative Description (FGD).

Introduction

This paper describes bilingual lexicographical processing of support verb constructions (SVCs) in a recently launched project of a machine-readable XML-based Swedish-Czech lexicon of predicate nouns. The lexicon is meant to capture delexicalized uses of the commonest lexical verbs in SVCs, in which the verbs show an evident tendency to grammaticalization (as defined by (Hopper, 1987) and further analyzed by (Heine, Claudi and Hünnemeyer, 1991)).

Support Verb Constructions, Support Verbs, Predicate Nouns

Support verb constructions are combinations of a lexical verb and a noun containing a predication. From the semantic point of view, the noun seems to be part of a complex predicate rather than the object (or subject) of the verb, despite what the surface syntax suggests. Support verbs are understood as verbs occurring in SVCs. Predicate nouns are in general nominal components of complex predicates (including SVCs).

An SVC is usually semantically transparent. Its meaning is concentrated in the noun phrase, while the semantic content of the verb is reduced or generalized. The matching verb is unpredictable, though often a metaphorical motivation can be traced back. Implicitly, SVCs affect the foreign language production rather than the reception (Heid, 1998), (Malmgren, 2002) and (Schroten, 2002).

If we look upon SVCs as collocations, the noun is the base, while the verb is the collocate; cf. e.g. (Malmgren, 2002), (Čermák, 2003) and (Schroten, 2002). Even in the cross-linguistic perspective it is the noun that constitutes the common denominator for equivalent support verb constructions, as empirically shown by (Fontenelle, 1992), whereas the support verbs do not necessarily match.

Important Features of the Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns

Besides itemizing the commonest SVCs and giving their Czech translation equivalents, the lexicon aims at providing the users with relevant SVC-construction rules for varying communication needs with special regard to event structure. The lexical evidence is always corpus-based.

Lemmatizing nouns both enables the enumeration of all verbs semantically related to the given noun together at one place and a more systematic description of restrictions in morphological number, article use and adjectival or pronominal modifications in the nouns. Inspired by (Hopper and Thompson, 1980), (Lindvall, 1998) and (Bjerre, 1999), we believe that morphosyntactic behavior of the noun together with lexical features of the support verb and of the event described by the predicate noun determine the event structure of the entire SVC employed in context.

Describing Valency in Predicate Nouns: Functional Generative Description

The lexicon displays the valency of the lemmatized predicate nouns within the FGD framework – a dependency-based formal stratificational language description framework that goes back to the functional-structural Prague School. For more detail see (Panevová, 1980) and (Sgall, Hajičová and Panevová, 1986). The theory of FGD has been implemented in the Prague Dependency Treebank project (Sgall, Panevová, Hajičová, 2004), a syntactically parsed corpus of Czech.

FGD can capture valency in the underlying syntax (the so-called *tectogrammatical language layer*). It enables listing of complementations (syntactically dependent autosemantic lexemes) in a valency lexicon, regardless of their surface (morphosyntactic) forms, providing them with semantic labels (*functors*) instead. It also regards coreference, ellipsis and topic-focus articulation. Implicitly, a complementation present in the tectogrammatical layer can either be directly rendered by the surface shape of the sentence, or it is omitted but can be inferred from the context or by common knowledge. A valency lexicon describes the valency patterns of a given lexeme (verb, noun, adjective or adverb) in form of *valency frames*. In a valency lexicon the frames roughly correspond to lexical units in ordinary lexicons.

The lexicon of predicate nouns was significantly inspired by the closely related valency lexicons of Czech verbs VALLEX (though machine-readable, also designed for human use) and PDT-VALLEX (a supporting tool for treebank annotation, interlinked with the corpus data) – cf. (Straňáková-Lopatková et. al., 2002) and (Hajič et al., 2003). Though the lexicon of predicate nouns is primarily meant for human use, the quite rigid structure of (PDT-)VALLEX, whose both variants have originated from the needs of NLP-applications, seems to be helpful in remaining consistent when describing complex linguistic phenomena. On top of that, the Swedish part is very likely to prove useful in a possible annotation of an FGD-based Swedish treebank, when interlinked with the data in the same way as PDT-VALLEX.

Ordering Support Verbs under the Predicate Noun Lemmas: Lexical Functions

The lexicon of predicate nouns is also called a "combinatorial" one to show the acknowledgement for existing collocational dictionaries that have paid systematic attention to support verb constructions, e.g. (Benson, Benson and Ilson, 1986) and especially (Mel'čuk et al., 1984, 1992) and (Mel'čuk and Žholkovsky, 1984), modeling "institutionalized" lexical relations by the so-called Lexical Functions.

Lexical Functions are part of the Meaning-Text-Theory developed by Igor Mel'čuk and his collaborators (Mel'čuk, 1988), (Kahane, 2003). There are two elementary types of LFs – paradigmatic and syntagmatic – and this paper concerns only the latter. In terms of collocations, when two lexical units are collocates, one is usually the base that "selects" the other lexical unit to render a certain meaning together. The MTT captures it by the mathematical functional notation: LF_i (X) = Y, where X is called the keyword (the collocational base) and Y the value of the LF_i (the collocate). LFs can assign one value or a set of values to a given keyword. The values stand in the same lexical relation towards the keyword but they are not necessarily synonymous. The LFs describe the semantic relation between the keyword and the values. For examples and more details see (Wanner, 1996).

The following LFs are specific to SVCs; their keywords are the predicate nouns and their values are by definition verbs: **Oper$_1$**, **Oper$_2$**, **Labor$_{1,2}$**, **Copul** and **Func**.

In **Oper**, the predicate noun is a direct object of a transitive support verb, e.g. *pay attention*) or a prepositional object of an intransitive support verb, e.g. *get in touch*.

In **Labor**, the predicate noun is a prepositional object of a transitive verb, e.g. *subject sb to an interrogation*.

In **Copul**, the noun (or the adjective) is part of the predicate, in which a lexical verb has acquired a copula-like meaning, e.g. *fall ill. (= start to be ill)*.

In **Func**, the predicate noun is the subject of the verb, e.g. *The accusation came from John*.

(The example sentences originate from (Wanner, 1996) and (Macleod, 2002).) The numbers denote indexes of the complementations (participants) of the events described. No. 1 is the Actor, No. 2 is the Patient. When an LF is specified by 1, it means that the Actor of the verbal event is identical with the Actor of the event described by the noun. When an LF is specified by 2, it means that the Actor of the verbal event is identical with the Patient of the event described by the noun.

Entry Structure

On the topmost level, the lexicon is divided into word entries. Each word entry relates to one predicate noun lemma and its possible spelling variants. Homonyms get each an indexed word entry. Each entry comprises valency frames of the given predicate noun. The frames regard the noun simply as an abstract noun standing outside any SVCs. Fig. 1 shows two valency frames of the lemma *kritik*. The noun governs two complementations with functors. Their surface forms are also listed.

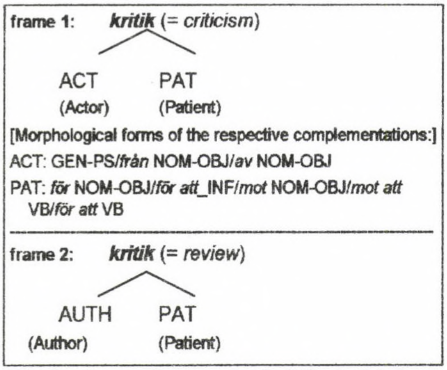
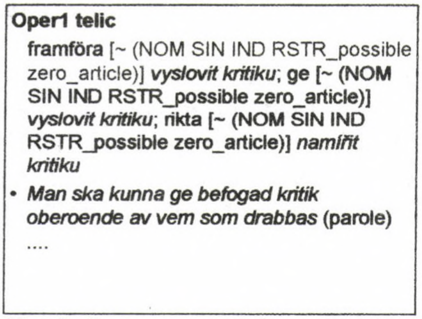

frame 1: ***kritik*** (= *criticism*)
ACT
(Actor)
PAT
(Patient)
[Morphological forms of the respective complementations:]
ACT: GEN-PS/*från* NOM-OBJ/*av* NOM-OBJ
PAT: *för* NOM-OBJ/*för att*_INF/*mot* NOM-OBJ/*mot att* VB/*för att* VB
frame 2: ***kritik*** (= *review*)
AUTH
(Author)
PAT
(Patient)

Fig.1: A word entry for kritik *(criticism) with two valency frames. The first frame includes the surface forms of the complementations by means of the SUC tagset (Ejerhed et al., 1992). Only the first frame renders the lexeme* kritik *as a predicate noun, the second frame shows* kritik *as an artefact, which cannot be a predicate noun.*

Each valency frame that renders a predicate noun (i.e. not the case of frame 2 in *kritik*) lists relevant SVCs, grouped according to LFs. These groups are technically called SVC-frames. An example is given by Fig. 2.



Oper1 telic

framföra [~ (NOM SIN IND RSTR_possible zero_article)] ***vyslovit kritiku***; **ge [~ (NOM SIN IND RSTR_possible zero_article)]** ***vyslovit kritiku***; **rikta [~ (NOM SIN IND RSTR_possible zero_article)]** ***namířit kritiku***

- ***Man ska kunna ge befogad kritik oberoende av vem som drabbas*** **(parole)**

....

Fig. 2: One support verb construction frame nested in the first valency frame of kritik. *It is defined by the Lexical Function* **Oper**$_1$. *It includes telicity marking, description of morphosyntactic characteristics of the predicate noun* kritik *in combination with the support verbs* framföra, ge *and* rikta *(in square brackets after each verb), and Czech translation equivalents (in italics). Also an example sentence with reference to the PAROLE-corpus (http://spraakbanken.gu.se) is attached.*

Describing Event Structure in Swedish SVCs

SVCs are often referred to as one means of specifying event structure in non-aspectual languages as Swedish. A kind of event structure opposition is assumed between a SVC and its corresponding synthetic predicate (when there is any). Support verbs add further semantic features to the event described by the given predicate noun, such as inchoativity, durativity, terminativity and causativity (called *aspectual, diathetic* and *modal values* by (Fontenelle, 1992), or simply *aktionsart* by others, e.g. (Šmilauer, 1972)). However, this gives no direct correspondence to the Slavic category of aspect, which apparently is the product of more event structure features in combination, one of which being telicity. Also (Hopper and Thompson, 1980) emphasize the difference between aktionsart (which they call "lexical aspect", telicity and perfectivity (grammatical aspect).

Telicity, introducing the values "telic" and "atelic" should be regarded as independent of "aspect"/"perfectivity"/"boundedness" with its values "perfective" and "imperfective". More to this issue see (Nakhimovsky, 1996): *"A verb lexeme is telic if a simple declarative sentence in the past tense in which that lexeme is the main predicate is a telic sentence. A sentence is telic if it describes a telic process. A process is telic if it has a built-in terminal point that is reached in the normal course of events and beyond which the process cannot continue."* Nakhimovsky's claim that telicity is a lexical feature (i.e. semantically inherent to the verb in question) while aspect is inferred from semantico-syntactic relations in each given sentence, corresponds to (Pustejovsky, 1991), who, speaking of event-types, claims that *"the lexical specification of a verb's event-type can be overridden as a result of syntactic and semantic compositionality of the verb with other elements in the sentence"* and (Hopper and Thompson, 1980): *"Whereas telicity can be determined generally by a simple inspection of the predicate, perfectivity is a property that emerges only in discourse".*

To summarize it, aktionsart and telicity are two different quantities, though they both are lexical features. Besides that, they both are to be discriminated from the grammatical aspect, whose morphological form they probably co-determine in aspectual languages as Czech.

The lexicon of predicate nouns captures aktionsart by attaching complementary LFs to the basic LFs listed above. It is phasal LFs – **Inc** (inceptive, inchoative), **Cont** (continuative) and **Fin** (finishing, terminative), causative LFs – **Caus** (causation), **Liqu** (causing to stop) and **Perm** (permitting to continue). Two more complementary LFs are employed, i.e. **Prox** (to be on the verge of) and **Anti** (negation). For details see (Wanner, 1996). The Anti-LF is mainly stated when the negation of the predicate noun is not allowed to negate the SVC and other means have to be used instead, such as the negation of the verb or using a support verb with the opposite meaning. The Anti-LF is not being stated consequently due to the lacking lexical evidence.

Issues of Telicity Marking in SVCs

It is to be stressed that SVCs are built as compositional events consisting of a "verbal" and a "nominal" subevent. Yet the "verbal" event does actually never "take place" due to the semantic depletion in support verbs (cf. (Fillmore, Johnson and Petruck, 2003)). The given support verb only passes some semantic features on to the "nominal" event. Durative events are by definition atelic (e.g. *have problems*), with the reservation that multiple telic "nominal" events combined with a durative atelic support verb express iterativity, e.g. *suffer from attacks*. (Below the "verbal" event corresponds to *subevent1* and the "nominal" event to *subevent$_2$*.)

SVCs denoting transitions (i.e. changes of state) are regarded as telic (cf. (Pustejovsky, 1991)), no matter what telicity value the given support verb would have if used as a lexical verb outside the SVC. This approach is based on (Bjerre, 1999). Bjerre puts it this way: "*SVCs denoting transitions are invariably achievements*[1], *either inchoatives or causatives [...], the SV always denotes an underspecified subevent$_1$. [...] Not surprising* terminative *is the negative counterpart of* inchoative. */*Situationen kom ud af kontrol – (Situation_the came out of control)*] denotes a situation in which the resultant state is the negative of that in [*Situationen kom under kontrol = Situation_the came under control*] above. [...] This may be paraphrased: (subevent$_1$:) The situation was under control when something happened as a result of which (subevent$_2$:) the situation was out of (=not under) control".* Bjerre notes that support verbs denoting transitions are either achievement verbs with inherently underspecified subevent$_1$ (*come, bring* etc.), or they are verbs of motion or location which lose their specific relation when used as support verbs.

[1] Transitions are further divided into two subtypes. In *achievements* the subevent$_1$ is underspecified, unlike in *accomplishments*, e.g. *Carl built a house* (accomplishment) × *The expedition reached the top of a mountain* (achievement). See (Bjerre, 1999).

For the purpose of the lexicon of predicate nouns, an SVC is thus marked as telic when:

a) both the subevent described by the predicate noun and the subevent described by the support verb are telic, e.g. *fatta beslut (take a decision)*

b) the subevent described by the support verb is atelic and the subevent described by the predicate noun is telic, e.g. *dra en slutsats (draw a conclusion)*

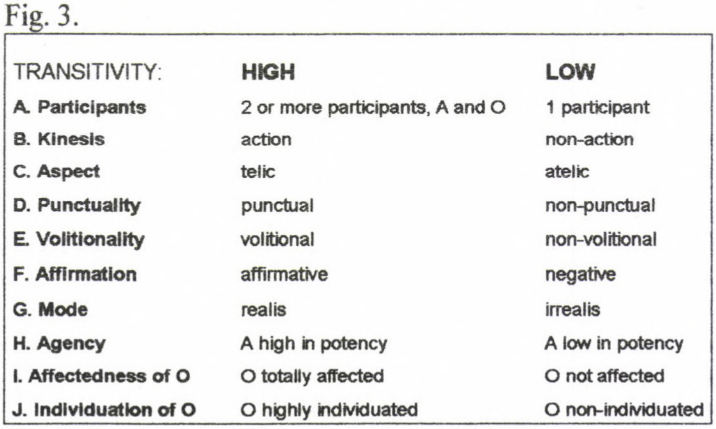
c) the subevent described by the support verb is telic and the subevent described by the predicate noun is atelic, e.g. *få besvär (get problems).*

The event a) describes the termination of a process, and so does the event b) while the event c) describes the onset of a state, thus is inchoative (inceptive).

Perfectivity as a Transitivity Component

Our attempt to make a link between the Swedish and the Czech ways of specifying event structure is based on (Lindvall, 1998) and on (Lindvall, 2001), a summarizing article. Lindvall has performed a comprehensive parallel-corpora based comparison of Greek, Polish and Swedish to look into verbal boundedness and object definiteness as two interacting components of Transitivity. We make use of her inferences regarding Swedish and we assume that her inferences regarding Polish will also apply to Czech, as Czech and Polish are tightly related languages.

Lindvall's point of departure is (Hopper and Thompson, 1980). By comparison of many unrelated languages they analyze Transitivity, a universal linguistic phenomenon, intuitively understood as transfer of an activity from an Agent to a Patient, producing some effect. Hopper and Thompson isolate component parts/ parameters of the Transitivity notion with regard to the information structure of the given utterance, concluding that Transitivity is a continuum. Their parameters of Transitivity suggest each a scale according to which clauses can be ranked – see Fig. 3.



TRANSITIVITY:	**HIGH**	**LOW**
A. Participants	2 or more participants, A and O	1 participant
B. Kinesis	action	non-action
C. Aspect	telic	atelic
D. Punctuality	punctual	non-punctual
E. Volitionality	volitional	non-volitional
F. Affirmation	affirmative	negative
G. Mode	realis	irrealis
H. Agency	A high in potency	A low in potency
I. Affectedness of O	O totally affected	O not affected
J. Individuation of O	O highly individuated	O non-individuated

Fig. 3. Components of Transitivity proposed by Hopper and Thompson. The letter A *means Agent,* O *means Object.*

Hopper and Thompson further claim that the component features of Transitivity *"CO-VARY extensively and systematically [...] whenever an obligatory* *pairing of two Transitivity features occurs in the morphosyntax or semantics of a clause, THE PAIRED FEATURES ARE ALWAYS ON THE SAME SIDE OF THE HIGH-LOW TRANSITIVITY SCALE".* They introduce the Transitivity Hypothesis: *"If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any of the features A-J, then, if a concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity."*

Lindvall has proved that the Transitivity Hypothesis applies even cross-linguistically, having shown on Greek (a language employing both morphological aspect and noun definiteness) that utterances with high Transitivity tend to have perfective verb forms and definite objects, while utterances with low Transitivity tend to have imperfective verb forms and indefinite objects. Then she compared translations between Swedish (a noun-definiteness language) and Polish (an aspectual language) in both directions. It proved evident that in utterances with high Transitivity, Polish translations from Swedish tend to have perfective verb forms and Swedish translations from Polish tend to have definite noun forms, while low Transitivity utterances tend to have imperfective verb forms (Polish) and indefinite noun forms (Swedish). The observed noun definiteness was not confined to morphosyntactic features but resulted from the semantics of the noun phrase, which, on the other hand, was very often reflected by morphosyntax. This is why the lexicon of predicate nouns includes a detailed description of the morphosyntactic behavior of the predicate nouns in SVCs. For more details on the data structure of the lexicon see (Cinková and Žabokrtský, 2005).

A special feature of SVCs is that telicity is not determined by the verb but by the "nominal" event (yet modified by the support verb, cf. above). It is again the definiteness of the predicate noun that co-determines perfectivity. This will become apparent in selections of verb aspect forms in Czech translations of Swedish utterances and in noun definiteness in Swedish translations of Czech utterances. We assume that SVCs, especially those denoting transitions, have potentially rather high Transitivity also according to other parameters. Just to name a few, predicate nouns in SVCs characterized by all LFs except **Func** and **Copul** are morphosyntactic objects totally affected by the support verbs – the "nominal" events "come into existence" only by being named together with the given support verb – cf. the discussion of "effected objects" in (Barón and Herslund, 1998). Besides that, SVCs used as a means of transforming a state or a process into a transition imply discourse foregrounding. Yet the degree of Transitivity of an utterance in discourse shifts with other parameter values, especially with volitionality, affirmation and mode (see Fig. 3), which could explain the rather high morphosyntactic variation in predicate nouns captured in the lexicon.

Conclusion

The Swedish-Czech Combinatorial Valency Lexicon of Predicate nouns is an attempt to make use of the Transitivity Hypothesis, cross-linguistically applied by

(Lindvall, 1998), in order to describe the potential of event structure modifications in Swedish SVCs for Czech learners. The ultimate objective is to help Czech learners of Swedish with bridging the mental gap between an aspectual and a non-aspectual language by better understanding and active usage of the lexical mechanisms that affect event structure in Swedish SVCs.

Acknowledgements

This work has been kindly supported by GAUK 489-2004 and GAČR 405/04/0243.

References

I. Barón and M. Herslund. 1998. *Support Verb Constructions as Predicate Formation.* In " The Structure of the Lexicon in Functional Grammar". John Benjamins.

M. Benson and E. Benson and R. Ilson. 1986. *The BBI Combinatory Dictionary of English.*

T. Bjerre. 1999. *Event Structure and Support Verb Constructions.* In "Proceedings of the ESSLLI Student Session 1999.

S. Cinková and Z. Žabokrtský. 2005. *Treating Support Verb Constructions in a Lexicon: Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns.* In "Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes – Saarland University". pages 22-27. Saarbrücken.

F. Čermák. 2003. *Abstract Nouns Collocations: Their Nature in a Parallel English-Czech Corpus.* In "Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora". Birmingham.

E. Ejerhed et al. 1992. *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project, Version 4.31.* Publications from the Department of General Linguistics, University of Umeå, no. 32.

Ch. J. Fillmore, Ch. R. Johnson and M. R. L. Petruck. 2003. Background to FrameNet. *FrameNet and Frame Semantics. International Journal of Lexicography* (Special Issue, Guest Editor: T. Fontenelle)16: 235-250.

T. Fontenelle. 1992. *Cooccurrence Knowledge, Support Verbs and Machine Readable Dictionaries.* In "Proceedings of the 2nd International Conference on Computational Lexicography, COMPLEX'92, Budapest, Hungary". Linguistic Institute, Hungarian Academy of Sciences. pages 137-145. Budapest.

J. Hajič et al. 2003. PDT-VALLEX: *Creating a Large-coverage Valency Lexicon for Treebank Annotation.* In "Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Växjö, Sweden, November 14 - 15, 2003". pages 57-68.Växjö.

U. Heid. 1998. *Towards a corpus-based dictionary of German noun-verb Collocations.* In "Actes EURALEX'98 Proceedings". pages 301-312. Liège.

B. Heine, U. Claudi and F. Hünnemeyer. 2001. *Grammaticalization. A conceptual framework.* Chicago.

P. Hopper. 1987. Emergent Grammar. *BLS,* 13:139-157.

P. Hopper and S. A. Thompson. 1980. Transitivity in Grammar and Discourse. *Language*, 56:251-299.

S. Kahane. 2003. *The Meaning-Text Theory.* In "Dependency and Valency. An International Handbook on Contemporary Research". Berlin.

A. Lindvall. 1998. *Transitivity in Discourse. A Comparison of Greek, Polish and Swedish.* Lund.

A. Lindvall. 2001. Grund, aspekt och definithet – en studie i morfologi i grekiska, polska och svenska. In: "Postskriptum. Språkliga studier till minnet av Elsie Wijk-Andersson". pages 170-184. Uppsala.

C. Macleod. 2002. *Lexical Annotation for Multi-word Entries Containing Nominalizations*. In "Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Canary Islands, Spain". pages 943-948.

S.-G. Malmgren. 2002. Begå *eller* ta självmord*? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000.* Göteborg.

I. A. Mel'čuk et al. 1984, 1992. *Dictionnaire explicatif et combinatoire du français contemporain*, Volume I and II. Montreal.

I. A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice.* New York.

I. A. Mel'čuk and A.K. Žholkovsky. 1984. *Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka.* In "Wiener Slawistischer Almanach. Sonderband 14." Vienna.

A. Nakhimovsky. 1996. *A Case of Aspectual Polysemy, with Implications for Lexical Functions.* In: "Lexical Functions in Lexicography and Natural Language Processing". Studies in Language Companion Series (SLCS), Vol. 31. pages 169-179. Amsterdam-Philadelphia.

J. Panevová. 1980. *Formy a funkce ve stavbě české věty.* Praha.

J. Pustejovsky. 2000. *Syntagmatic Processes.* In "Handbook of Lexicology and Lexicography". de Gruyter.

J. Pustejovsky. 1991. The Syntax of Event Structure. *Cognition*, 41:47-81.

J. Schroten. 2002. *Light Verb Constructions in bilingual dictionaries.* In "From Lexicology to Lexicography". pages 83-94. Utrecht.

P. Sgall, J. Panevová and E. Hajičová. 2004. *Deep Syntactic Annotation: Tectogrammatical Representation and Beyond.* In "Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference", pages 32-38.

P. Sgall, E. Hajičová and J. Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht, Prague.

M. Straňáková-Lopatková et. al. 2002. *Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation.* In "LREC2002, Proceedings, vol. III. ELRA, pages 949-956.

V. Šmilauer. 1972. *Nauka o českém jazyku.* Praha.

L. Wanner. 1996. (Ed.) *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language Companion Series (SLCS), Vol. 31. Amsterdam-Philadelphia.

Norsk Ordbok 2014 from manuscript to datebase - standard gains and growing pains

ODDRUN GRØNVIK

Norsk Ordbok 2014, University of Oslo
P.O.Box 1021 Blindern, N-0315 Oslo
oddrun.gronvik@iln.uio.no

This paper will present a review of the digitisation process of of a major academic dictionary through the initial phase of the Project Norsk Ordbok 2014 (2002 – early 2005). The hypothesis at the start was that a thorough revision of editorial practice, linked to creating a stringent digitised dictionary writing system, would create a more reliable and consistent dictionary, with clearer procedures for processing source materials and composing entries. An efficient Dictionary writing system (DWS) application would also help train new editors and make them productive in less time than what has traditionally been assumed necessary.

Having publishing the first volume after the project started, and being well under way with the next one, it can be shown that the major goals described below on the whole have been achieved. The paper will discuss some areas in depth, look at the advantages, but also point out some possible pitfalls and some lasting difficulties.

Background

Norsk Ordbok ('The Norwegian Dictionary') was started in 1930 with the aim of providing a scholarly and exhaustive account of the vocabulary of Norwegian dialects and the written language Nynorsk, one of the two official written standards for Norwegian. The model was that of the big academic dictionaries for English, German, Swedish and Danish.

However, Norsk Ordbok differs from most academic documentary dictionaries for European languages in using records of spoken language as well as literature for its source material. Determining an etymology and suggesting a standard form for words documented only through dialect transcripts of necessity forms part of the lexicographical work, and adds to its complexity, but has to be seen as an integral part of Nynorsk lexicographical tradition (Grønvik 1992)

The task was underestimated from the start both in terms of complexity and effort. The work on Norsk Ordbok started with a lengthy phase of material collection and basic material collation. Real editing started in 1947 and had by 2001 reached into the letter h (in volume 4 out of 12 planned volumes). The progress rate through the alphabet had then slowed down steadily, while lexicographical treatment grew more and more detailed. Around 2000, at the then rate of progress, Norsk Ordbok could look towards a final publication date for the last volume after 2060, which could be read as another way of saying that the dictionary would never be finished.

Project refinancing, revision and terms

A project with the aim of completing Norsk Ordbok by 2014, in time for the bicentennial celebrations of the Norwegian Constitution, was started in 2002, with financing guaranteed by the Norwegian Storting (Parlia.ment) and by the University of Oslo.

The project is called Norsk Ordbok 2014 with the abbreviation NO 2014.

The project plan is based on the following conditions (from the funders)

1. NO 2014 must be completely digitised (materials, tools, manuscript)
2. Editorial methods and rules for NO 2014 must be revised to fit (a) digitisation, (b) training a large number of new editors in a very short time, without any loss of academic standard or research quality in the dictionary manuscript.

3. NO 2014 must prove itself useful to linguistic research beyond the purposes of dictionary itself, and it must fit into the larger strategies for research and academic development at the University of Oslo.
4. All digitised materials and research results developed within NO 2014 must be made generally available to the public as soon as possible within the course of production.

The only way of meeting these demands was to edit the the dictionary directly into a relational database, from which an xml file could be generated and modified to produce a correct print page. The chosen software was Oracle, already used in digitising the main language collections of NO 2014.

Critical philological review combined with computational analysis

The present paper deals with how these demands were met through a long term intensive cooperation between the NO 2014 and the Unit for Digital Documentation (EDD) at the Arts Faculty, University of Oslo. The following deserne particular mention: Dr. Christian Emil Ore, Lars Jørgen Tvedt, Dr. Daniel Ridings. Without their inspired commitment, NO 2014 would not stand where it is now.

In this cooperation, the major component was a detailed analysis of editorial practice in volume 1 - 4 by the senior editors (working from inside knowledge of the Norsk Ordbok tradition) and by key staff members at EDD (extracting structure by developing a parser for volume 1- 4, and forcing analysis and discussion of each entry component by programming for maximum data integrity).

This process will be illustrated by four case sketches:

1. Entry structure (linearity of running text versus tree structure)

The formal body structure of an entry in volume 1-4 was supposed to have four levels of sense units, marked by (a) upper case letter in bold (only shown if more than one), (b) Arabic numeral in bold, (c) lower case letter in bold:

A

    1

        a

There was also the possibility of using (d) Arabic numerals for ordering the meanings of polysemous idioms within a sense unit: 1,2,3 …

In addition, the following markers were used as separators within the sense unit:

// (double slash) for sub-definitions and for fixed phrases and idioms followed by their own definitions,

/ (single slash) for quotations (followed by source references and comments)

; (semicolon) for a part of a definition with a different shade of meaning.

It was intended that elements should be ordered so that double slash marked a stronger division than slash, which again marked a stronger division than semicolon. However, the complex material, set against insufficient editorial rules, left a wide field for individual judgment and improvisation.

The manuscript parsing performed by EDD gave this result for these separators:

Separator can be followed by	**Double slash //**	**Single slash /**	**Semicolon ;**
New definition within sense unit	X	X	x
Idiom (with one or more (numbered) definitions (1, 2, 3))	X	X	x
Quotation or editorial example with comment	X	X	
Sub-definition	X	X	
Introduction to idiom(s)	X		x
Introduction to list of compounds with headword as final element	X		x
Quotation or editorial example without comment		X	x
Introductory comment to quotation or editoral example (mostly style marker)		X	
Cross reference			x
Part of comment after quotation			x
Part of definition			x
Part of definition after idiom			x
Variant information after idiom			x
Etymological information after idiom			x

In short, (a) the entry format was more finely graded than provided for by the editorial rules, (b) all separators had multiple uses in order to cover all needs (c) the descriptive elements of the entry were to some extent created to meet the complexities of the material at hand.

Further, the parsing showed that the marking up of an entry structure to a considerable extent was **relational**, i.e. determined by the relative weight of materials for that particular entry, and not by general criteria for different linguistic categories. The result was a fluid presentation which often read well, but was lacking in hierarchy and consistency above entry level.

With the evidence from the manuscript parsing on the table it was easy to agree on wanting (a) a full revision of field system and entry structure, (b) restraints on the entry structure which ensure an open tree structure, maximum four levels and the necessary restraints to ensure consistency, (c) explicit editorial rules for each entry component.

The result is the new DWS **sense unit**, constructed from (sets of) interlinked tables. An entry body can have an unlimited number of sense units in the **A1a-**structure, but only the sense unit at the end of a tree branch can exploit the format to the full.

The sense unit format now has four major components in fixed order:

1. Main definition followed by examples
2. One or more sub-definitions followed by examples
3. One or more sub-entries for lexicalised phrases
4. (Illustrative) compounds where the headword is the first or the last element.

The real innovation is nr 3, the sub-entry for lexicalised phrases, which in turn has forced us to deal systematically with phraseology as a sub-discipline of linguistics. This development has been pushed forward by the creation of a (so far) ca 20 million word corpus in addition to our older collections.

2. Multiple use of materials and fuzzy documentation

A historic and documentary academic dictionary depends on its use of sources, not only in terms of documentation but in terms of consistency. An important part of NO source materials is word collections from Norwegian dialects, from ca 1600 until today, in manuscript form and as printed books. Another important sub-set of sources are written accounts of tradition and country life, often written in dialect-marked, non-standard language. Finally, NO 2014 also has a large collection of transcribed dialect words from our own informants. For these, and for other unquestionable dialect items in our collections, only the place of origin is given as a source.

The original editorial rules for handling literary versus geographical sources were not stringent enough for the growing collections, and also practice changed over time. Further, a general shortage of coverage for many words could tempt editors to over-exploit sources, by f. i. listing an 17th century dialect form both as a historic form and as a speech form, or by using a quotation from non-standard language to show a change in the written standard. Nynorsk is a young written language, standardised through consecutive reforms from 1848 until 1981, and that influence from spoken Norwegian on the written standard is considered legitimate (Vikør 2001: 104).

The parsing process (by EDD) revealed multiple and inconsistent use of sources in volume 1-4, especially within the categories older versus newer sources and standard language versus rendering of speech. We needed to create a system that would (a) prevent wrong use of sources, (b) save time for new editors unfamiliar with the language collections.

As a result, a strict classification of all source materials was carried out, where each source was classified for age, genre and use within NO 2014. A database containing a reference bibliography of more than 5000 works for the UiO language collections existed before 2001. This database has been used to mark up each source according to its classification, and it is linked directly to the various source fields in the DWS application. The bibliographical classification is then used to extract specialised sub-bibliographies for f.i. etymological sources, historical sources, dialect sources etc, expressed in the DWS application as fixed menus, so that mistaken use of sources to a large extent is precluded.

Through our dictionary administrative system, the bibliography database is also used to advise editors on whether a word deserves an entry. If f. i. all sources for a word are (bilingual or special) dictionaries, or a word is shown to be a literary hapax legomenon, editing is not recommended.

Our current experience is that the internal control system offered through the bibliography database is popular with the editors because it saves them a lot of time and effort. Getting to know the sources used to be a long and slow process, and consistency in handling sources is hard to achieve. The integration of the bibliography database into the DWS speeds up editing and prevents mistakes.

An important section of NO 2014 written sources consists of dictionaries covering local or regional speech from after 1900, i.e. dialect dictionaries or glossaries. Information from these dictionaries can be listed with a reference to the place where the word is used, or giving the book itself as a source, depending on the category of information used. If this system should prove too complex, it can be tightened up through the bibliography database, but before we do that, we want to see that there is a problem that needs solving.

3. The purpose and logic of cross referencing

Historic and documentary dictionaries are often rich in cross references. Although the practice of explicit cross referencing was well controlled in NO 2014, the purpose and function of cross referencing had never been clearly defined.

In planning the restructuring of the dictionary format we also found that the database designers saw potential and needs for other types of cross referencing than the traditional ones, some of which have been integrated into the DWS application.

Cross references in NO 2014 – from, to, when and why		
	To point in tree structure in entry, i.e.	
	Head word (show head word and homograph number)	**Sense unit (show head word, homograph number and number of sense unit)**
From cross reference entry	a) from less important to more important standard form (where entry is found) b) from dialect variant to standard form	
From entry head	for irregular paradigms where each form has an entry, from inflected form to entry	
From Etymology table	Point to origin of derivations and constituting elements of compounds	
From definition text		Synonym definitions, hyperonyms
From cross reference field after definition		“compare”
From cross reference field after example		“compare”
From Compound list	Pointer from naked compound to (unprinted) database entry and digitized materials	

This table shows the cross reference system built into the DWS application. The "From" column to the left is the place where the cross reference appears. The column titles to the right say what points in the tree structure of an entry you can cross reference to, and for each type it is briefly indicated why this type is included.

Cross references now constitute direct links between entries and sense units in the database. Each entry and sense unit has its own id number. Traditional cross referencing of the "compare" type is minimised. Instead, we encourage editors to put more effort into writing good definitions. Cross referencing is now used primarily to secure that (a) etymological relations are clearly stated, (b) idioms and fixed phrases are defined under only one headword, (c) defining vocabulary is itself defined, and circular definitions avoided, (d) compounds which have to be excluded from the printed.dictionary are linked to their digital entries, which in turn link with the digitised collections that NO 2014 rest on.

The cross references covered by type c, i.e. implicit cross references embedded in definition text, are particularly important in an academic dictionary covering both dialectal variation and a written standard. A case in point is the range of names for common plants, all of which are defined by the official botanical name. The cross referencing system in the database shows cross referencing both ways – at the top of an entry's the tree structure , one can look at a list of entries that contain a cross reference to the entry in question, and thus get a view of f. i. all dialect names of a plant, a bird, together with its official name.

We also see a tendency among editors to use this function of implicit (invisible) cross referencing on the key word, the hyperonym of a definition. This is not something insisted on at present, but it is possible, and it can be inserted at any time. It is logical to use a dictionary lik our to build semantic hierarchies. This is one way in which the project can become useful in linguistic research beyond lexicography proper.

The cross referencing system is not the easiest part of the application to use. But it provides safeguards against cross referencing to non-existent or unprinted entries or sense units, and once a link is correctly entered it stays in place although the entry structure may be changed at either end. It also carries with it the possibility of overviews and insights that paper based editing fails to provide, and we considerate a constitutional part of our DWS.

4. Direct control of sorted and edited materials from the entry back to the digital archives

Ideally a historic and documentary dictionary like NO 2014 should be generated from below, from an exhaustive system of carefully classified individual items of linguistic information, all from solid and verifiable sources. Further, the language to be described should be fully developed and thoroughly standardised, og of course exhaustively decribed in a huge meta literature from every possible angle.

This is not the case for Nynorsk. The written sources of the language are scanty and diverse, the influence from speech rich and contradictory, the orthography has been revised a number of times, and any real standardisation of the written language can only be looked for after 1945. In order to organise the collections more efficiently and speed up editing, NO 2014 has – together with EDD - created a headword index - the Meta dictionary - in the form of a database to which all electronic sources are linked, lemma by lemma. This database is expandable both as to posts and as to the number of sources that can be linked to it.

A thorough revision of the language collections via the Meta dictionary led to an entry reduction of about 20 % (from 0,7 to 0,55 million), and has proved an essential tool in organising the materials on which the dictionary is based.

The Meta dictionary, as well as our major individual collections, are digitised and freely available on the web, cf. URL below. It is therefore possible for all to check an entry in NO 2014 against available materials and evaluate the product.

It is also possible for editors to go straight from the editing format to the Meta dictionary entry, and look at each item of information as they edit. Once an entry has been generated in the NO 2014 database, a link has been created to the Meta dictionary.

However, No 2014 wants to take care of the work that editors do when they sort materials and structure their entries, and to make this hidden background work visible through the electronic version of the dictionary database (at present available only inhouse). Work on a semantic sorter has been going on for some time and will be implemented in 2005. This semantic sorter will allow linking each quotation and each item of information to its relevant sense unit in the entry. The NO 2014 corpus will be included in this system via the Meta Dictionary.

Conclusion – advantages, pitfalls and points to watch

The process of establishing a digitsl platform for all editorial work with the project NO 2014 has forced the project leadership to look at weaknesses and inconsistencies in the handling of source materials and of editorial practice, and to decide how to handle such problems on a "best practice" methodology. In a word, the digitisation process, combined with revised editorial rules, has forced the creation of a stringent editorial DWS application and more explicit editorial rules, which in turn has resulted in a more lucid and consistent dictionary. Furthermore, current experience suggests that Norsk

Ordbok can be ready in 2014 as planned, in spite of having to train twenty plus editors from scratch in three to four years.

The database system created for NO 2014 makes training of new editors a much easier task. Newly recruited editors become productive after a few months of training, and do not seem to feel daunted by the complexity of the project NO 2014. One third of volume 5 is written by editors who started training in the summer of 2003.

In brief, reworking the format of NO 2014 through passing former practice through the sieve of the DWS database designers has led to:
Clearer delimiting and desvription of linguistic categories
Firmer and more predictable formats
A more consistent and searchable dictionary
A dictonary that is easier to work with
More focus on the job that only properly trained editors can do, i.e. analyse and describe the materials from a linguistic and lexicographical point of view.

The chief pitfall for a project like NO 2014 is to lean back and leave design, solutions and testing to the software designers. One point is that project safety depends on inhouse mastery of the product that has been ordered. That is certainly important – you can't become a good cook if you stay out of the kitchen. But the really important loss would be to miss the intensive and critical overhaul of traditional assumptions and ideas about lexicography, linguistics and the art of categorization, which goes well beyond any individual academic discipline.

Literature

Vikør, Lars S. (2001): The Nordic languages. Their Status and Interrelations. 2. rev. ed. (1. ed. 1993, 2. ed. 1995). Novus, Oslo.
Grønvik, Oddrun (1992): The Earliest Dictionaries of Nynorsk in the Light of Presentday Dictionary Typology. Seventh International Conference of Nordic and General Linguistics 1989. In: The Nordic Languages and Modern Linguistics I-II p xx-xx.

URLs:

Norsk Ordbok 2014: http://no2014.uio.no/tekster/ordboka/index.php

NO2014 nynorskkorpus (tagged and lemmatized): http://folk.uio.no/danielr/tagged-nn-alpha.html

Orthographic Disambiguation in Chinese and Japanese Dictionary Lookup

Jack Halpern
The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001, Japan
jack@cjk.org

Abstract

The orthographical complexity of Chinese, Japanese and Korean (CJK) poses a special challenge to the developers of computational linguistic tools, especially in the area of dictionary lookup and intelligent information retrieval. These difficulties are exacerbated by the lack of a standardized orthography in these languages, especially the highly irregular Japanese orthography. This paper focuses on the typology of Chinese and Japanese orthographic variation, provides a brief analysis of the linguistic issues, and discusses why lexical databases should play a central role in the disambiguation process.

1 Introduction

Various factors contribute to the difficulties of dictionary lookup and information retrieval. To achieve truly "intelligent" retrieval many challenges must be overcome. Some of the major issues include:

1. The lack of a standard orthography. To process the extremely large number of orthographic variants (especially in Japanese) and character forms requires support for advanced IR technologies such as **cross-orthographic searching** (Halpern 2000).
2. The accurate conversion between Simplified Chinese (SC) and Traditional Chinese (TC), a deceptively simple but in fact extremely difficult computational task (Halpern and Kerman 1999).
3. The morphological complexity of Japanese poses a formidable challenge to the development of an accurate morphological analyzer. This performs such operations as canonicalization, *stemming* (removing inflectional endings) and *conflation* (reducing morphological variants to a single form) on the morphemic level.
4. The difficulty of performing accurate word segmentation, especially in Chinese and Japanese which are written without interword spacing. This involves identifying word boundaries by breaking a text stream into meaningful semantic units for dictionary lookup and indexing purposes. Good progress in this area is reported in Emerson (2000) and Yu et al. (2000).
5. Miscellaneous retrieval technologies such as lexeme-based retrieval (e.g. 'take off' + 'jacket' from 'took off his jacket'), identifying syntactic phrases (such as 研究する from 研究をした), synonym expansion, and cross-language information retrieval (CLIR) (Goto et al. 2001).
6. Miscellaneous technical requirements such as transcoding between multiple character sets and encodings, support for Unicode, and input method editors (IME). Most of these issues have been satisfactorily resolved, as reported in Lunde (1999).
7. Proper nouns pose special difficulties for IR tools, as they are extremely numerous, difficult to detect without a lexicon, and have an unstable orthography.
8. Automatic recognition of terms and their variants, a complex topic beyond the scope of this paper. It is described in detail for European languages in Jacquemin (2001), and we are currently investigating it for Chinese and Japanese.

Each of the above is a major issue that deserves a paper in its own right. Here, the focus is on **orthographic disambiguation,** which refers to the detection, normalization and conversion of orthographic variants. This paper summarizes the typology of Chinese and Japanese orthographic variation, briefly analyzes the linguistic issues, and discusses why lexical databases should play a central role in the disambiguation process.

2 Orthographic Variation in Chinese

2.1 One Language, Two Scripts As a result of the postwar language reforms in the PRC, thousands of character forms underwent drastic simplifications (Zongbiao 1986). Chinese written in these simplified forms is called **Simplified Chinese (SC)**. Taiwan, Hong Kong, and most overseas Chinese continue to use the old, complex forms, referred to as **Traditional Chinese (TC).**

The complexity of the Chinese writing system is well known. Some factors contributing to this are the large number of characters in common use, their complex forms, the major differences between TC and SC along various dimensions, the presence of numerous orthographic variants in TC, and others. The numerous variants and the difficulty of converting between SC and TC are of special importance to Chinese IR applications.

2.2 Chinese-to-Chinese Conversion The process of automatically converting SC to/from TC, referred to as **C2C conversion**, is full of complexities and pitfalls. A detailed description of the linguistic issues can be found in Halpern and Kerman (1999), while technical issues related to encoding and character sets are described in Lunde (1999). The conversion can be implemented on three levels in increasing order of sophistication, briefly described below.

2.2.1 Code Conversion The easiest, but most unreliable, way to perform C2C conversion is on a codepoint-to-codepoint basis by looking the source up in a mapping table, such as the one shown below. This is referred to as **code conversion** or **transcoding.** Because of the numerous one-to-many ambiguities (which occur in both the SC-to-TC and the TC-to-SC directions), the rate of conversion failure is unacceptably high.

Table 1. Code Conversion

SC	TC1	TC2	TC3	TC4	Remarks
门	們				one-to-one
汤	湯				one-to-one
发	發	髮			one-to-many
暗	暗	闇			one-to-many
干	幹	乾	干	榦	one-to-many

2.2.2 Orthographic Conversion The next level of sophistication in C2C conversion is referred to as **orthographic conversion,** because the items being converted are orthographic units, rather than codepoints in a character set. That is, they are meaningful linguistic units, especially multi-character lexemes. While code conversion is ambiguous, orthographic conversion gives better results because the orthographic mapping tables enable conversion on the word level.

Table 2. Orthographic Conversion

English	SC	TC1	TC2	Incorrect	Comments
telephone	电话	電話			Unambiguous
We	我们	我們			Unambiguous
Start-off	出发	出發		出髮 齣髮 齣發	one-to-many
Dry	干燥	乾燥		干燥 幹燥 榦燥	one-to-many
	阴干	陰乾	陰干		Depends on context

As can be seen, the ambiguities inherent in code conversion are resolved by using an orthographic mapping table, which avoids false conversions such as shown in the **Incorrect** column. Because of segmentation ambiguities, such conversion must be done with the aid of a morphological analyzer that can break the text stream into meaningful units (Emerson 2000).

2.2.3 Lexemic Conversion A more sophisticated, and far more challenging, approach to C2C conversion is called **lexemic conversion**, which maps SC and TC lexemes that are **semantically,** *not* orthographically, equivalent. For example, SC 信息 (*xìnxī*) 'information' is converted to the semantically equivalent TC 資訊 (*zīxùn*). This is similar to the difference between *lorry* in British English and *truck* in American English.

There are numerous lexemic differences between SC and TC, especially in technical terms and proper nouns, as demonstrated by Tsou (2000). For example, there are more than 10 variants for 'Osama bin Laden.' To complicate matters, the correct TC is sometimes locale-dependent. Lexemic conversion is the most difficult aspect of C2C conversion and can only be done with the help of mapping tables. Table 3 illustrates various patterns of cross-locale lexemic variation.

Table 3. Lexemic Conversion

English	S C	Taiwan TC	Hong Kong TC	Other TC	Incorrect TC (orthographic)
Software	软件	軟體	軟件		軟件
Taxi	出租汽车	計程車	的士	德士	出租汽車
Osama bin Laden	奥萨马本拉登	奧薩瑪賓拉登	奧薩瑪賓拉丹		奧薩馬本拉登
Oahu	瓦胡岛	歐胡島			瓦胡島

2.3 Traditional Chinese Variants Traditional Chinese does not have a stable orthography. There are numerous TC variant forms, and much confusion prevails. To process TC (and to some extent SC) it is necessary to disambiguate these variants using mapping tables (Halpern 2001).

2.3.1 TC Variants in Taiwan and Hong Kong Traditional Chinese dictionaries often disagree on the choice of the standard TC form. TC variants can be classified into various types, as illustrated in Table 4.

Table 4. TC Variants

Var. 1	Var. 2	English	Comment
裏	裡	Inside	100% interchangeable
敎	教	Teach	100% interchangeable
著	着	Particle	variant 2 not in Big5
為	爲	For	variant 2 not in Big5
沉	沈	sink; surname	partially interchangeable
泄	洩	leak; divulge	partially interchangeable

There are various reasons for the existence of TC variants, such as some TC forms are not being available in the Big Five character set, the occasional use of SC forms, and others.

2.3.2 Mainland vs. Taiwanese Variants To a limited extent, the TC forms are used in the PRC for some classical literature, newspapers for overseas Chinese, etc., based on a standard that maps the SC forms (GB 2312-80) to their corresponding TC forms (GB/T 12345-90). However, these mappings do not necessarily agree with those widely used in Taiwan. We will refer to the former as

"Simplified Traditional Chinese" (STC), and to the latter as **"Traditional Traditional Chinese"** (TTC).

Table 5. STC vs. TTC Variants

Pinyin	SC	STC	TTC
Xiàn	线	綫	線
Bēng	绷	綳	繃
cè	厕	厠	廁

3 Orthographic Variation in Japanese

3.1 One Language, Four Scripts The Japanese orthography is highly irregular. Because of the large number of orthographic variants and easily confused homophones, the Japanese writing system is significantly more complex than any other major language, including Chinese. A major factor is the complex interaction of the four scripts used to write Japanese, resulting in countless words that can be written in a variety of often unpredictable ways (Halpern 1990, 2000).

Table 6 shows the orthographic variants of 取り扱い *toriatsukai* 'handling', illustrating a variety of variation patterns.

Table 6. Variants of *toriatsukai*

Toriatsukai	Type of variant
取り扱い	"standard" form
取扱い	okurigana variant
取扱	All kanji
とり扱い	replace kanji with hiragana
取りあつかい	replace kanji with hiragana
とりあつかい	All hiragana

An example of how difficult Japanese IR can be is the proverbial "A hen that lays golden eggs." The "standard" orthography would be 金の卵を産む鶏 (*Kin no tamago wo umu niwatori*). In reality, *tamago* 'egg' has four variants (卵, 玉子, たまご, タマゴ), *niwatori* 'chicken' three (鶏, にわとり, ニワトリ) and *umu* 'to lay' two (産む, 生む), which expands to 24 permutations like 金の卵を生むニワトリ, 金の玉子を産む鶏 etc. As can be easily verified by searching the web, these variants frequently occur in webpages. Clearly, the user has no hope of finding them unless the application supports orthographic disambiguation.

3.2 Okurigana Variants One of the most common types of orthographic variation in Japanese occurs in kana endings, called 送り仮名 *okurigana,* that are attached to a kanji base or stem. Although it is possible to generate some okurigana variants algorithmically, such as nouns (飛出し) derived from verbs (飛出す), on the whole hard-coded tables are required. Because usage is often unpredictable and the variants are numerous, okurigana must play a major role in Japanese orthographic disambiguation.

Table 7. Okurigana Variants

English	Reading	Standard	Variants
publish	*kakiarawasu*	書き表す	書き表わす 書表わす 書表す
Perform	*Okonau*	行う	行なう
handling	*Toriatsukai*	取り扱い	取扱い 取扱

3.3 Cross-Script Orthographic Variants Japanese is written in a mixture of four scripts (Halpern 1990): **kanji** (Chinese characters), two syllabic scripts called **hiragana** and **katakana,** and **romaji** (the Latin alphabet). Orthographic variation across scripts, which should play a major role in Japanese IR, is extremely common and mostly unpredictable, so that the same word can be written in hiragana, katakana or kanji, or even in a mixture of two scripts. Table 8 shows the major cross-script variation patterns in Japanese.

Table 8. Cross-Script Variants

Kanji vs. Hiragana	大勢　おおぜい
Kanji vs. Katakana	硫黄　イオウ
Kanji vs. hiragana vs. katakana	猫　ねこ　ネコ
Katakana vs. hybrid	ワイシャツ　Yシャツ
Kanji vs. katakana vs. hybrid	皮膚　ヒフ　皮フ
Kanji vs. hybrid	彗星　すい星
Hiragana vs. katakana	ぴかぴか　ピカピカ

3.4 Kana Variants Recent years have seen a sharp increase in the use of katakana, a syllabary used mostly to write loanwords. A major annoyance in Japanese IR is that katakana orthography is often irregular; it is quite common for the same word to be written in multiple, unpredictable ways which cannot be generated algorithmically. Hiragana is used mostly to write grammatical elements and some native Japanese words. Although hiragana orthography is generally regular, a small number of irregularities persist. Some of the major types of kana variation are shown in Table 9.

Table 9. Katakana and Hiragana Variants

Type	English	Reading	Standard	Variants
Macron	computer	*konpyuuta* *konpyuutaa*	コンピュータ	コンピューター
Long vowels	maid	*Meedo*	メード	メイド
Multiple kana	team	*Chiimu* *Tiimu*	チーム	ティーム
Traditional	big	*Ookii*	おおきい	おうきい
づ vs. ず	continue	*Tsuzuku*	つづく	つずく

The above is only a brief introduction to the most important types of kana variation. There are various others, including an optional middle dot (*nakaguro)* and small katakana variants (クォ vs. クオ), and the use of traditional (じ vs. ぢ) and historical (い vs. ゐ) kana.

3.5 Miscellaneous Variants There are various other types of orthographic variants in Japanese, which are beyond the scope of this paper. Only a couple of the important ones are mentioned below. A detailed treatment can be found in Halpern (2000).

3.5.1 Kanji Variants Though the Japanese writing system underwent major reforms in the postwar period and the character forms have by now been standardized, there is still a significant number of variants in common use, such as abbreviated forms in contemporary Japanese (才 for 歳 and 巾 for 幅) and traditional forms in proper nouns and classical works (such as 嶋 for 島 and 發 for 発).

3.5.2 Kun Homophones An important factor that contributes to the complexity of the Japanese writing system is the existence of a large number of homophones (words pronounced the same but written differently) and their variable orthography (Halpern 2000). Not only can each kanji have many *kun* readings, but many *kun* words can be written in a bewildering variety of ways. The majority of *kun* homophones are often close or even identical in meaning and thus easily confused, i.e., *noboru* means 'go up' when written 上る but 'climb' when written 登る, while *yawarakai* 'soft' is written 柔らかい or 軟らかい with identical meanings.

4 The Role of Lexical Databases

Because of the irregular orthography of Chinese and Japanese, lexeme-based procedures such as orthographic disambiguation cannot be based on probabilistic methods (e.g. bigramming) alone. Many attempts have been made along these lines, as for example Brill (2001) and Goto et al. (2001), with some claiming performance equivalent to lexicon-based methods, while Kwok (1997) reports good results with only a small lexicon and simple segmentor.

These methods may be satisfactory for pure IR (relevant document retrieval), but for orthographic disambiguation and C2C conversion, Emerson (2000) and others have shown that a robust morphological analyzer capable of processing lexemes, rather than bigrams or *n*-grams, must be supported by a large-scale computational lexicon (even 100,000 entries is much too small).

The CJK Dictionary Institute (CJKI), which specializes in CJK computational lexicography, is engaged in an ongoing research and development effort to compile comprehensive CJK lexical databases (currently about 5.5 million entries), with special emphasis on orthographic disambiguation and proper nouns. Listed below are the principal components useful for intelligent IR tools and orthographic disambiguation.

1. **Chinese to Chinese conversion.** In 1996, CJKI launched a project to investigate C2C conversion issues in-depth, and to build comprehensive mapping tables (now at 1.3 million SC and 1.2 million TC items) whose goal is to achieve near 100% conversion accuracy. These include:
    a. SC-to/from-TC code-level mapping tables
    b. SC-to/from-TC orthographic and lexemic mapping tables for general vocabulary
    c. SC-to/from-TC orthographic mapping tables for proper nouns
    d. Comprehensive SC-to/from-TC orthographic/lexemic mapping tables for technical terminology, especially IT terms
2. **TC orthographhc normalization tables**
    a. TC normalization mapping tables
    b. STC-to/from-TTC character mapping tables
3. **Japanese orthographic variant databases**
    a. A comprehensive database of Japanese orthographic variants
    b. A database of semantically classified homophone groups
    c. Semantically classified synonym groups for synonym expansion (Japanese thesaurus)
    d. An English-Japanese lexicon for CLIR
    e. Rules for identifying unlisted variants

Conclusions

IR tools have become increasingly important to information retrieval in particular and to information technology in general. As we have seen, because of the irregular orthography of Chinese and Japanese, intelligent information retrieval and dictionary lookup require not only sophisticated tools such as morphological analyzers, but also lexical databases fine-tuned to the needs of orthographic disambiguation.

Few if any IR tools perform orthographic disambiguation. For truly "intelligent" IR to become a reality, not only must lexicon-based disambiguation be supported, but such emerging technologies as CLIR, synonym expansion and cross-homophone searching should also be implemented.

We are currently engaged in further developing the lexical resources required for building intelligent Chinese and Japanese information retrieval tools and for supporting accurate segmentation technology.

References

Brill, E. and Kacmarick, G. and Brocket, C. (2001) *Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs.* Microsoft Research, Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.

Emerson, T. (2000) *Segmenting Chinese in Unicode.* Proc. of the 16th International Unicode Conference, Amsterdam

Goto, I., Uratani, N. and Ehara T. (2001) *Cross-Language Information Retrieval of Proper Nouns using Context Information.* NHK Science and Technical Research Laboratories. Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan

Jacquemin, C. (2001) *Spotting and Discovering Terms through Natural Language Processing.* The MIT Press, Cambridge, MA

Halpern, J. (1990) *Outline Of Japanese Writing System.* In "New Japanese-English Character Dictionary*"*, 6th printing, Kenkyusha Ltd., Tokyo, Japan (www.kanji.org/kanji/japanese/writing/outline.htm)

Halpern, J. and Kerman J. (1999) *The Pitfalls and Complexities of Chinese to Chinese Conversion.* Proc. of the Fourteenth International Unicode Conference in Cambridge, MA.

Halpern, J. (2000) *The Challenges of Intelligent Japanese Searching.* Working paper (www.cjk.org/cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan.

Halpern, J. (2001) *Variation in Traditional Chinese Orthography.* Working paper (www.cjk.org/cjk/cjk/reference/chinvar.htm), The CJK Dictionary Institute, Saitama, Japan.

Metaphors and Meanings: a Lexicographical Approach

Patrick Hanks

Brandeis University
415 South St.
Waltham, MA 02454-9110
patrick@cs.brandeis.edu

Berlin-Brandenburg Academy of Sciences[1]
Jägerstr. 22/23
10117 Berlin
hanks@bbaw.de

Abstract

Metaphors and similes present particular challenges not only for cognitive scientists, philosophers, and literary critics, all of whom have contributed to the present somewhat biased state of our knowledge, but also for linguists and lexicographers. How far can – and should – a dictionary go in representing figurative language? This paper offers a radical, corpus-based look at the relationship between literal meanings, metaphors, and similes. It investigates the respective roles of metaphors and similes in language. It proposes an elaboration of the theory of resonance (Black 1979) in the context of the theory of norms and exploitations (Hanks 1994, 2004), and discusses the extent to which a dictionary definition can explain figurative as well as literal uses of terms.

1. Introduction

A large monolingual dictionary is a record of the conventional meanings and uses of the words of a language. Within that broad definition, different interpretations have grown up in different cultures of what counts as a 'meaning', what uses are worth showing, and what a dictionary should say. That is, there are different conventional presentations of the linguistic conventions. Monolingual dictionaries of German and modern Greek, for example, give plentiful information about the phraseology associated with particular words. By contrast, English and (in particular) American dictionaries focus primarily on the meaning of the *concepts* that lie behind the words and have little – sometimes astonishingly little – to say about the phraseology associated with the words themselves. In neither case are the links between meaning and use elaborated in any detail.

[1] This work was supported in part by the Wolfgang Paul Prize awarded to Christiane Fellbaum at the Berlin-Brandenburg Academy of Sciences, as part of the Alexander von Humboldt Foundation's Zukunftinvestitionsprogramm.

Lexicographical attempts to describe phraseology have been bedevilled by at least two problems: 1) a failure (at a deep theoretical level) to distinguish between normal usage and all possible usage, and 2) the use of introspection as a source of evidence. Nowhere are these problems more apparent than in the treatment of conventional metaphors and figurative language.

A product of the first problem is that the normal phraseology of a language becomes buried in a welter of postulated usages that may not actually occur, even though (in some people's theories at least) they might possibly occur. This has been a characteristic of much work in American linguistics during the past fifty years. The picture presented is often more complex than it really is (or perhaps it would be truer to say that the picture is of a large number of fictitious or extremely rare complexities, while some of the real complexities are overlooked). Many linguists now recognize that some form of empirically well-founded statistical, probabilistic, or stochastic model is necessary, a model designed to reflect two simple facts: a) any theory must include some means of distinguishing the normal from the bizarre, and b) some usages are more normal than others. However, statistical, probabilistic, and stochastic models have (so far) had little to say about metaphors and similes.

With respect to the second problem, it is now clear from systematic comparisons of phraseologically oriented pre-corpus dictionaries and the evidence of large corpora that there is something of a mismatch between collocations in use and collocations predicted by introspection.

The study of metaphor is flawed by a similar imbalance. In the 1960s and 70s, metaphor was discussed mainly by philosophers. A pivotal moment in the modern discussion of metaphor was Black's (1962) "interaction model", which was elaborated further in Black (1979). This argued that metaphors have meaning by virtue of interaction between the meaning of a **primary subject** and some salient feature or features in the semantics of a **secondary subject**. Thus, Black argues that in Wallace Stevens's metaphor "A poem is a pheasant", readers interpret the primary subject (*poem*) in terms of some salient feature of the secondary subject (*pheasant*). One or more attributes of a poem (beauty and showy language, perhaps) are interpreted in terms of one or more attributes of a pheasant (e.g. beauty and bright, showy plumage).

Davidson (1978) disputed Black's account and argued that all metaphors are trivially false, like lies, while all similes are trivially true. This leads him to push the understanding of metaphors and similes out of semantics altogether, into pragmatics, where there are no necessary truths.

Since 1980, the discussion has been largely led by cognitive scientists. The role of metaphor as a component of human cognition has been much discussed – e.g. by Lakoff and Johnson (1980), Gibbs (1994), Katz (1998), Glucksberg (2001), Giora (2003), and many others. Correspondingly little attention, however, has been paid to how metaphors are actually used. Thus, to take just one example, Giora (2003) offers a "graded salience hypothesis" as an explanation for how we understand word meanings and metaphors in context, but she does not mention the contrast between social salience (usage) and cognitive salience (meaning or belief). Her focus is on cognition. The focus of empirical linguists such as Sinclair (1985) and Gross (1993) is on usage. They have little or nothing to say about metaphors or cognition; indeed Sinclair (p. c.) denies that metaphor is anything other than a diachronic concept. It is time to bring the two together.

Our starting point is a brief look at an influential theory of metaphor. I then go to examine the distinction between literal and metaphorical meaning, and the distinction between conventional and creative language use. Metaphors are often said to be examples of creativity in language use. To what extent is this true? If it is true, a dictionary can have nothing to say about them. But maybe there are conventions of metaphorical language that a dictionary can and should report?

2. Metaphorical and Literal Meaning

A powerful influence on thinking about metaphor in recent years has been Lakoff and Johnson (1980; hereafter 'L&J'). L&J present invented examples in support of their thesis that "our ordinary conceptual system is fundamentally metaphorical in nature" (p. 3) and that we "structure one experience in terms of another" (p. 77). They argue that metaphors are based on (I would say **exploit**) cognitive concepts that are '**experiential gestalts**'. They distinguish between two kinds of experience: direct experience of the world and indirect experience. Metaphors present indirect experiences. Thus, they see metaphor as a property of the human conceptual system, rather than specifically of language. They do, however, concede that the details of metaphorical conceptual structures may differ from culture to culture. Their direct/indirect distinction can be related to two kinds of convention in language: the literal meaning of a word or phrase, and metaphorical or other exploitation of that meaning. They offer, as a typical **structural metaphor**, ARGUMENT IS WAR (with subsidiary examples such as "Your claims are indefensible" and "He attacked the weak points in my argument"), and as a typical **orientational metaphor**, HAPPY IS UP. They argue that there is an overall systematicity in the organization of metaphorical concepts and that metaphors are grounded in an "experiential gestalts". They also mention the possibility of "cultural gestalts". At least three things need to be said about the work of L&J in the present context:

1. In the first place, the notion **metaphor** can only be meaningful if it is contrasted with something. The obvious contrast is with **literal**. If there are no literal meanings, there can be no metaphors. But what is a literal meaning? L&J do not say much about literal meaning. They do not offer

criteria for distinguishing literal from metaphorical. They do not elaborate on the association, if there is one, between "direct experience" and literal meaning. They do not say why one word or phrase or sense of a word or phrase should be classed as metaphorical and the other as literal. For example, instead of ARGUMENT IS WAR, why not postulate that WAR IS ARGUMENT? Our first step, therefore, will be to propose criteria for distinguishing literal and metaphorical meaning.

2. L&J are ambivalent about the relationship between experiential gestalts and cultural gestalts. They acknowledge that "we do not know very much about the experiential base of metaphor" (p. 19), but nevertheless they simply assume that metaphors *have* an experiential base. Our second step will be to suggest that the base of metaphor lies in the language itself, and that it is unsatisfactory to represent this in terms of an experiential gestalt.
3. L&J say nothing about similes. Although there is a broad area of overlap between the linguistic roles of metaphors and of similes, there are, as we shall see, also significant differences. For one thing, similes have a much more significant role to play in linguistic creativity. I have examine a sizeable body of actual examples of both metaphors and similes in the British National Corpus (BNC), only a small part of which can be presented here, and I draw some tentative conclusions. With reference to both metaphors and similes, a central question will concern distinctions between conventional and creative instances.

3. Criteria for Literalness

Each word in a language has one or more conventional uses. It is a large part of the task of the lexicographer to discover what these conventional uses are and what they mean. Some words also have conventions of exploitation, which can also be accounted for lexicographically. Metaphor is a classic example of one type of exploitation. However, systematicity is generally lacking in lexicographic accounts of metaphor.

It is widely assumed that (perhaps because the notion of literalness is intuitively satisfying) criteria for defining and determining literal meanings are well established and universally accepted. This is not the case.

If a metaphor represents a conventional use of a word or phrase, why should we class it as metaphorical at all? Why not simply say that some expressions have two meaning – two literal meanings, if you like – reserving the term **metaphor** for original creative uses that have no place in a dictionary? In fact, that is exactly what some linguists and many English dictionaries – British and American – do. For example, the verb ***backfire*** is listed in most English dictionaries as having just two meanings: "1. (of an action) to have an unintended and undesirable effect" and "2. (of a car or other vehicle) to make a sudden loud noise as a result of a mistimed explosion in the engine or exhaust system."

I would like to argue – and this is controversial – that there is a good case for classifying the first sense of ***backfire*** as literal and the other as a conventional metaphor. But on what grounds can such a distinction be made? Frequency? Historical priority? Abstract reference rather than concrete? All these criteria may play a role, but none of them are necessary or sufficient to distinguish metaphorical meanings from literal ones.

Frequency

If a word has two senses, is the most frequent sense the literal one? This is clearly not the case with the verb ***backfire.*** In all the general English corpora that I have looked at, the 'unintended effect' sense is approximately twenty times more frequent than the the one involving explosions in the exhaust system of a vehicle, but I do not think that anyone would want to claim that the 'explosion' sense is a metaphor based on the idea of a plan or action backfiring. Either we are dealing with two independent literal senses here or the sense 'unintended effect of a plan or action' sense is the metaphorical one. It is the only candidate.

Historical priority

Is the oldest meaning of a word always the literal one? This criterion works fine for ***backfire***, but not for other words. It would be stretching a point to say that all modern uses of ***awful*** are metaphorical, because the literal meaning is 'full of religious reverence'—though in fact such a claim is sometimes made by traditionalists. These examples are among the many that show that literal meanings change from time to time and that therefore the oldest meaning is not necessarily always the literal one. Anyway, if historical priority is to be the criterion, how far back in time should it go? To Latin? To Ancient Greek and Proto-Germanic? To Indo-European? For literate English speakers up to the 19th century, Latin and Greek were an inseparable part of their literary heritage Metaphorical transfers that resulted in the current meanings of some modern English words took place two or three thousand years ago in Latin and other ancient languages. The English word ***ardent*** means 'enthusiastic or passionate'. It is derived from a Latin word (*ardens*) which originally meant 'burning' or 'on fire'. Are we obliged to say that the literal meaning of English ***ardent*** is 'burning', a sense in which it has never been used in English? That seems at least one step too far in pursuit of literal meaning.

And what about the meaning of ***literal*** itself? The oldest meaning of ***literal*** is 'of or pertaining to letters'. So historical priority as a criterion for literalness would commit us to the proposition that the literal meaning of a word is the meaning of the letters of which it is composed. This is palpable nonsense.

Concrete, not Abstract

If a word has two meanings, one abstract and the other concrete, is the abstract meaning always metaphorical and the concrete meaning literal?

Consider the English noun ***object.*** A *physical object* is by definition concrete, while the *direct object* of a verb (a grammatical term), is equally undeniably abstract. The object of an investigation, in the sense 'goal or purpose', is likewise abstract. Does anyone want to claim that the abstract senses are metaphors based on the notion of a concrete physical object? This seems untenable. Common sense dictates that these three senses of the word at least are to be classified as literal and independent of one another. There is no metaphorical relationship among them. If there is such a relationship, it antedates the evolution of English by many centuries. An extreme historicist analysis might perhaps claim that all noun senses are ultimately of metaphorical origin in Latin, based on *objectus* 'something thrown down in front of someone or something', past participle of the verb *objicere* 'to throw before'. This hardly seems relevant to an understanding of any of the meanings of the modern English word ***object,*** though it is in fact what unites them.

Resonance

The most convincing criterion for metaphoricity is perhaps what Max Black (1979) called **resonance**, which is akin to L&J's notion of **reverberation**. A use of a word or phrase is metaphorical if its interpretation is in some way enhanced by a more basic sense, namely the literal sense. Our understanding of ***backfire*** in the sense of an action having an undesirable effect is enhanced by the image of an undesired explosion in a vehicle's exhaust system. Typically, but not necessarily, the literal sense is more frequent, older, and concrete, but necessarily only the metaphorical sense can resonate. If the literal sense is not more frequent, it must at least not be obsolete.

The notion of resonance needs far more detailed exploration than is possible here. Just four points need to be made in the present context:

1. Resonance is directional. For example, your understanding of what Mr Blunkett wanted to say about what was written about him in certain newspapers is enhanced by his description of it as 'garbage', invoking a resonance with the horrible stinking mess in your rubbish bin. This is a one-way relationship; it

cannot be reversed. Thus, if ***garbage*** is used in the literal sense of 'stuff thrown away', our understanding of it is not enhanced by any kind of resonance with what is written in newspaper articles.

2. Resonance requires that both senses be active in the contemporary language. Garbage in the sense 'nonsense' can resonate with the notion of the contents of a rubbish bin, because that sense is still currently active. But it cannot resonate with the notion of the discarded innards of butchered animals (historically, the oldest sense), because that sense is obsolete.
3. Resonance is a gradable. Some metaphors are more resonant than others, but it is not clear how resonance or metaphoricity can be measured. For one reader (me) "bringing the world down on one's head" is much more resonant than "bringing something to light" or "breaking up with someone." The latter two expressions are highly conventional: for a casual reader their resonance may be activated only lightly or indeed may not be activated at all.
4. Resonance is a potential, not a necessary condition for successful interpretation. One reader may read about someone in financial difficulties struggling to **keep his head above water** and think only in terms of continuing solvency. The image of drowning in real water may not occur at all to such a reader. In that case, there is no metaphoricity. Another reader may visualize a graphic image of a person drowning and think in terms of a ***sea of debts*** (not just ***debts***), and so on. Resonance enhances interpretation, but absence of resonance does not absolutely destroy the possibility of successful interpretation. It merely diminishes it.

None of the four criteria discussed in this section ((a) frequency, (b) historical priority, (c) concreteness, (d) absence of resonance) is sufficient in itself to identify a meaning as literal and distinguish it from a metaphorical or allusive one, but any combination of two or more of them can do it. Thus a car engine backfiring is justifiably classed as the literal sense of ***backfire***, even though it fails criterion (a) – it is much less frequent than the sense of a plan going wrong –, because it satisfies criteria (b), (c), and (d): it is the older sense historically, it is a concrete event rather than an abstract one, and it does not have the potential to resonate. The literal meaning of ***camera*** is an apparatus for taking photographs, even though this sense fails criterion (b), historical priority, because it satisfies criteria (a), (c), and (d).

4. Most Metaphors are Conventions

When we come to study actual usage, we find, rather surprisingly, that newly created metaphors in prose are very few and far between. Metaphors are everywhere, but almost all of them have been seen before. They activate secondary senses of the relevant words and phrases – established senses –, rather than creating new meanings.

To take an example at random, Appendix I reproduces a short article from *The Guardian* newspaper about the background to the resignation in December 2004 of the British Home Secretary, David Blunkett. The story consists of 443 words; there are 22 expressions that may be classified as metaphors (underlined in Appendix I). Dividing the number of metaphors by the number of words in the article, we may say that the article has a **resonance quotient** (RQ) of 22/443 (almost exactly .05). This is neither unusually high nor unusually low. A rather longer article on the same page by Sandra Laville, dealing with the same events, has an RQ of 33/552 (.06). A commentary on the same events by Polly Toynbee on page 16 of the same issue of *The Guardian* has an RQ of 52/560 (.09).

Three points should be noticed about the otherwise unremarkable article in Appendix I and the other articles mentioned.

1. All of the metaphors in them are conventional. None of them are newly created. We have seen all of them before. Some are well established as idiomatic phrases in English, others are secondary senses of existing words.
2. It is often hard to decide whether a particular secondary use of a word is a metaphor at all. If a man and his wife have separated, is it literal or metaphorical to say that they have 'broken up'? The phrase is a perfectly conventional way of referring to the end of a marital partnership—it is almost literal—but it still has the potential to resonate with the concept of breaking or destroying something, in a way that 'separating' and 'moving on' do not have. In Appendix I, the benefit of the doubt is given to metaphoricity, i.e. words and phrases have been underlined if they seemed to the reader (me) to have the **potential** to resonate with another sense of the same word or phrase (the 'literal' sense). This judgement is very subjective, of course: short of asking subjects to report their introspections (a notoriously unreliable research technique), there is no way of telling whether any given resonance is activated or not in the mind of any other reader.
3. Quantity is not the only measure of resonance. Even if resonance is activated, it can differ greatly in quality, though this too is subjective. Contrast two sentences referring to different aspects of the same event:
    a. His sacrifice means he faces a lonely future. (Laville)
    b. Sleeping with the enemy, he fell among the most frivolous rightwing effete scoundrels of the Westminster political scene. (Toynbee)

It could be argued that the resonance of (b) is much higher in quality than that of (a). Toynbee's sentence activates not only a more dramatic metaphorical resonance, but also the resonance of intertextual references ("Sleeping with enemy" is the title of a 1991 film; "He fell among thieves" is a biblical quotation). This is no place to digress into a detailed discussion of intertextuality, but its role in resonance theory – the theory of linguistic-cultural gestalts—cannot be disregarded.

If resonance quotients are measured, we find that different genres and different texts—and indeed different parts of the same text—have vastly different RQs. Some texts—scientific abstracts and financial reports, for example—have a very low RQ with few or no metaphors, idioms, and similes. Some chapters of some novels have a very high RQ.

An even more important point about these three articles is that, with respect to the distinction between conventional and newly coined metaphors, they are typical – typical, that is, of most prose writing. There are in these texts plenty of words and phrases used in secondary senses, which we may or may not class as metaphors, according to taste, and there are plenty of allusions, but there is little or no evidence of original creation of metaphors. A wider search in both literature and corpora shows that the vast majority of metaphors outside of poetry are highly conventional. If they are conventions, they can be reported in dictionaries. But are they?

If this is right, then the evidence does not seem to support the claims of writers such as Ortony (1975) and Katz (1998) that "metaphors may not only be nice, they may in fact be ... intrinsically related to the human ability to invent new concepts" (Katz, p. 21). Metaphors do not normally do this. Possibly, similes do. If I want to describe an unfamiliar object or situation in graphic and easily comprehended language, I may do so in terms of a simile, but I am much less likely to create a metaphor for this purpose. Thus, I might well say, "He was hunched over his desk like a frog". I would be less likely to report the same situation as "A frog was hunched over the desk". This point is considered in more detail in section 6 below.

As readers, we can easily convince ourselves that a metaphor is creative, either because it seems graphic and striking, or because we have not come across it before, or more simply because the very nature of resonance demands interpretative input by the reader, so that even a highly conventional metaphor can generate an impression of creativity. Generally, however, the creativity is all on the part of the reader. The writer has done no more than make use of an existing convention. Creativity on the part of the writer does occur, but it is rare, and it is governed by constraints that have not been properly explored.

5. Conventional Metaphors and Linguistic Gestalts

Some words lend themselves peculiarly well to the creation of metaphorical senses. A characteristic of such words is that they denote something that has a cognitively salient property: the hardness of ***iron***, the coldness of ***ice***, the brightness of the ***sun***, the vastness of the ***sea***, the confusion of a ***jungle***, the barrenness of a ***desert***. Although these are familiar concepts, there is often also something slightly exotic about them. How many of us have actually spent time in a jungle or a desert? Indeed, their exotic quality is a very part of their suitability for metaphorical exploitation. They are not mundane. They stimulate the imagination of the reader or hearer.

Let us examine the metaphorical uses of one such word, ***oasis,*** in a little more detail. First, let us look at the words with which oasis is most associated, in both literal and metaphorical uses. The salience test measures word association statistically by comparing the observed frequency of co-occurrence of a pair of words in a corpus with chance (i.e. the predicted frequency of co-occurrence if the words were randomly distributed, rather than being structured in paragraphs, sentences, and phrases). Table 1 shows the salience scores for the most associated collocates of ***oasis,*** based on BNC.

Collocate	**No. of co-occurrences**	**Salience score**
Desert	13	20.8
Calm	7	14.1
Greenery	3	9.6
Welcome	4	7.9
Green	4	6.6
Tranquillity	2	6.0
Peaceful	3	5.9
Peace	3	5.9
Pleasant	3	5.5

Table 1. Salience scores for the most associated collocates of ***oasis***
(based on WASPS, http://wasps.itri.bton.ac.uk)

Not surprisingly, ***desert*** figures prominently among the statistically significant collocates of ***oasis***. It is found in both literal and metaphorical uses. Rather more surprising as significant collocates are ***calm*** and ***tranquillity.*** The set of collocates associated with a word may be regarded as a **linguistic gestalt** (which may or may not map onto L&J's 'experiential gestalts'). I would like to discuss this in a little more detail, as it lies at the heart of the phenomenon of conventional metaphor. Five points may be singled out.

1. Firstly, ***oasis*** has very positive vibes – **positive semantic prosody**, Sinclair would call it – resulting from collocation with words such as ***calm, tranquillity, peace, pleasant***, and other, statistically less significant but still important collocates, such as ***cool, lush, luxurious, pool, water, trees,*** and ***palm trees.***
2. The positive vibes of the word have nothing to do with the real world. My much-travelled colleague Christiane Fellbaum tells me that oases, in reality, are noisy, smelly, crowded places, full of hooting lorries and bustling people. This reality is at odds with the implications of the collocations of ***oasis*** as a word of English. Regardless of such prosaic realities, the English language persists in classifying oases as green, tranquil, and peaceful. This is an important point, affecting all conventional metaphors and allusive

uses of words. A content word in a language may have a linguistic-cultural gestalt that has nothing to do with either scientific truth or the everyday experience of any individual . L&J come closest to acknowledging this when they say (pp. 180-181), "Our categories of experience and the dimensions out of which they are constructed not only have emerged from our experience but are constantly being tested through ongoing successful functioning by all members of our culture."

3. ***Oasis*** is one of a set of words that are particularly productive of conventional metaphors. Over 40% of uses of ***oasis*** in BNC are figurative, and the majority of these are metaphors rather than similes, idioms, or other tropes. Metaphorical uses are shown in Appendix II.
4. Metaphoricity is **signalled syntagmatically** in a variety of ways. Most noticeably, the preposition 'of' often signals that the oasis in question is metaphorical: e.g. *an oasis of calm, an oasis of tranquillity, an oasis of common sense*. This contrasts with the naming use of 'of', as in *the oasis of Bahriyah, over 200 miles from Cairo*, where of course the meaning is literal. A similar indicator of metaphoricity is the use of an adjective from an unrelated or incompatible domain, e.g. *the corporate jungle, a psychological jungle*. The normal use of ***corporate*** and ***psychological*** is in domains unrelated to the geographical ***jungle***. Other clues to metaphoricity are not so straightforward. For example, a locality, if mentioned, can determine one of two incompatible meanings: *an oasis of calm in the centre of Leeds* can only be metaphorical, while *an oasis in the Libyan desert* can only be literal and is probably the very opposite of calm and tranquil.
5. There is a **cline of allusiveness** in the actual use of the word ***oasis***. At one extreme, it denotes a location in a desert where water and vegetation are found. At the other extreme are cases where the oasis is not a physical location at all, but something abstract, e.g. *an oasis of common sense*. Here, the property of *isolatedness* is exploited. In between are cases where the oasis is indeed a location, but not a location in a literal desert, for example *an oasis in the city centre*, where the desert is metaphorical and ***oasis*** inherits the metaphoricity. In such cases the culturally assigned properties of *calm* and *tranquillity* are exploited.

6. Many Similes are Unconventional

It may well be, as some cognitive scientists tell us, that metaphors structure thought, and it seems equally plausible that similes play a particular role in structuring new thoughts. But there is more to it than structuring thought. Metaphors and similes are part of the structure of *language*, not merely of thought. Both metaphors and similes are deeply embedded in the particular language through which they are expressed as well as in the folk beliefs of its users, governed by conventions that have not been adequately studied. As we have seen, these beliefs may have little to do with scientific theory or everyday reality. When we turn to a corpus analysis of similes, we find that they play a much more active role in making meanings than metaphors, but sometimes in unexpected ways.

Unreal comparisons

Many similes are highly conventional, and present no difficulty for analysis. Consider:

> "A sigh went through him like a wave."

The image is obvious. Waves are familiar to most people, either from going to the seaside or from television. They are part of our everyday experience. Up to this point, L&J's explanation of metaphor (and, by implication, similes) as structuring one experience in terms of another is secure. And so it is with many similes – "he shook her like a rag doll", "a white BMW which looks more like a bathroom cabinet than a car", and so on. See Hanks (forthcoming) for a summary of the most widely used vehicles for similes in BNC and further discussion of similes in general.

However, in another large class of similes, also discussed in Hanks (forthcoming), the primary subject is compared, not to something familiar, but to something that does not exist in the tangible physical universe (*a nightmare, a banshee, a ghost, a witch, a fairy-tale princess*, etc.). Even when the secondary subject of the simile is something perfectly familiar to most readers (*a frog, a lighthouse, a man*), it is sometimes modified in an unexpected way (*like a broiled frog, like a demented lighthouse, like a man with a swarm of bees in his underpants*).

In Charles Dickens's *Pickwick Papers* (1837; Oxford World Classics edn. p. 64), Mr Miller, who has disgraced himself at the card table by revoking, is described as being "*as out of place as a dolphin in a sentry box.*" No human being has ever seen a dolphin in a sentry box, but with a little imagination the image is entirely interpretable. If Dickens had used a metaphor ("Mr Miller was a dolphin in a sentry box"), he would have left his readers stranded in incomprehension. Similes, however far-fetched, invite imaginative co-operation. A simile, however far-fetched, invites the question "In what respect?" But far-fetched metaphors simply create bewilderment. Thus:

> "He was a dolphin in a sentry box."
> READER'S RESPONSE: Eh?
>
> "He was like a dolphin in a sentry box."
> READER'S RESPONSE: In what respect?

And in fact, as so often happens, the writer specifies in the immediate context the respect in which Mr Miller was like "a dolphin in a sentry box", namely being out of place.

A modern writer who makes particularly productive use of similes, often (like Dickens) for comic effect, is Sue Townsend. A few examples from *Adrian Mole and the Weapons of Mass Destruction* (2004) may be cited:

> A tall woman with a face like a pretty pig joined us at the counter. (p. 22)
>
> "Coffee then?" I said. – "Coffee?" she said, as though I had suggested fresh pig's blood. (p. 27)
>
> It [a home-made scone] tasted as if it had been baked in AD 1307 over a fire made of twigs and dried cow dung. (p. 86)

Have you ever seen a pig with a pretty face? (Of course, beauty is in the eye of the beholder, and maybe you have strange tastes …) Have you ever drunk fresh pig's blood? Have you ever tasted a scone baked in AD 1307 over a fire made of twigs and dried cow dung? Of course not. These similes work, not by appealing to experience, but by activating an interaction among culturally and linguistically determined prototypes (strictly speaking, stereotypes).. Similes have the unique ability to activate a synthesis not only among separate beliefs, but even among incompatibles. Metaphors cannot do this.

Syntactic Displacement

A phenomenon that is sometimes (not rarely) found in similes is syntactic displacement. An example is the following:

> He looked like a broiled frog, hunched over his desk, grinning and satisfied.

Taken out of context, this simile may seem uninterpretable, but in context it works fine. How? Firstly, the writer specifies the respects in which "he" was like a frog, namely being hunched (when compared to human beings, the normal posture of a frog appears hunched or crouched) – and grinning (the mouth of a frog is proportionally wider than that of a human being). The writer might also have mentioned (but didn't) large bulging eyes.

But secondly, what is this about "broiled"? It is inconceivable that a frog, when broiled, would grin or look satisfied. Moreover, surely a broiled frog is not part of anyone's direct experience? The adjective ***broiled*** is now mainly an American cookery term, but it survives in occasional British usage referring to an unpleasant pink skin colouring resulting from sunburn. Examination of the wider context in BNC reveals that the primary subject has recently returned from the Canary Islands or some other warm climate. No doubt he was sunburned. So the logical structure of this sentence is something like this:

> He looked broiled and like a frog.

Syntactic displacement of this kind is not uncommon in metaphors. The word demented provides some classic examples. A lighthouse, not having a mind or a personality, cannot possibly suffer from dementia, so the expression "a demented lighthouse" is literally incoherent. It could not be used successfully in a metaphor. And yet we find:

The presence of a single woman in their midst acts like a demented lighthouse, enticing hapless men onto the rocks.

7. A Cultural-Linguistic Gestalt

Appendix III shows a proportionally selected concordance from BNC for ***demented***.

Literally, this is a medical term meaning 'suffering from dementia; losing one's mind', typically in the collocations *demented patients, demented elderly*, and *demented geriatrics*. However, only 21% of uses of this word have the literal sense. 36% of uses of ***demented*** are classified as metaphors. These are applied to human beings in a specified role:

herds of demented idiots vandalising the scenery
a demented nanny
a demented warrior
a person demented with grief
the shrill, demented choir of wailing shells

or to human actions:

a demented grin on his pock-marked face
roaring with demented laughter
my demented request
the bereaved mother's demented scream
jumping into a demented tango

or to animate creatures:

the only sound in the room apart from a demented fly.

A double metaphor may be observed in *the shrill, demented choir of wailing shells:* a metaphor embedded within a metaphor. Here, ***demented*** is applied metaphorically to a word denoting a human group (***choir***), which is itself being used metaphorically.

The remaining uses of ***demented*** – 45% – almost half of all uses of this word in this general, "balanced and representative" corpus – are similes. Examples include:

howling like a demented banshee
I look like a demented barber
the idea of God pursuing a whole family like a demented genealogist
Arthur'll jump up and down like a demented rabbit

Historically, ***demented*** is a derivative of Latin *mens, mentis* 'mind', and originally meant 'having lost one's mind'. It might seem reasonable, therefore, to stipulate that, to be described correctly as 'demented', something must have had a mind in the first place. On this reasoning, inanimate objects cannot be demented. This reasoning is evident in dictionary definitions, but I think it is misplaced. It has not in the least deterred writers from applying ***demented*** to inanimate objects in similes:

> ***ticking over figures like a demented computer***
> ***my script looks like demented knitting***
> ***a single woman in their midst acts like a demented lighthouse***
> ***thrashing plastic like a demented clock spring***
> ***racketing against her ribs like a demented steam train***
> ***the paddle ... thrashing like a demented washing machine***
> ***rising and falling like a demented yo-yo***

Are such uses solecisms? Are they evidence of a meaning change in progress? Or are they evidence that similes permit greater logical licence than other structures? A definitive answer is not possible, but certainly there is evidence of a convention that is ignored by dictionaries and linguists alike. Syntactic displacement is here, and very little reality. This is not a safe place for truth-conditional semanticists.

There is also genre specificity. The metaphors and similes cited above are almost all from fiction texts, evidence of a strong association between ***demented*** and the fiction genre. But sentences in fiction are as much rule-governed as any other kind of writing. If we look carefully, we shall see that there is a coherent set of semantic properties and cultural beliefs associated with the word, which can be reported in a dictionary. The common thread is that the English word ***demented***, in both metaphors and similes, regularly denotes an association with violently energetic, purposeless, and often alarming activity. Dictionaries could say this, but none of them do.

Merriam Webster's Collegiate and *Collins English Dictionary* both define the word as "mad; insane". That's all; end of story. The *American Heritage* and *Random House* dictionaries both say much the same thing, but in two separate definitions, suggesting a slight difference in semantic scope: "**1.** Mentally ill; insane. **2**. Suffering from dementia."

The *Oxford Advanced Learner's Dictionary, 6th Edition,* places the figurative sense first (thus according it literal status), and defines it with surprisingly narrow semantic scope: "*(especially BrE)* behaving in a crazy way because you are extremely upset or worried."

The *New Oxford Dictionary of English* (1998; hereafter 'NODE') made a heroic but unsuccessful attempt to capture the figurative sense. After defining the literal meaning as "suffering from dementia", it continues:

> **informal** driven to behave irrationally due to anger, distress, or excitement.

The NODE definition structure was set up to capture subtle interrelations between literal and metaphorical uses, such as are found with ***demented***, but sadly, here as elsewhere, there is a failure of execution, due in part no doubt to the extraordinary pressures under which that dictionary was compiled. With the benefit of hindsight, the NODE definition could have been written to reflect the linguistic and semantic facts, as follows:

> **demented** ***adjective.*** suffering from dementia: *demented elderly patients.*
>
> **figurative** exhibiting or associated with violently energetic, ill-directed, and often alarming activity: *roaring with demented laughter, ... thrashing like a demented washing machine.*

Inclusion of a simile as a dictionary example is justified when the evidence shows clearly that one of the main conventional uses of the word is to form similes.

8. What should a Dictionary say about Dogs?

Let us now look at how corpus evidence for figurative usage can be used to improve the explanation of a core term in the vocabulary. ***Dog*** is a core term. Some anthropologists maintain that the close symbiotic relationship between humans and dogs has been largely responsible for the evolutionary success of both species. Dogs are good at things that humans are not good at, and vice versa. We may expect, therefore, that if any word has a claim to a central role in human culture and a rich, complex, and deeply embedded cultural gestalt, it will be ***dog***.

Is it useful to define the English word ***dog*** as *Canis familiaris*? The question sounds rhetorical, perhaps, but there are several possible serious answers. Maybe some people really do want to know about the status of dogs in the classification of the animal kingdom by taxonomical zoologists. But we should not imagine that assigning a place in an IS-A hierarchy ("a dog is a canine, is a mammal, is a creature, is a physical object, is an entity ...") is all that needs to be said about a word.

Readers may have other questions about dogs, which a dictionary, as a collective cultural index., could aspire to answer. What is so special about dogs? What do dogs do? What, if anything, does a dog symbolize for humans? What does it mean to be "treated like a dog"? (There is nothing in the semantics of either ***treat*** or ***dog*** to justify the assumption that it means 'to be treated badly', so how do we know that that is what it means?) . What does it mean to "work like a dog"? If someone is compared to a dog with a bone, what is implied? What is the relationship between dogs and humans? What are dogs for, or rather what do humans use them for? What is a mad dog?

There are over 12,000 hits for ***dog*** in BNC – far too many for any human being to analyse, so some form of sampling is necessary. Random sampling is one possibility, but motivated sampling is another. Two kinds of motivated sampling are: 1) measure collocations statistically and evaluate the highest scoring ones; 2) select the similes and evaluate the semantic gestalt as something that other things are compared to. Here, I report on both these strategies.

Table 2 shows the collocations that are most associated (per salience score) with ***dog*** in BNC Anyone who looks at salience tables regularly will see immediately that these salience scores are unusually high, compared with those of some other words, e.g. ***oasis*** (see Table 1 above) or ***hare***. For ***hare***, the only collocates of anything like comparable significance are words denoting other hunted animals (***fox*** 19.8, ***deer*** 17.6, ***rabbit*** 17.5, and ***stag*** 13.2) – and their pursuer, ***hound*** (10.7). This is a comparatively impoverished set. We may conclude that the cognitive gestalt of ***dog*** is richer and stronger than that of ***hare*** in the English language.

The collocations of Table 2 form the basis of the linguistic gestalt for ***dog***. What do dogs do? They bark, sniff, whine, eat (voraciously, as it happens), howl, savage intruders, chase cats, etc. What kinds of dogs are there? Well, in addition to different breeds, there are guide dogs, sniffer dogs, guard dogs, pet dogs, tracker dogs, and so on. What do people do with dogs, well (in Britain at least) it seems that mostly they take them for walks, on a leash or a lead. The linguistic gestalt even points to idioms. It may seems slightly strange that ***wag*** does not occur in the list of verbs of which ***dog*** is a significant subject, since a dog wagging its tail is a familiar concept, but even weirder is that ***dog*** is a significant direct object of ***wag***. This is a direct consequence of the idiom ***the tail wagging the dog,*** denoting a state of affairs in which someone has got the wrong order of priorities.

The use of a word in a comparison or simile directly exploits some aspect of a linguistic gestalt, so to flesh out our linguistic gestalt for ***dog*** we may also look at the similes most associated with the word. These are shown in Appendix IV, sorted by the semantic attributes or inferences based on them.

Putting together the new evidence from significant collocates and from similes, along with old evidence from existing dictionaries, books about dogs, and recent research in anthropology, we may attempt an account of the linguistic-cultural gestalt of ***dog*** in English somewhat as follows.

Table 2: Most Significant Collocates for ***dog*** in BNC

COLLOC.	HITS	MI	COLLOC.	HITS	MI	COLLOC.	HITS	MI
IN PP			AND/OR			MODIFIED BY		
~ on leash	10	11.3	cat	209	31.4	guide	137	29.6
~ on lead	11	11.1	horse	57	14.7	sniffer	19	20.2
			bitch	23	11.8	guard	46	18.4
~ at heel	8	11.2	wolf	16	11.0	pet	31	16.2
~ for walk	44	21.5				tracker	12	15.0
~ for blind …	17	16.2				collie	10	14.7
~ in manger	5	11.0	MODIFIES			shepherd	14	14.3
~ in kennel	5	10.5	~ handler	25	18.0	mongrel	9	14.1
			~ dirt	17	16.5	lap	20	13.3
GOVERNED BY 'OF'			~ warden	30	15.2	puppy	13	12.9
Barking of ~	12	17.9	~ barking	10	14.6	hunting	26	12.8
Pack of ~	24	14.9	~ collar	22	14.5	Alsatian	6	11.7
Breed of ~	17	11,2	~ food	74	14.3			
			~ turd	8	13.1			
LIKE ~	281	2.3	~ shit	13	12.8			
			~ owner	85	12.8			
SUBJECT OF VERB			~ biscuit	16	12.3			
Bark	118	32.7	~ kennel	7	11.4			
Sniff	24	16.6	~ breeder	12	10.6			
Whine	15	14.8						
Eat	50	14.6						
Howl	17	14.6						
Savage	11	13.3						
Chase	29	13.2						
Bite	30	12.6						
Yap	6	11.4						
Bound	15	11.0						
OBJECT OF VERB								
Walk	116	22.1						
Wag	16	14.6						
Train	41	9.5						
Let	50	9.1						

> **dog noun 1.** a domestic animal with four legs and a tail, species *Canis familiaris.*
> •Dogs were probably domesticated from wolves in Mesolithic times; some anthropologists attribute the evolutionary success of the two species (humans and dogs) to their long and close association. There are many different breeds of dog. Some are kept as pets, while others are trained for tasks such as hunting, guarding buildings, finding things by smell, and guiding blind people. Dogs are noted for their acute sense of smell, their loyalty to individual humans, their obedience (even when badly treated), their potential for aggression, and their ability to be tenacious and aggressive. Dogs bark, snarl, and sometimes howl; small dogs yap and whine. Dogs wag their tails when happy, eat food voraciously, and typically gnaw at bones. A ***mad dog*** is one with rabies and is extremely dangerous.

The above may look a bit like an encyclopedia article, but it isn't. Every word in it is linguistically justified, based on a close examination of how the word is used, compared with text books and scientific discussion. The entry could go even further: more details could be added. But the above summarizes all the main points. Much of this information could be got across in other ways, for example by a set of hypertext links to selected actual uses of the word in a corpus or to a table of significant collocates.. The presentation proposed here is designed as a succinct and factual presentation of the gestalt in a form that could go into a one-volume dictionary.

It must be acknowledged that ***dog*** is an extreme case. Probably because of the animal's long symbiotic relationship with humans, the word has an exceptionally rich linguistic-cultural gestalt in English, and no dobut the same is true of its equivalent in other languages. Other entries may need no more than a single word or phrase to elaborate their salient semantic attribute(s) and improve the chances of allusions to them being understood: ***hornets*** have a painful sting, ***hares*** run very fast in a zigzag course when pursued, ***frogs*** are perceived has having a crouching or hunched posture, ***oases*** are perceived as calm and peaceful, ***deserts*** are perceived as barren, ***jungles*** are thought to be tangled and confusing.

The current status of English dictionaries is that salient attributes are mentioned for some words but not always. Systematicity in the presentation of linguistic gestalts is called for. Discrepancies between attributes in the linguistic gestalt and attributes in scientific classification or everyday reality can be catered for by hedges such as 'perceived as' and 'considered to be'.

8. Conclusions

This paper has made the following points:

1. Meanings of many important content words form linguistic gestalts. These are related to cultural gestalts, which in turn may be related to L&J's 'experiential gestalts'.

Linguistic gestalts are independent both of everyday reality and scientific truth. Rather, they represent the conventional beliefs associated with words in a linguistic community.

2. The main features of linguistic gestalts can be identified quite precisely by corpus analysis, and can be (but all too often are not) reported systematically in dictionaries.

3. Linguistic gestalts play a central role in the exploitation of literal meanings in figurative language (metaphors and similes).

4. Most metaphors are conventional. They are classed as metaphors in part because they exploit some aspect or aspects of a word's linguistic gestalt in a way that resonates with its primary, literal sense. Measuring resonance is a challenge for future research in this area.

5. Similes have a particular facility to be used creatively. We have explored some aspect of the mechanics of similes, metaphors, and literal meanings and proposed criteria for each. Both metaphors and similes work by activating the reader's or hearer's imagination in a way that literal meanings do not. This is more true of similes, especially unreal or impossible similes, than of metaphors.

Bibliography

Black, Max. 1962. *Models and Metaphors.* Cornell University Press.

Black, Max. 1979. 'More about metaphors' in A. Ortony (ed.): *Metaphor and Thought* (2nd edition 1993). Cambridge University Press.

Davidson, Donald (1978): 'What Metaphors mean' in *Critical Inquiry*, 5.

Geyken, A. forthcoming (2006). 'Lexicon grammars' in K. Brown (ed.): *Encylopedia of Language and Linguistics,* 2nd edition. Elsevier.

Giora, Rachel. 2003. *On Our Mind.* Oxford University Press.

Gibbs, Raymond W. 1994. *The Poetics of Mind.* Cambridge University Press.

Glucksberg, Sam. 2001. *Understanding Figurative Language.* Oxford University Press.

Gross, Maurice. 1993. 'Constructing lexicon grammars' in B. T. S. Atkins and A. Zampolli (eds.): *Computational Approaches to the Lexicon.* Oxford University Press.

Hanks, Patrick. 1990. 'Evidence and Intuition in Lexicography' in Jerzy Tomaszczyk and Barbara Lewandowska-Tomaszczyk (eds.), *Meaning and Lexicography.* John Benjamins Publishing Company.

Hanks, Patrick. 1994. 'Linguistic Norms and Pragmatic Explanations, or Why Lexicographers need Prototype Theory and Vice Versa' in F. Kiefer, G. Kiss, and J. Pajzs (eds.), *Papers in Computational Lexicography: Complex '94.* Research Institute for Linguistics, Hungarian Academy of Sciences.

Hanks, Patrick. 2004. 'The Syntagmatics of Metaphor' in *International Journal of Lexicography*, 17:3.

Hanks, Patrick. Forthcoming. 'Similes and Sets' in R. Blatná and V. Petkovic (eds.). Festschrift for Professor Čermák.

Katz, Albert N. 1998. 'Figurative language and figurative thought: a review' in A. Katz, C. Cacciari, R. Gibbs, and M

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago University Press.

Ortony, Andrew. 1975. 'Why metaphors are necessary and not just nice' in *Educational Theory* 25.

Sinclair, John. 1985. 'Lexicographic evidence', in R. Ilson (ed.): *Dictionaries, lexicography, and language learning*. Pergamon.

Sweetser, E. 1991. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure.* Cambridge University Press.

Appendix I. Metaphors in a newspaper article

Press coverage 'was garbage'

by Owen Gibson: The Guardian, December 16 2004, p. 3.

David Blunkett last night hit out at the "terrible garbage" written in the newspapers over the past three weeks, which he claimed had increased the pressure on him to resign.

The Sheffield Brightside MP angrily reiterated that he was not responsible for bringing the affair that ultimately led to his resignation to light in the press.

Highlighting the central role of the media in propelling the story towards last night's conclusion, Mr Blunkett also revealed that his ex-wife Ruth had turned down a £50,000 offer from a Sunday newspaper to talk about their relationship. The disclosure that a newspaper had offered his former wife money for her story will surprise few, given that a series of revelations in the media have consistently added to the pressure on Mr Blunkett.

The News of the World first revealed on August 14 that the then home secretary was having an affair with a married woman and the story has snowballed since then, driven by a series of revelations from "friends" and "close sources" in the rival Blunkett and Quinn camps.

Meanwhile, both continued to insist in public that they did not wish to discuss their private lives.

Further disclosures in the Sunday Telegraph and the Daily Mail moved the story out of the private arena. Once allegations that Mr Blunkett fast-tracked a visa application for Ms Quinn's former nanny came to light, the coverage also spread to parts of the media that had hitherto ignored it on the basis that it concerned the politician's private life.

It had been continually speculated that Mr Blunkett was himself the source for the original story, a claim he angrily rebutted last night. "I was not involved and I would not bring the world down on my head or my little boy's head," he said.

He hit out at the "lies" that had been printed about him, particularly relating to claims originally published in the London Evening Standard that he had demanded a DNA test on Ms Quinn's child. Mr Blunkett also paid tribute to his ex-wife Ruth, who is the mother of his three eldest children. She had, he said, been "absolutely superb" throughout his difficulties.

He said she had turned down a £50,000 offer from a Sunday newspaper to talk about the collapse of their marriage. "I am proud to have friends like that, even those I have broken up with, and I am a very lucky man."

And dismissing a story in yesterday's Daily Mail, which claimed that he may have fast-tracked a second visa application for his former lover's nanny, he said: "Today's article was another example of the vitriol that has been poured upon me."

443 words.

22 metaphors, all conventional, including some recognized idioms.

Resonance Quotient: 20

Resonance Quality: Low.

Appendix II. 'oasis': metaphorical uses (BNC)

```
METAPHORS
Stoke Mandeville station is a little oasis; clean and bright and friendly.
the idea of the home. A safe private oasis far from the madding crowd, a spher
d steaks. New Town Hotel. A relaxing oasis for professional and business men.
  that she now regards her job as an oasis in a desert of coping with Harry 's
 son or daughter they love can be an oasis in the desert of their week if they
  the intellectuals, described as an oasis in the midst of this desert of feud
  that went with it had been such an oasis in the alarming wilderness of doing
 counts. Poiana Brasov is a skier 's oasis in Romania just like the small glas
  which she had found a haven was an oasis in the East End 's desert, which co
 ought that St Jude 's garden was an oasis in the wilderness. Sit down before
 at Driffield - which was a pleasant oasis in the East Riding of Yorkshire -
planned opencast site was a pleasant oasis in a decaying industrial landscape.
 CHILDREN have helped create a green oasis in the heart of an industrial pollu
 arable crops, leaving Harwell as an oasis in the middle of the fields. I thin
ld her to watch out: love hurts. The oasis is just a mirage. It might be thril
  pretty creatures who loved my city oasis of colour and scent. Bees : the bum
nute. Rustic appeal: Create your own oasis of rustic charm with an intimate an
 kta, our first place of rest, is an oasis of delight. Four days from the near
ormal gardens, it remains a peaceful oasis of eighteenth-century civilization,
 resort of Portals Nous, there is an oasis of calm, tranquillity and romance,
 mile, Chapel-le-Dale is reached, an oasis of greenery in a bleak landscape. H
aslemere 's Parish Church provide an oasis of colour in a week of otherwise gr
upon Portugal Place which remains an oasis of timeless calm only a few paces f
verpool Canal and the River Aire. An oasis of calm in the centre of Leeds, at
 et en route. Try to reserve a small oasis of quietness and time to relax inst
 of extraordinary beauty, a half-lit oasis of repose amid the vigour of the re
  it. This soft green womb formed an oasis of peace in the chaotic tumble that
  them to the Lorrimores who were an oasis of silence in the chattering mob an
 le can escape the hurly-burly to an oasis of calm and do what the like best:
 Faridabad on Sunday seems a distant oasis of calm. Earlier this week I warned
 ends who kept this small area as an oasis of tradition. The other Emirates up
 le can escape the hurly-burly to an oasis of calm and do what the like best:
 rming setting for their Studio - an oasis of style and inspiration. From here
 tan pleasures yet a world apart. An oasis of tranquillity with dramatically s
 re lay in its midday calm, an urban oasis of greenery and Georgian elegance r
 red by lovely trees and gardens, an oasis of greenery in marked contrast to t
 omes with the brief stopover in the oasis of a job. More so at that time when
   Ken 's wife, Betty, was making an oasis of beauty in the flat fieldscape. B
urtyard with two ancient wells in an oasis of serenity amidst the bustling cit
re there is a delightful garden - an oasis of serenity just 10 minutes from th
arming setting for their Studio # an oasis of style and inspiration. From here
```

```
 merely admiring the garden. Such an oasis of peace in all these bricks # And
nd think only: these last days, this oasis of friendship, it is all over. Ever
lough your wild land. We produced an oasis of civilized industry in this fever
 s still pleasant air, an unexpected oasis of prosperous history, almost a tin
 t empty dining table, with a little oasis of bottles, coffee pot and cheesebo
  to be in chaos # instead, it 's an oasis of calm. You could easily run a hou
  in the heat of the day had made an oasis of quiet shadow, a source of energy
 k immediately or forfeit that small oasis of security she 'd won for herself.
  all of creation was welcome to the oasis of her home. I think that often peo
  stone house somewhere to go, a dry oasis of white solitude, a scrap of deser
 its and now the orderly room was an oasis of gloom. Even mcphee, the Assistan
 e contrast to the outside world, an oasis of peace in a world of battle, le r
ared his tongue and lips to the tiny oasis of moisture, and as he drew erotica
ires that have turned a jail into an oasis of colour. Richard Barnett reports:
 a country previously regarded as an oasis of economic success in east Africa.
  the journalese has it. It 's about oases of control where there should be no
ite, sandy beaches dotted with small oases of civilisation, and we were certai
ic Gardens), one of several splendid oases of green in the city. before crossi
  of a Centreparc. There are welcome oases of common sense, however. In the re
 , window-sills and balconies can be oases of green all year with containers f
al Trust gardens, then, are not just oases of beauty and tranquillity or examp
uty in the spring and summer, little oases of wild flowers." He gave the infor
," she said. These Sundays were the oases of human contact in the desert of m
and not only their feet. These brief oases of super-wealth were a direct resul
ould continue to live in their green oasis on the outskirts of Bradford, West
or secluded sunbathing. A lush green oasis on one side boasts a selection of s
modern housing estate and is a sheer oasis out of the past. The setting might
into the muddy trench which forms an oasis round the stem. Ignoring my presenc
rom the centre of town, it is a cool oasis set in a beautiful, mature garden w
  Black Dog presented a brightly lit oasis suggesting welcome and hospitality.
form music into another transcendent oasis. The world is a treasure house of f
  fireside chairs beckoned as a cosy oasis. Through the first tall window one
M5 yet remote from it, is a peaceful oasis undisturbed by tourists who, in the
 t marsh, now happily salvaged as an oasis within a new housing estate. In oth
that out of place here in this green oasis with its feeling of calm and utter
 virtuosity and clarity is a welcome oasis in the current clouded debate: whit
```

Appendix III. 'demented' (adjective): selective concordance (BNC)

```
LITERAL USES
om symptomatic failure occurred in a demented 81 year old man who presented
nts. On examination in August he was demented and ataxic with generalised m
the number of moderately to severely demented elderly people living in the
 needs of the frail elderly, and the demented elderly in particular, had be
 subjects were institutionalised and demented geriatrics who were too far g
  Owen was caring for her moderately demented husband, and coping well with
 a bearing on the integration of the demented include: i the proportion of
  the river. She was watching a poor demented man, obviously having a bad
 , 18, was grabbed in Bangkok by the demented man, who insisted he wanted
osis for caring for mentally ill and demented patients moves to the communi
ntal records, 1217 of which were for demented patients aged 40 to 64 years
pressed patients find more disturbed demented patients distressing. Geriatr
yburgh) but again the most disturbed demented patients are excluded. d Opti
ole numbers of mildly and moderately demented patients, or that the private
use most likely to benefit (severely demented people living alone and witho
rrent. When an old person becomes so demented that she or he can not be con
hen, as is common, the old person is demented. The experience of daily livi
 segregated with almost all patients demented. The position of # functional
ner # Most clients were too severely demented to have a very full understan
ted to non-demented and the ratio of demented to non-demented residents up
tes arises when an old person who is demented turns night into day, and per
          METAPHORS
```

```
snatched away, amidst " the shrill, demented choir of wailing shells ." So
  or death, and the music from their demented fiddling will drive a man mad
 only sound in the room apart from a demented fly. It had been woken from i
 his hands and, with his eyes on the demented fly, expressed disbelief by t
with his eyes fixed on Gabriel and a demented grin on his pock-marked face,
  Mr Cubbage was by this time almost demented, he had to know Coleen 's dec
   sight used to be that of herds of demented idiots vandalising the scener
 'd said, trying to make him see how demented it was. You have to be in the
 Curtis 's cold voice cut across the demented killer 's ranting. On that oc
  to the door by Garvey roaring with demented laughter, slapping his thighs
 the family and acting in a suitably demented manner; noise effects include
   one observer in court thought him demented) Mathews repeated all the lie
  the unspeakable creep Mathews, the demented McLachlan and the absurd Bott
 5, out Nov 10 . The woman here is a demented nanny who is surreptitiously
 s be discounted as the ravings of a demented old man. Nonetheless, he may
and without a place to stay. Are you demented, or what?" "Do you think an
  parliamentary draftsman was drunk, demented or determined to perpetrate
 controls over the extent to which a demented public servant can make a p
  she 'll say. She 'll go absolutely demented." "Really, Constance, I ca
our. The central role is that of the demented Renata " half-saint, half-w
nders of the world. On hearing of my demented request, the other father, A
or the atmosphere pulsed to the same demented rhythm that thrummed in her
 she broke into a terrible, loud and demented scream. Charlie, Charlie, wh
 never forget the bereaved mother 's demented scream. Tell your cousins wh
rying Circus was broadcast. With its demented self-referentiality, its abr
went on, "I think I drive my sister demented. She says she ca n't think w
textures jumping dramatically into a demented tango, remained the best of
ughter. At times she seemed actually demented. Tell Uncle Ashot you want
  herself and went upstairs, feeling demented. Tony picked her up in his
 great hallmarks is the stare: not a demented warrior eyeball-to-eyeball
ived her daughter 's story; although demented with grief, she was unable
 it not for Mrs Hepwood being nearly demented with anxiety she would never
 allow Mike Patton to take us to his demented world -- a warming thought
  mouth, burst into the room, gave a demented yelp, and rushed at Flossie.
         SIMILES
 ding about the yard, cackling as if demented and taking sudden rushes in
uck up the mountain of rubble like a demented animal, and returned to the
nsell howled then, like some kind of demented animal. When he spoke again,
 ed in her mind. Spittals, like some demented bailiff, would soon be forec
 ise until it started howling like a demented banshee." "Howling?" "Som
 's supposed to be me. I look like a demented barber. Who did this drawing
 abbed a pitchfork in and out like a demented barman trying to get the las
   more like brick houses in which a demented builder is steadily, day and
 beams, abseiling to the rescue like demented camouflaged commandos in the
 clambering from tree to tree like a demented chimp. The journey was not
 ctory with thrashing plastic like a demented clockspring which has diseng
 ain was ticking over figures like a demented computer." "So that was wh
piled high on either side as if some demented creature had clawed the corp
 This often sounds like the rap of a demented DJ: the way she moves has go
anced around like, as he put it. "a demented fairy, asking for a bit." A
  from the insect, and jumped like a demented flea half a trillion times,
  island-hopped the Caribbean like a demented flea. I 've plotted your cou
 were like tiny aircraft following a demented flight-path, having lost the
  God pursuing a whole family like a demented genealogist seemed grossly
 now almost full, reeled like a pale demented ghost. His mind stretched ou
 d forward and glaring at her like a demented goldfish. As a host it was
 slamming against her ribcage like a demented hammer. He stepped closer,
 sloshes around the whitewash like a demented house-painter. The star 's
 ng cowboy boots and grinning like a demented howdy-doody does n't turn a
 writing. My usual script looks like demented knitting, but among my manus
 le woman in their midst acts like a demented lighthouse: enticing hapless
 typical Indian fighter flits like a demented moth around a light bulb, th
 tools and horse tack jangled like a demented musical band, and each time
 to job to domestic chores like some demented nursemaid on speed. Give him
 apier approach to cut but an almost demented obsession with the decade th
whereas a physically fit but totally demented person has little left to gi
 t in and out of his backside like a demented piston. All the while her th
  sounded more like the ravings of a demented priest. Occasionally he 'd b
  Arthur 'll jump up and down like a demented rabbit. And can we erm hire
 e described me as behaving # like a demented Scotsman # I felt I had good
 over the map of south London like a demented snake # BR disposed of its B
 fcase down, Maisie 's fingers, like demented spiders in a bath, ran this
  was bellowing down the line like a demented station announcer. If I had
 , racketing against her ribs like a demented steam train. Her skin was bu
 nmoved. Polly, stop behaving like a demented Tory at Calais # You are the
 . Fifteen pounds # he asked, like a demented ventriloquist. He was paraly
 the paddle portholes thrashing like demented washing machines, the engine
 t pity. Mother went on like someone demented when she found him. It was
 ussion shuffles like dogs sniffing, demented woodpecker noises, thumps li
 aunch was rising and falling like a demented yo-yo . The engine slowed an
```

Appendix IV. 'like a dog': selective concordance (BNC)

THE CONCEPTUAL STEREOTYPE "DOG" IN THE FOLK CULTURE OF ENGLISH

1. DOGS HAVE A SYMBIOTIC RELATIONSHIP WITH HUMANS

1.1 DOGS ARE OBEDIENT AND LOYAL. THEY GO FOR WALKS WITH THEIR MASTERS

```
 place. There he would wait, like a dog grown lonely at its master 's absen
ke kindly to walking to heel like a dog, and prefer to walk alongside you.
 to hands and knees, obeying like a dog, without question. Trusting her. An
 Albert in an expectant way, like a dog waiting to be given a biscuit. I wo
d I had to follow him like a little dog, all the way. And it was rough goin
the office, howling like a bereaved dog. The door slammed behind him. Georg
rouched at his knee like a faithful dog. Another chair was fetched from the
o his sides and stood like a little dog, found out, and admitting " Yes, I
 the ground, crouched like a little dog, sat Orme, one of Watkin the dung-c
for him and Kevin, like an obedient dog, ended up nodding slowly as she tol
ght on after him " like an obedient dog or something " A moment later and h
t you hanging around her like a pet dog" Nahum said to him after the servic
He was too trusting, like a willing dog. Another runner threw up his hands
 has to respond like a well-trained dog to demands shouted at him. The smal
tthew sat there like a well-trained dog, ready to make a given response at
 " Action " And like a well-trained dog I stuck an arm in the air and said
  "It was easy."  "Like taking your dog for a walk. With Arnold I was real
```

1.2 DOGS ARE OFTEN TREATED BADLY BY PEOPLE (AND GENERALLY ARE LOYAL NEVERTHELESS)

```
bogey, and there is nothing like a dog (the only animal which trusts human
eading look in his eyes now, like a dog that knows it 's going to be kicked
ecords, that Rab was treated like a dog. I 'll tell that vet a thing or two
ere you 'd have been treated like a dog. Someone else should have been allo
 civil to me, not treated me like a dog. You go in and take your rights; I
m him as she could, cringing like a dog, her tail clamped down hard, her no
ake me out and shoot me down like a dog, old buddy, I was forgetting. Grab
 find that man, and kill him like a dog." Later the same evening Dorian Gra
s weapon and sat down like a beaten dog. Soon after, Dr Livesey rode away o
 the floor, crouching like a beaten dog. Athelstan knelt beside him. Simon
ad down on my knee, like a penitent dog creeping to its angry master when
im suffering. I feel like a whipped dog," he said, in a rare moment of cand
ar. Poor Rory looked like a whipped dog. Funny, was n't it? Such a big stro
```

2. DOGS BARK, GROWL, SNARL, HOWL

```
 talent: the ability to bark like a dog. He began to do so with a desperat
 ear starts to howl and bark like a dog, and the room shakes with the vibr
 e rare trick such as barking like a dog. In our pragmatic and knowing cent
to amuse everyone by barking like a dog during the screening of a film (fi
ing silly and bloody barking like a dog and miaowing like a cat. Excuse me
t of his voice, from a growl like a dog warning its master that it has a s
down the hill, and he growls like a dog if anyone suggests a car. But the
blazing and Cezanne snarling like a dog and then walking out of Aix with h
e theatre. He wanted to howl like a dog and hear the echoes all around him
d his scrawny neck. I howled like a dog, struggling against my captors, un
the office, howling like a bereaved dog. The door slammed behind him. Geor
howls with boredom like a neglected dog. And though Flaubert aggressively
          3. DOGS EAT VORACIOUSLY. THEY GNAW AT BONES
ead. He ate it in two bites, like a dog, and put me back on the gravestone
 greasy soup. I ate, gulping like a dog, and then changed my clothes. The
ure English breed." Even so, like a dog at the bone, The Times was driven
 eyeless man took hold of it like a dog biting a bone. He pulled me violen
ry. Fox said, "He was like a hungry dog let loose in a butcher 's shop; no
cial camaraderie, he worried like a dog with freshly stolen meat over the
          4. DOGS CAN WORK TIRELESSLY AT A TASK
Jannie" "And I make him work like a dog". "I drive you hard because I dri
hat the devil, he had worked like a dog all day, and it was true, Karl wo
Laverne trots back and forth like a dog that wo n't give up, outside an e
't bear to lose. They 'll be like a dog with a rag, tugging away at it, w
sa 's mind, for some reason, like a dog with a bone, could not let go of
ress on particular policies "like a dog after a bone," an adviser claims.
 then, was she pursuing this like a dog with a bone." They had been throw
and beyond, never giving up, like a dog with a dirty old carcass. Anyway,
          5. DOGS CAN BE AGGRESSIVE
ned and threw himself again, like a dog, going for Tuan 's throat. This
 short cruel chain like a dangerous dog. A postcard showing the Wicklow
 draught chased up the steps like a dog snapping at their heels, and the
 's face. Gaveston stiffened like a dog ready to attack. No, of course we
at people, my dear; you look like a dog that's baring its teeth. Drink up
          6. DOGS FIGHT CATS
 together. We 'd fight like cat and dog. Hell, we still have our spats.
 eyes. And they fought like cat and dog at times. Even that last morning
ood Irina and I fought like cat and dog. In our old age we have found a
e nor there. We fought like cat and dog the whole time we were together
          7. DOGS HAVE A HIGHLY DEVELOPED SENSE OF SMELL
e uplifted to the dawn like a happy dog sniffing at the air, euphoric wit
that they perceive them, not like a dog sniffing its way along a scent pa
 it was a sizing-up process. Like a dog checking another dog 's scent. Sm
nd I still think so." He was like a dog sniffing at a weak scent, dodging
was nosing out the territory like a dog in a new home; no objective in vi
          8. DOGS SHAKE THEMSELVES TO GET DRY
 stood up and shook himself, like a dog emerging from water. He dragged a
 off him, and shakes himself like a dog. The water sprays off him that hi
r John woke, shaking himself like a dog, mouthing the most terrible curse
ain growled and shook itself like a dog coming out of the sea. Reality be
          9. DOGS WAG THEIR TAILS WHEN THEY ARE HAPPY
then they waggle their tails like a dog and then start cleaning their fea
          10. DOGS PANT TO PERSPIRE
he top of Pachamama, panting like a dog in the torrid heat and thin air,
 she came once more, panting like a dog, licking at his face. They rolled
          11. A MAD DOG IS ONE WITH RABIES AND IS VERY DANGEROUS
e stood and glared at me like a mad dog. "What?" he croaked; and just the
g on Isabella Rossellini like a mad dog with velvet, scissors, and mask.
" is generally treated like a rabid dog, something acknowledged as being
 frothing at the mouth like a rabid dog in walking boots with the effort
n he had turned on her like a rabid dog, snapping, destroying, infecting.
```

Extracting collocations and their contexts from corpora

Ulrich Heid, Julia Ritz
Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
Azenbergstraße 12
70174 Stuttgart
{heid, ritzja}@ims.uni-stuttgart.de

Most collocation extraction tools for corpus based lexicography rely on statistical sorting to identify significant cooccurrences of word forms or of lemmas. This gives access to word or lemma pairs, but it does not keep track of any linguistic properties of the combinations extracted. However, morphosyntactic properties, typical combinations of collocations, their insertion into the syntax of a sentence, etc. all merit to be described in a detailed dictionary, and thus there is a need for tools extracting data to support lexicographers in describing such phenomena. We present such a tool in this paper: it is based on a multiparametric analysis of data from a chunked corpus of German. The tool is conceived as a suite of different modules. The chunked text is searched by means of pattern-matching, and both collocation candidates and context parameters in the above mentioned sense are extracted, sentence by sentence. These data are stored in a database. For each collocation type (i.e. combination of base and collocate), all context parameters found in all example sentences can be compared and preferences calculated.

1. Introduction

1.1 Collocations

Our work is based on a lexicographic notion of collocations, as it is advocated, among others, by Hausmann (2004): collocations, according to our working definition, are binary lexically determined word combinations composed of a base and a collocate. The base lexical unit selects the collocate, and the two elements are in a defined syntactic relationship (e.g. verb+object, verb+subject, noun+attributive adjective, etc.). Examples of collocations are *give* + *talk*, *question* + *arise*, or *high* *hopes*, respectively (bases underlined). The two elements may be linked by grammatical words (e.g. prepositions, articles: *zur Sprache bringen* ("mention")); both base and collocate may themselves be multiword expressions (as in *Stein und Bein schwören* (Hausmann (2004) "swear by all that is holy")).

Many collocations are relatively frequent and their cooccurrence may be statistically significant in texts, which is why association measures (such as the Log Likelihood Ratio Test, Dunning (1993), for details: http://www.collocations.de/AM) have been used to extract collocation candidates from corpora; but not all significant word pairs are necessarily collocations.

1.2 Collocations in context

If the objective of collocation extraction from corpora is to provide raw material for the creation of a detailed collocation dictionary, listing the word combinations is necessary, but not sufficient. There are both lexicographic and linguistic reasons for this. From the (user-oriented) viewpoint of lexicography, a collocation dictionary should be a text production dictionary. For this function, it is necessary that the dictionary gives as many and as explicit hints as possible, regarding the syntagmatic use of the items it describes, i.e. the way in which they are syntagmatically inserted into a text (see section 2 for a detailed discussion). In descriptive linguistic terms, it seems that many collocations show quite marked preferences with respect to certain syntagmatic properties: morphosyntactic, syntactic, combinatorial (see the results in section 4). We claim that this type of phenomena must be covered in a dictionary, and we show in the following how data for these phenomena can be extracted from large corpora (section 3).

2. Syntagmatic phenomena with collocations

Many collocations show distributional preferences, i.e. an idiosyncratic, non-predictable distribution with respect to morphosyntactic properties. The following are examples for German noun+adjective and verb+object collocations; we indicate in each item the targeted morphosyntactic dimension and (in parentheses) possible values; we then give German examples (Tutin (2004) describes a similar data collection for French):

- Number (singular/plural):
  –*Hoffnung + sich machen: s. Hoffnungen machen* (pl.) (entertain + hope)
  –*Hilfe + medizinisch: medizinische Hilfe* (sg.) (medical help)
- Case (nomin., genitive, dative, accusative):
  –*Hoffnung + gut: guter Hoffnung (sein)*(gen.) (be pregnant)
  –*gemessen + Schritt: gemessenen Schritts*(gen.) (measured (pace))
- Definiteness (definite, indefinite, null article):
  –*Ende + finden: ein Ende finden* (indef.) (come to an end)
  –*Kraft + treibend: die treibende Kraft* (def.) (driving force)
- Active/passive in Support Verb Constructions:
  –*Hoffnung + enttäuschen: H. wird enttäuscht* (pass.) (destroy + hope)
  –*Angebot + nutzen: Angebot wird genutzt*(pass.) (accept + offer)

Furthermore, collocations which involve nouns (in particular verb+object collocations) need to be syntactically inserted into the sentence. Many of the nouns have their own subcategorisation properties (e.g. *proposal* may take an infinitive, outside as well as within a collocation, such as *make+proposal*). The choice of the determiner used in the collocation (*make a/the proposal ...*) is then dependent, among others, on the subcategorisation of the noun (thus: *make the* proposal *to... INF ...*). In so far as such facts are not fully predictable from grammar, they need to be covered by a good collocation dictionary. A similar example is *take the trouble to + INF* vs. *have {} trouble to + INF*.

Finally, collocations may be combined: two collocations may share a base, as in *einstimmiges Urteil fällen* (= *einstimmiges Urteil* + *Urteil fällen* ("pass an unanimous sentence")). Such combinations of collocations form lexical triples (or even larger groups) of significant occurrence frequency (cf. (Zinsmeister/Heid (2003)) and need to be mentioned, perhaps as additional information, in a detailed collocation description. German examples of high frequency are *scharfe Kritik üben* (= criticize massively), or *klare Absage erteilen* (= reject clearly).

3.Extracting data from corpora for a contextual description of collocations

To extract data from text corpora which would support a detailed description of collocations within their syntagmatic contexts, a word-pair based approach is not sufficient. Not only would it either abstract away from the morphosyntactic properties of the collocation components (if based on lemmas) or dilute statistical data to the point of artificially discarding large amounts of low-frequency data (if based on word forms, cf. Evert/Heid/Spranger 2004), but it would also, obviously, not capture larger sequences (triples, sentence contexts, etc.).

3.1 Requirements

An extraction tool designed to adequately capture the phenomena described above should at least have the following properties:

- syntagmatic orientation: the extraction should cover the whole sentence;

- multiparametric extraction: the tool should be able to extract and store a considerable number of facts derived from the analysis of a sentence (e.g. morphosyntax of the noun group in a verb+object collocation, active/passive of the sentence, presence/absence of a complement clause, etc.); we call this extraction multiparametric (in accordance with Spranger (2004), because the tool is intended to extract, along with the targeted collocations, also an (a priori open) number of context parameters;

- robustness: the tool should be robust enough to cope with complex sentences.

3.2 Extraction architecture

To satisfy the above requirements, an architecture is needed which allows for a flexible combination of different corpus linguistic tools. We propose a stepwise approach which is an instance of the by now well-established general architecture for collocation extraction, as used and described a.o. by Smadja (1993), Heid

(1998), Krenn (2000) or Evert (2004a):

- Preprocessing of corpora: tokenizing, part-of-speech-tagging, possibly (recursive) chunking or robust parsing. This step is supposed to provide a rough syntactic analysis and annotation of lexical items and chunks or phrases at the level of morphosyntax.

- Query-based extraction: regular expressions over the annotated material, relying on all types of annotation. This step should produce not only a word pair, but also a number of attribute/value pairs describing the context: active/passive, singular/plural, etc. Not all attribute/value pairs are necessarily binary: for instance, the tool should also capture lemmas of adjectives modifying the noun in a verb + object collocation (e.g. *einstimmig* in *einstimmiges Urteil fällen*). The results of this extraction should be stored in a table or in a database. Thus, we get more than the syntactically homogeneous word pair types produced by standard collocation extraction tools: we allow for some (controlled) (morpho-)syntactic variation, and we capture this variation in a controlled way. Thus, each occurrence in the corpus of a given collocation has one entry in the data collection, which also lists its context parameters.

- Interpretation: As the data collection is particularized down to the level of specific corpus instances of a given linguistic phenomenon, some abstraction is needed. This is provided by interpretation tools; some such tools provide a statistical interpretation, e.g. of the observed quantative preferences of a given collocation with respect to a binary feature. A calculus for such cases has been published in Evert (2004b). Linguistic interpretation could be, among others, a classification of compounds and derived words in terms of their heads and their morphological models. This latter device would allow us to compare collocational behaviour across word classes (*drink heavily - heavy drinker - heavy drinking*) or to compare collocations of compounds with collocations of their heads (*Pause einlegen - Rauchpause einlegen* (have a (smoking) break), cf. Zinsmeister/Heid (2004)).

  The interpretation tools take the table entries and group them, cluster them, calculate distributions, etc., and feed their results back to additional tables.

Figure 1 schematizes the architecture. It is represented in the form of a pipeline for the sake of simplification, ignoring thus the fact that the interpretation steps may be iterative (e.g. first, a morphological analysis of compound nouns in verb+object-collocations is performed, then statistical preferences are calculated on the basis of items from morphological families).

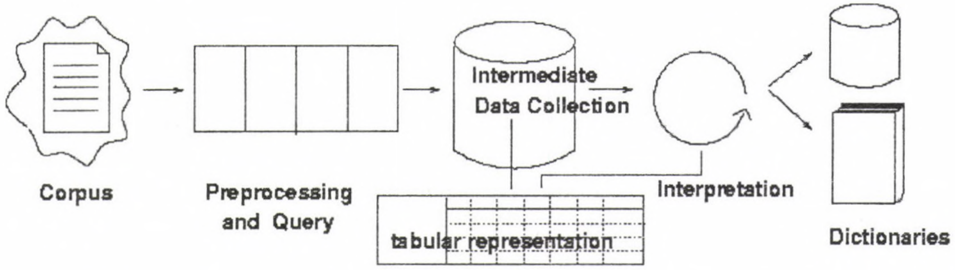


Figure 1: Architecture of an extraction system for collocations in context

3.3 Context-aware extraction of linguistic data

The approach described here is by no means limited to collocations. On the contrary, an extraction and adequate representation and interpretation of context parameters is crucial if we want to get a more detailed picture of many linguistic properties of lexemes. Cases in point are the syntactic and semantic interpretation of nominalizations of verbs (of the *-ation*-type), or syntactic subcategorization.

For the former, Aldinger (2005) proposes a detailed contextual analysis: for each analyzed occurrence of a nominalization in *-ung* of a German verb, she identifies, among others, whether the nominalization is in the singular or in the plural, in a definite or in an indefinite NP, under what kind of grammatical function it is embedded, whether the genitive by which it is followed is in the singular or plural, etc. For syntactic subcategorization, Klotz (2000) has shown the close interrelationship between certain subcategorization patterns and their lexical filling: the collocations *ask + assistance*, *ask + help*, *ask + asylum* clearly select the subcategorization pattern *sb asks (sb) for sth* (*ask for help, etc.*), whereas *ask + opinion* and *ask + permission* prefer the pattern *sb asks sth (from sb)*, (cf. Klotz 2000: 176, on the basis of the *Bank of English* corpus).

These examples seem to indicate that there is a serious need for a type of corpus-based extraction of linguistic data that pays close attention to context parameters, along with the targeted phenomena. These context parameters may be useful as clues for a fine-grained interpretation (as with nominalizations, in Aldinger (2005)), as elements of subregularities which allow us to refine existing classifications of linguistic data, or simply as usage information, for a more detailed lexical description (as with collocations). We see the extraction and interpretation of context parameters alongside the main target of a corpus-based lexical extraction task as a step towards the corpus-based description of the idiomatic nature of language. Where we find marked preferences, these are a sign of idiomatization.

4. Implementation and first results

4.1 Preprocessing and Chunking

As a corpus basis, we used ca. 200 million words from a collection of German newspaper texts of the time frame between 1987 and 1993. As only prenominal participle constructions of the type *der aus dem Amt scheidende Minister* ("the minister who is about to retire from office") have been used to extract collocations of the verb+PP-type, a total of ca. 150.000 candidate sentences have been analyzed. For the preprocessing of the German corpora, we use tokenizing, tagging/lemmatization (Schmid, 1994) and recursive chunking (Kermes 2003, Spranger 2002/2).

A recursive chunker is used, because it covers large amounts of text robustly, providing not only structural and lemma information, but also morphosyntactic feature annotations: at each lexical form, morphosyntactic features are annotated disjunctively, and feature unification is used to provide the adequate information at the heads of chunks or phrases: <np_agr |Gen:M:Pl:Def|> *der Hunde* </np_agr>

4.2 Intermediate data repository

The extraction tool is based on pattern matching within the annotated corpus; all available word annotations (pos, lemma, wordform, full morphosyntactic decoration) and all annotations at chunk and/or phrase level can be used as matching conditions. The query component is specified in the CQP language (which is part of the IMS CorpusWorkbench, cf. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/). It feeds its results into a relational data base which lists all analyzed instances of participle groups along with their context parameters.

For all features with a fixed range of values, we use the algorithm and the R implementation proposed by Evert (2004b) to sum up data from sentences illustrating the same collocation candidate, and to calculate the lower and upper bound of the conservative estimate that indicates preferences. These data lead to a second table which contains absolute figures and preference indications for the following context parameters:

- number of the noun: singular, plural
- case of the noun
- determination of the noun: definite, indefinite, null, demonstrative, quantifier
- modification of the noun: adjective, participle construction, relative clause, genitive-np
- fusion of preposition and article in verb+pp-collocations: yes, no
- type of participle: present, past
- negation: yes/no

The morphological interpretation of complex words will be based on the SMOR two-level morphology system, which handles compounding and derivation (Schmid/Heid/Fitschen 2004).

4.3 First results

The following tables summarize a few results of the multiparametric analysis. For the sake of a convenient presentation, we reproduce, in each table, only the parameters under discussion; for example, table 1 groups combinations with a strong preference for the singular (lower bound of the preference estimate: at least 70%): we consequently indicate the total absolute figure of occurrences (sg) as well as the lower (p_l) and upper (p_u) bound of the singular preference. The collocations discussed in all tables are numbered (first column), for reference; collocations are alphabetized per table, by bases.

	combination type	total	sg	p_l	p_u
1	Abschluß+stehen+vor	29	29	90.19	100.00
2	Ausdruck+bringen+zu	12	12	77.91	100.00
3	Bedrängnis+geraten/raten+in	108	108	97.26	100.00
4	Boden+liegen+an	214	214	98.61	100.00
5	Boden+schießen+aus	17	17	83.84	100.00
6	Boden+stampfen+aus	16	16	82.93	100.00
7	Kraft+setzen+in	89	89	96.69	100.00
8	Kraft+treten+in	933	930	99.17	100.00
9	Mode+kommen+aus	68	68	95.69	100.00
10	Mode+kommen+in	68	68	95.69	100.00
11	Papier+bringen+zu	18	18	84.67	100.00
12	Rate+ziehen+zu	12	12	77.91	100.00
13	Welt+setzen+in	14	14	80.74	100.00

Table 1: Singular, lower bound $(p_l) > 70$

Table 1 contains collocations with a strong preference for the singular (even though most of the nouns could have plural forms, outside the collocation). A comparison of *in Bedrängnis geraten* (line 3) with the almost synonymous *in Schwierigkeiten geraten* (line 19) shows the effect of morphological idiomatization: one expression requires the singular, the other the plural (the verb alternative in 19 is due to a lemmatization ambiguity).

	combination type	total	pl	p_l	p_u
14	Bedürfnis+zuschneiden+auf	16	16	82.93	100.00
15	Depression+leiden+unter	20	18	71.74	98.19
16	Droge+stehen+unter	14	14	80.74	100.00
17	Erwartung+liegen+über	18	18	84.67	100.00
18	Fuge+geraten/raten+aus	42	42	93.12	100.00
19	Schwierigkeit+geraten/raten+in	111	111	97.34	100.00
20	Schwierigkeit+stecken+in	65	65	95.50	100.00

Table 2: Plural, lower bound $(p_l) > 70$

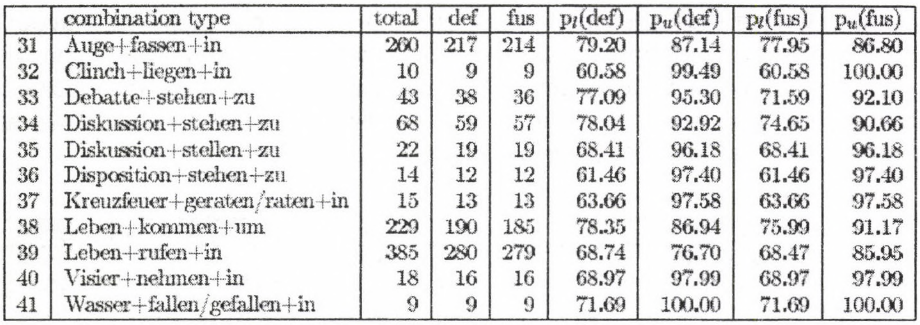
Table 3 shows preferences for a null determiner (typically with singular nouns):

most cases seem to be rather idiomatic. Next to *in Gang kommen* ("be set in motion", singular, null determiner), there is another idiom *in die Gänge kommen* ("get organized", plural, definite article), which is less frequent. Cases of this type (same lexeme, different morphosyntactic preferences, distinguishing two collocations or idioms), which are rather rare, cannot be distinguished by our system.

	combination type	total	nulldet	p_l	p_u
21	Alkoholeinfluß+stehen+unter	22	22	87.27	100.00
22	Alkoholeinwirkung+stehen+unter	10	10	74.11	100.00
23	Angriff+nehmen+in	81	79	92.43	99.56
24	Anspruch+nehmen+in	36	33	79.85	97.69
25	Eis+legen+auf	22	20	74.05	98.36
26	Eis+liegen+auf	32	30	81.61	98.88
27	Frage+kommen+in	202	198	95.53	99.32
28	Gang+bringen+in	23	22	80.98	99.78
29	Gang+kommen+in	103	99	91.33	98.66
30	Gang+setzen+in	53	50	86.02	98.44

Table 3: Null determiner, lower bound (p_l) > 60

Table 4 contains evidence characterized by the simultaneous occurrence of two preferences: the preference for a definite article and for a fused form of preposition and article (*zur, zum, ins, ums*). Examples like *ins Kreuzfeuer geraten* (lit. "come under fire from all sides") show this phenomenon (line 37). Note that this collocation most frequently comes with a genitive complement (e.g. *ins Kreuzfeuer der Kritik geraten*, "be criticized from all sides") forming a combination of collocations.

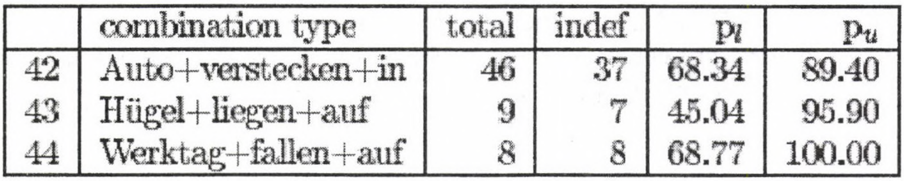


	combination type	total	def	fus	p_l(def)	p_u(def)	p_l(fus)	p_u(fus)
31	Auge+fassen+in	260	217	214	79.20	87.14	77.95	86.80
32	Clinch+liegen+in	10	9	9	60.58	99.49	60.58	100.00
33	Debatte+stehen+zu	43	38	36	77.09	95.30	71.59	92.10
34	Diskussion+stehen+zu	68	59	57	78.04	92.92	74.65	90.66
35	Diskussion+stellen+zu	22	19	19	68.41	96.18	68.41	96.18
36	Disposition+stehen+zu	14	12	12	61.46	97.40	61.46	97.40
37	Kreuzfeuer+geraten/raten+in	15	13	13	63.66	97.58	63.66	97.58
38	Leben+kommen+um	229	190	185	78.35	86.94	75.99	91.17
39	Leben+rufen+in	385	280	279	68.74	76.70	68.47	85.95
40	Visier+nehmen+in	18	16	16	68.97	97.99	68.97	97.99
41	Wasser+fallen/gefallen+in	9	9	9	71.69	100.00	71.69	100.00

Table 4: Definite, fusion, lower bounds (p_l) each > 60

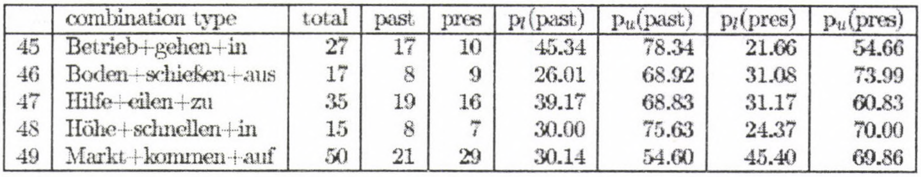
If a preference for a definite article or for a null article is quite widespread in the data, preferences for an indefinite article are not. Table 5 shows the few cases we identified, only one of which is collocational (line 44: *auf einen Werktag fallen*, “fall on a weekday”):

Table 5: Indefinite, lower bound (p_l) > 45



	combination type	total	indef	p_l	p_u
42	Auto+verstecken+in	46	37	68.34	89.40
43	Hügel+liegen+auf	9	7	45.04	95.90
44	Werktag+fallen+auf	8	8	68.77	100.00

As we extracted the collocation data from prenominal participle constructions, only two kinds of participles (present and past) are present in the data. Even with respect to this option, which seems to be unrelated at first sight, most collocations show preferences; table 6 lists the few cases where both kinds of participles are found. Most other collocations do have marked preferences.



	combination type	total	past	pres	p_l(past)	p_u(past)	p_l(pres)	p_u(pres)
45	Betrieb+gehen+in	27	17	10	45.34	78.34	21.66	54.66
46	Boden+schießen+aus	17	8	9	26.01	68.92	31.08	73.99
47	Hilfe+eilen+zu	35	19	16	39.17	68.83	31.17	60.83
48	Höhe+schnellen+in	15	8	7	30.00	75.63	24.37	70.00
49	Markt+kommen+auf	50	21	29	30.14	54.60	45.40	69.86

Table 6: Instances occurring in both past participle and present participle, lower bounds (p_l) each > 20

The results show the degree to which collocations display an uneven distribution of morphosyntactic preferences. For lexicography, we consider it vital to be able to provide lexicographers with data of the kind presented here, as these preferences need to be described in the dictionary. Similarly, the absence of preferences in certain collocations (e.g. *Frage* + *stellen*, "ask + question") is also a lexicographically relevant fact in itself. We expect to be able to signal for each collocation candidate those morphosyntactic properties which are unevenly distributed, so that the lexicographers get easy access to these preferences.

5. Conclusion

In this paper we emphasized the need for keeping track of context parameters of collocations. We presented an architecture for extracting data on morphosyntactic preferences from text corpora, as well as a few first results. These results have not yet been cross-classified with data produced by association measures which identify lexeme combinations as collocations. Thus a few trivial combinations (e.g. lines 42, 43) and a few idiomatic expressions (e.g. lines 25, 26, 41) are found in the data presented. An open question is how idiomatic expressions and collocations can be separated in this approach.

Future work will also address the use of the architecture presented here on other types of sentences, to extract more collocation data than was possible by using prenominal participles. Further research concerns other types of targeted data (nominalizations (Aldinger 2005), subcategorization), and the detection of interdependencies between context parameters. This should provide a solid basis for the provision of data supporting the systematic description of lexico-grammatical subregularities.

References

(Aldinger 2005)
Nadine Aldinger: "Corpus-driven genitive disambiguation", to appear in: *Proceedings of Corpus Linguistics 2005*, Birmingham, 2005.

(Dunning 1993)
Ted Dunning: "Accurate Methods for the Statistics of Surprise and Coincidence". Computational Linguistics, 19/1, 61-74, 1993.

(Evert 2004a)
Stefan Evert: "The statistics of word coocurrences - word pairs and collocations", Diss., Stuttgart, 2004

(Evert 2004b)
Stefan Evert: "The Statistical Analysis of Morphosyntactic Distributions", in: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC), pp. 1539 - 1542

(Evert et al. 2004)
Stefan Evert, Ulrich Heid and Kristina Spranger: "Identifying Morphosyntactic Preferences in Collocations", in: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC), pp. 907 - 910

(Hausmann 2004)
Franz Josef Hausmann: "Was sind eigentlich Kollokationen?" in: Karin Steyer (Ed.): *Wortverbindungen - mehr oder weniger fest* [= Institut für Deutsche Sprache: Jahrbuch 2003], 2004, pp. 309 - 334

(Heid 1998)
Ulrich Heid: "Building a Dictionary of German Support Verb Constructions", in: *Proceedings of the 1st International Conference on Linguistic Resources and Evaluation, Granada, May 1998*, 1998, pp. 69 - 73.

(Kermes 2003)
Hannah Kermes: *Offline (and Online) Text Analysis for Computational Lexicography*, Diss., Stuttgart, (Stuttgart: IMS), AIMS

(Klotz 2000)
Michael Klotz: *Grammatik und Lexik. Studien zur Syntagmatik englischer Verben*, (Tübingen: Stauffenburg Verlag), 2000.

(Krenn 2000)
Brigitte Krenn: *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*, Saarbrücken: DFKI & Universität des Saarlandes, 2000.

(Schmid 1994)
Helmut Schmid: "Probabilistic Part-of-Speech Tagging Using Decision Trees", in: *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44 - 49, Manchester, UK, 1994

(Schmid/Heid/Fitschen 2004)
Helmut Schmid, Ulrich Heid, Arne Fitschen: "SMOR: A German computational morphology covering derivation, compounding, and inflection", in: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC), pp. 1263 – 1266.

(Seretan et al. 2004)
Violeta Serețan, Luka Nerima, Eric Wehrli: "A tool for multi-word collocation extraction and visualization in multilingual corpora", in: *Proceedings of the 11th Euralex International Congress*, (Lorient: UBS), 2004, Vol. 2, 755 - 766

(Smadja 1993)
Frank Smadja: "Retrieving Collocations from Text: Xtract", in: *Computational Linguistics*, Vol. 19, Nr.1, 1993, pp. 143-177 [= Special Issue on Using Large Corpora I]

(Spranger 2002)
Kristina Spranger: *A lexically informed chunking analysis as a starting point for the extraction of linguistic information and terminology from Dutch text*, (Stuttgart: Univ. Stuttgart, IMS), 2002 [= Diploma Thesis], 115pp.

(Spranger 2004)
Kristina Spranger: "Beyond Subcategorization Acquisition - Multi-Parameter Extraction from German Text Corpora", in *Proceedings of the 11th Euralex International Congress*,

(Lorient: UBS), 2004, pp. 171-175

(Tutin 2004)
Agnès Tutin: "Pour une modélisation dynamique des collocations dans les textes", in: *Proceedings of the 11th Euralex International Congress*, (Lorient: UBS), 2004, Vol. 1, pp. 207-221

(Zinsmeister/Heid 2003)
Heike Zinsmeister, Ulrich Heid: "Significant Triples: Adjective+Noun+Verb Combinations", in: *Proceedings of Complex 2003, Budapest*, 2003.

Lexical vs. Dictionary Databases

Design Choices of the MorDebe System

Maarten Janssen
Instituto de Linguística Teórica e Computacional
Rua Conde de Redondo, 74-5 – Lisboa, Portugal
maarten@janssenweb.net

Many lexical databases are modelled simply as digital version of paper dictionaries. However, for many purposes the demands on a lexical database are different from those on a dictionary database. Therefore, the MorDebe database system deviates from the design of dictionary databases in a number of important ways. Firstly, it puts different restrictions on the inclusion of words due to its lesser restrictions in size. Secondly, it does not list only lemmas, but complete inflectional paradigms. And thirdly, lemma separation is form-based rather than meaning-based. This article discusses the advantages and problems of this different approach.

1. Introduction

Over the last few decades, a large amount of new lexical resources have arisen: machine readable dictionaries, lexical databases, full-form lexicons, morphological databases, semantic networks, dictionary databases, etc. Most of these lexical systems have been modelled after lexicographic sources. This paper discusses the design of a lexical database system called *MorDebe*, and why its design differs in important respects for the traditional set-up of dictionaries and dictionary databases.

The term *dictionary database* will be used in this article for a database whose primary function is the compilation of lexicographic products. This can either be simply a digital version of a paper dictionary, from which the printed version is generated (often called a machine readable dictionary), or it can be a complex system from which a wide range of monolingual and bilingual dictionaries are derived, as is the case with the Van Dale Lexicographic Information System (VLIS).

A lexical database, on the other hand, is a lexical resource system meant primarily for computational exploitation. This can be the use in a search engine providing human users with lexical information, but also the use in NLP applications, computer aided language-learning systems, computer aided linguistic research, etc. The lexical database system described in this article is called *MorDebe,* a system which aims explicitly at the use of a single set of lexical data in a wide range of applications, including both NLP systems and human consultation.

This article will discuss the main points in which the MorDebe database differs from dictionary databases. This comparison will be strictly from the perspective of the formal properties – with three main sources of difference: the amount and type of information stored for each lemma, the number of lemmas recorded, and the separation of lemmas. The discussion will focus not only on the motivations for these differences, but also on the resulting problems. Although a strict separation is not possible, the semantic properties of lexical databases are largely ignored in this article.

2. MorDebe Set-up

MorDebe is a lexical database system, whose set-up is largely language-independent, but whose content is currently purely Portuguese. In its current version, MorDebe only specifies formal properties of words - the semantic component has not yet been developed. In the long run, MorDebe is intended to be integrated with the multilingual SIM*u*LLDA system (Janssen, 2002), providing formal, semantic, and cross-linguistic information.

The core of the MorDebe database consists of two related tables: the first is a table containing lemmas, defining for each lemma its citation form, its grammatical category, and when applicable, its compositional structure, and its terminological domain. The second is a table containing word-forms, specifying for each word-form its orthography, the lemma it belongs to, it inflectional form (number, gender, person, tense, aspect, etc.), and when available its syllabification and pronunciation.

The design of MorDebe aims at reusing the same set of data in a wide range of (linguistic) applications. To that end, the design is as much as possible theory and application independent. There are currently a number of ways in which the MorDebe database is used, including a part-of-speech tagger, and the analysis of derivational forms. Of these, two are particularly relevant for the current article:

MorDebe on-line consultation
The most direct use of the MorDebe database is its on-line consultation: there is a internet page that allows users to consult the MorDebe database, either via the lemma database, giving the stored information and the complete inflection, or via the word-form, giving the lemma it belongs to, as well as its inflectional form, the information about the lemma, etc. (more on this in section 3.2)

NeoTrack: neologism detection
The MorDebe database is used to generate the exclusion lexicon used in the semiautomatic detection of neologisms in on-line newspapers. The database is used bidirectionally: not only for the detection of neologism candidates, but new words encountered are also added to the database (more on this in section 4.1).

3. Lemma scope: full-form vs. lemma-only lexicons

Traditionally, dictionaries consist of (printed) lists of words, represented by their citation form, which the user can browse through. In some dictionaries, the key inflectional forms are represented along with the lemma of the entry, at least in the case of irregular inflection. But the dictionary is not commonly intended to be a complete source of inflectional information.

MorDebe, on the other hand, contains the full set of inflectional forms. This information is crucial to the lexical database, whereas it is largely irrelevant for traditional dictionary purposes. Other than what are usually called *full-form lexicons*, MorDebe does not merely list word-forms: all information is organised around lemmas. But for each lemma, its full inflection is provided.

3.1. Word-form driven database access

A full-form lexicon is clearly necessary for NLP use: the computer has to be told all inflected forms explicitly. But even for human consultation, there is a clear advantage of a full-form lexicon: database access is often most conveniently accessed by word-form, and not by lemma. Although it is common to look up words in dictionaries by their citation form, this is not always the most user-friendly solution. Especially for non-native users, it is sometimes hard to find words if the citation form is unknown – the word *mice* is hard to find if you do not know it is the irregular plural of *mouse*. And this is worse for prefixing language – for instance, it is hard for non-native users to find perfective verbs in Slavic languages. When trying to find the word produblyrovannyi in a the Oxford Russian-English dictionary, one has to know that it is a form of the verb dublirovatص (to duplicate), located at the other side of the dictionary.

In traditional dictionaries, this problem is often solved by putting irregular or hard-to-find inflections down as entries of their own. For instance, LDOCE lists *was* as a lexical entry, defined as “*1st and 3rd person sing. past tense of* be*”*. But although this solves the problem of retrievability, it is not the most elegant solution: it mixes lemmas and word-forms, and spreads related forms (are, is, was, being, are, am) around the dictionary. And it is has no clear demarcation criteria: should the Portuguese irregular 2nd person negative imperative *oiçais* of the verb *ouvir* (to listen) be included? Or the regular plural *sloegen* of the Dutch irregular past tense *sloeg* of the verb *slaan* (to hit)?

Access to the MorDebe database is primarily via the word-form: the user enters a word (string) he is looking for, and the database displays which form of which lemma it is. If the word appears in more than one inflectional paradigm, MorDebe lists all the lemmas the word belongs to. Since all the word-forms are explicitly listed, there is no difference in treatment between irregular forms, such as *was*, regular forms, such as *walked*, or cases where the inflectional form is identical to the citation form, such as the past tense *beat*.

A drawback of the inclusion of all inflected forms is that the set of word-forms becomes very large very quickly. MorDebe currently contains some 125.000 lemmas, but already slightly over 1,5 million word-forms for Portuguese, and it is growing steadily. This means that the possibility of browsing is virtually eliminated: MorDebe only provides access to the word-forms and lemmas via search queries. And despite the obvious advantages of searchable indexes, browseable lists have their own merits: people have a tendency of going through lists if they do not know exactly what they are looking for.

3.2. Storage vs. Computation

Storing all inflected forms explicitly is is not the most space-efficient way of storage: for the creation of the Portuguese data for MorDebe, a program was developed which generates verbal forms for all Portuguese verbs. In this applet, only the truly irregular forms are stored – all the rest, including transformations, are stored as rules. This applet implicitly contains all verb-forms in Portuguese, and stores them much more efficiently than the MorDebe database.

But although storage is less efficient, retrieval is much faster: to find all word-forms that are spelled as *walked*, a rule-based storage system would have to rely on morphological analysis to determine that it is the past tense of *walk*, whereas MorDebe can simply look up all the matching forms in the database. MorDebe even allows advanced search options, such as giving all word-forms ending on *-ked* or matching the pattern *wal**d*. In a rule-based system, these forms could only be retrieved by explicitly expanding all lemmas to find the matching word-forms.

Rule based system are effectively only useable in lemma-driven approaches, such as the CD version of the Houaiss dictionary (HoCD): when the user looks up a verb, the dictionary provides not only the normal definition, but also the full inflectional paradigm. But the access is always via the lemma. For a word-form driven system like MorDebe, explicit storage is the best solution.

4. Lexical Coverage

One of the main tasks of dictionary editors is lexical selection. Foremost, because the amount of lemmas presented in a dictionary is limited by physical boundaries, making a careful selection of the most relevant lemmas necessary. But also because it has to be assured that all lemmas included are well-established, correct, general language terms.

However, it is well known that the most frequent source of frustration of dictionary users is the absence of a word they are looking for. That is why Oppentocht & Schutz (2003) suggest that it might be useful to provide a much wider coverage in dictionaries, where possibly the words are not even supplied with definitions, since about 85% of all dictionary consultations are for checking spelling and word existence only.

This observation, although made from the perspective of dictionary design, describes much more the set-up of a lexical database than that of dictionary database. In the design of MorDebe, there are no reasons to reject words due to space limitations. And whereas the entries are intended to be adorned with semantic definitions, the lemma list and their definitions are stored separately, implying that it is not necessary for each lemma to have a semantic definition. When only the form and not the meaning is provided for a given lexical entry, it is nonetheless available for checking spelling and existence. Furthermore, the inflectional paradigm can still be provided on-line, making the system work as an orthographic guide. In that sense, MorDebe is exactly what Oppentocht & Schutz sketch as a future possibility.

As an orthographic guide, a lexical database only has real value if all recorded lemmas are thoroughly checked - if new words would be added too easily, the database reduces to an arbitrary list of words - likely to be correct, but not necessarily so. Therefore, new lemmas are only added to MorDebe after careful verification of existence and correctness. It is possible to add dubious words as well (and even incorrectly spelled words) but in MorDebe, this is only done when marking these lemmas explicitly as 'dubious' or 'wrong'. To ensure correctness of the database, MorDebe furthermore stores with each lemma its base of justification - either motivated by its occurrence in reliable dictionaries, or its occurrence in established sources - as well as when it was added and by whom.

Because of the virtual lack of limitations on number of lemmas, there is also a lessened restriction on generality of the term. Terminological words can be added to MorDebe, when explicitly marked as belonging to a specific terminological domain. In the interface, it is possible to restrict the search queries to only a specific domain, or only general language terms.

4.1. Neologisms

The inclusion of neologisms is a particularly difficult question in the design of dictionaries, as for instance described by Agens (1995). On the one hand, users expect new words to be in their dictionaries, on the other hand, the inclusion of new words is a labour and cost intensive process, and there always is a respectable time-lag between the observation of neologisms by lexicographers, and their availability in the written end-product.

This is less so for lexical databases: the MorDebe database was set-up explicitly for the observation and description of neologisms using a web-based utility called NeoTrack (Janssen, *forthcoming*): daily, two major Portuguese newspapers are checked for possibly new words – i.e. words that are not in the MorDebe database. These neologism candidates are manually verified against corpora to verify whether they are real neologisms or already established words – and either added to a neologism database, or to the MorDebe database. Although it stays a labour intensive process, it is much easier to keep a lexical database up-to-date in this fashion than it is for the traditional dictionary database. And there is no delay between the observation of new words and their on-line availability in MorDebe.

5. Lemma Separation

A major issue on which there are differences between dictionary databases and MorDebe is the question of when to put two word-senses under the same lemma, and when to create different lemmas for them. In a dictionary database, lemma separation is always done in such a way to optimise both compactness and information, driven largely by semantic considerations.

The space limitation in dictionaries sometimes even leads to clustering of different lemmas under a single entry in the case of semantically transparent word-senses, as in the case of run-ons. Strictly speaking, run-ons are morphological derivation, listed at the end of a lemma, with only a grammatical category indication, and no semantic explanation, as the entry " ~ **ly** *adv* " at the end of *royal* in the LDOCE dictionary - indicating that *royally* is the adverbial form of *royal* with the expected semantics, or the entry "~**ker**" at the end of *picnic* to indicate that a *picnicker* is someone who has a picnic.

In the case of zero-derivations, this clustering sometimes can go even further. The GDLP lists at the beginning of *beatão* (hypocrite) that it is either an adjective or a noun, as does PetRob for *réflexe* (reflex), clustering different word-classes under a single lexical entry. This clustering is a clear indication that lemma separation in dictionary databases is based primarily on semantic motivations.

5.1. Inflection based lemma separation

The central focus on inflections in MorDebe shifts the perspective on lemma separation - making it much more form-based. In MorDebe, the inflectional forms are seen as an integral part of the lemma. Therefore, two word senses with different inflections cannot be treated under the same lemma. So in MorDebe, there have to be two different lemmas for the verb *to ring,* because depending on its meanings, the past tense is either *rang* (phone) or *ringed* (bird). And there have to be two lexical entries for *band* in Dutch, since its plural can either be *banden* (tyres) or *bands* (bands).

However, within a single inflectional paradigm, alternative forms may occur: the past tense of the Dutch verb *waaien* (to blow) is either *woei* or *waaide*. And the plural of *pixel* (pixel) in Portuguese can either be *pixels* or *pixéis*. The question whether alternative inflectional forms lead to lemma separation is dependent on whether the two variant forms are intersubstitutable in all circumstances.

From an inflection-based perspective, it is clear that words of different word classes can never be listed under the same lemma: different word-classes have different inflectional paradigms. But taken very strictly, inflection based lemma separation goes even further: in its meaning of the celestial body circling the earth, *moon* is a *singulare tantum.* But in its poetic use as a synonym for *month,* or its more general meaning as a satellite body, it is not. And the word *foot* does not have a plural form in its use as a measurement unit. So strictly speaking, there should be two entries for *moon* and *foot*, one with a plural and one without. More extremely, the same would hold for all words that can be used both as mass nouns and count nouns.

This problem becomes even bigger if what Booij (1995) calls *inherent inflections* are taken into account: word-forms which are in a sense 'between' inflection and derivation. Traditionally, nominal gender is seen as inflectional in many Romance languages: the Portuguese word *geradora* is seen as an inflected form of *gerador* (originator). But female forms only exist for animate nouns. Consider the Portuguese word *amarelo* (yellow). In its base meaning as the colour, it does not have a plural, but in its more liberal sense of 'shade of yellow' it does. And when denoting someone with a yellow complexion ('pale person'), it even has a female singular and plural form. A strictly paradigm-based lemma system would require at least three different entries for *amarelo*: one without a plural, one with, and one with a female form as well.

To resolve these undesirable consequences of strict inflection-based lemma separation, MorDebe allows the existence of *semi-defectives*: words that have a defective inflection in some of their meanings. There is only one word *agua* (water) in MorDebe, which has a plural form *aguas*. And the fact that in its mass noun reading this plural cannot be used is seen as a semantic restriction imposed by a specific reading of the word.

It should be observed that these problems with inflection and word-senses can only be ignored in dictionaries because inflection is often not explicitly treated. But for instance the DLPC does explicitly list female forms, and for this reason, it is forced to view *amarelo* as homonymous, listing the colour and the pale person reading as separate entries.

6. Conclusion

Although there are many ways in which the design of lexical databases, or at least the MorDebe database, resembles the design of dictionary databases, this article shows that there are points in which they differ, due to their different purposes. Firstly, where dictionary databases have no real need for inflected word-form, they are crucial to the set-up of a lexical database. Secondly, where the focus in dictionary database is on selection to preserve compactness, consistency, and correctness, the emphasis in lexical databases is more on completeness and coverage. And thirdly, where lemma separation is almost exclusively governed by semantic issues in dictionary databases, it is largely driven by formal considerations in the MorDebe design.

All three of these differences lead to an increased amount of data in the LDB with respect to the dictionary database. This means that where the dictionary can still be browsed, MorDebe can only be used via search queries. On the other hand, where the dictionary can only be accessed via the lemma, MorDebe can be accessed via any word-form.

References

1. Dictionaries

OxRus: Paul Falla (ed.). 2000. *Oxford Russian-English Dictionary*. Oxford: OUP.

LDOCE: Randolph Quirks (ed.). 1987. *Longman Dictionary of Contemporary English, 2nd Edition.* Essex: Longman.

HoCD: Antônio Houaiss (ed.). 2001. *Dicionário Houaiss Eletrônico*. Lisboa, Rio de Janeiro: Círculo de Leitores.

CED: Patrick Hanks (ed.). 1986. *Collins Dictionary of the English Language, 2nd*

Edition. London: Collins.

DLPC: João Malaca Casteleiro (ed.). 2001. *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa*. Lisboa: Verbos.

GDLP: Graciete Teixeira (ed.). 2004. *Grande Dicionário da Lingua Portuguesa*. Porto: Porto Editora.

PetRob: Paul Robert (ed.). 1989. *Le Petit Robert 1*. Paris: le Robert.

2. Other

Agnes, Michael. 1995. "Why It Isn't There: Practical Constraints on the Recording of Neologisms." *Dictionaries: Journal of the Dictionary Society of America*, vol. 16, p. 45 – 50.

Booij, Geert. 1995. Inherent versus contextual inflection and the split morphology hypothesis. *In:* Booij & van Marle (eds.) *Yearbook of Morphology 1995*. Dordrecht: Kluwer.

Janssen, Maarten. 2002. *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua*. PhD Thesis, Utrecht University.

Janssen, Maarten. 2004. Multilingual Lexical Databases, Lexical Gaps, and SIM*u*LLDA. *International Journal of Lexicography*, vol. 17: 137 - 154.

Janssen, Maarten. *forthcoming*. Orthographic Neologisms. Selection criteria and semi-automatic detection. *Submitted to Terminology.*

Oppentocht, Lineke & Rik Schutz. 2003. Developments in Electronic Dictionary Design. *In:* P. van Sterkenburg (ed.) *A Practical Guide to Lexicography.* Amsterdam: John Benjamins Publishing.

Lexicotext, Or How To Use A Statistical Tool In Dictionary Compilation

MARGARETA KASTBERG SJÖBLOM
ILF-CNRS
Bases, Corpus et Langage (UMR 6039)
UFR Lettres, Arts et Sciences humaines
98, Bd Edouard Herriot
06204 NICE Cedex 3, France
kastberg@unice.fr

The purpose of this paper is to explore lexicostatistical methods in dictionary compilation. The lexicostatistical methods and techniques have created new research opportunities for a wide variety of corpus analysis. The application of the statistical tool Hyperbase, which was indeed originally conceived for the treatment of large literary corpora has since been extended to include research on political, historical, commercial and political corpora and even oral data bases. Not only can the program process short texts in a rapid and efficient way, but its maximum output is achieved when exploring large corpora (exceeding one million tokens). The exploration and development of such statistical methods can also be very useful in lexicography, particularly in the study of dictionaries where it opens new avenues of research for a larger audience in the field of lexicography. *Lexicotext* is a specific adaptation of the Hyperbase tool for the treatment of dictionary corpora. The phraseology content of a dictionary, i.e. the sentences and the examples in the context of different words in the dictionary, is indeed a form of closed corpus and could even be regarded as a specific discourse or genre, which, after an adequate data-processing treatment, adapts perfectly to corpus linguistics. The quantitative method, which makes it possible to take into account the totality of the corpus simultaneously, gives a synthetic and impartial view of the language communicated by the dictionary.

Lexicostatistical methods

The lexicostatistical methods and techniques inspired mainly by the lexical statistics methods developed in the 1960's in France, – particularly by Pierre Guiraud and Charles Muller,– have created new research opportunities in a wide variety of corpus analysis.

Pioneer work in this field of research, especially in literary genres like the French classical theatre, developed a technique and work method that today not only applies to various literary corpora, but also defines a technical platform for studies in the general evolution of language; for generic typological analyses; and, even for studies of the political journalistic language. The application of the statistical tool Hyperbase, which was originally conceived for the treatment of large literary corpora, including the works of Balzac, Hugo or Proust, has since been extended to include research on political, historical, commercial and political corpora and even oral data bases such as opinion polls. Not only can the program process short texts in a rapid and efficient way, but its maximum output is achieved exploring large corpora (exceeding one million tokens).

The creation of large textual corpora, their compilation into exploitable data-bases for quantitative treatment, as well as for various linguistic research and stylistics, has been one of the most important developments in corpus linguistics over the last years. Today's techniques, through the integration of lemmatizers into the statistical tool, allow direct access to normalized and tagged textual data, as well as to the indispensable grammatical codes.

As compared with traditional lexicostatistics, the majority of corpora today are finally morpho-syntactically standardized, in order to widen research possibilities and move towards a more complete analysis of various aspects of the language, applying methods of what today in France is called *logometrics*. In fact, the development and the improvement of software tools has overcome earlier objections made against traditional lexical statistics and constitutes an important qualitative achievement in this field. This qualitative progress offers, for the first time, a tool for complete statistical language treatment (from chains of characters to semantic isotopies).

The Hyperbase tool

The Hyperbase tool, conceived and developed by Étienne Brunet and CNRS "Bases, Corpus et Langage" in France, is a complete tool for studies of large corpora of preset texts. Through its different functions the program allows a documentary as well as a statistical approach to the corpus. Documentary functions, such as concordance research, context of a word (or a chain of words) and a collocation function are available, based on the whole corpus or on a selected part. Different frequency dictionaries are provided automatically by the program and a variety of lists are available in a rapid and efficient way (it takes only a few seconds to generate a dictionary from a corpus of several million tokens). The distribution of a word (or a group of words) can be simultaneously studied in the whole corpus and also be visualized thanks to the graphic applications.

The statistical treatment opens up the possibility for a number of different analyses, such as lexical richness, the study of hapax legomena, lexical growth, lexical distance (or connexion), chronological correlation as well as for internal or external specificity studies. The external comparison source for the French language is the large corpus of *Le Trésor de la langue française* (*TLF*), containing 86 million tokens, and allowing the user to choose a certain period of time from the 15th century until today. A new multilingual version of Hyperbase is now under development, which opens up the way to corpus analyses of English, German and Italian corpora. The program will offer the user possibilities to explore normalized and tagged textual data, and explore the British National Corpus as an external comparison source for the English language.

A thematic function makes an inventory of all the terms located in the immediate environment of a given word and according to their statistically expected frequency. The program also makes it possible to extract a semantic pole around a specific word. Different graphic representations such as factor analysis, dendrograms/branching diagrams or tree analyses, give an almost "geographic" overview to the complex and multiple connections or oppositions amongst the different words of a corpus.

The exploration and development of these statistical methods can also be very useful in lexicography, particularly in the study of dictionaries where it opens new avenues of research for a larger audience in the field of lexicography. The creation and management of systematized and readily available dictionary corpora are indisputably very useful complements for the lexicographer, in particular within the development of bilingual dictionaries as well as for more theoretical studies such as comparative studies of dictionaries.

Lexicotext

Lexicotext is a specific adaptation of the Hyperbase tool for the treatment of dictionary corpora. The first version of this hypertextual data-base contains the French nomenclature extracted from a bilingual dictionary, *Norstedts stora fransk-svenska ordbok* (1998), one of the most recent French-Swedish dictionaries. The program enables documentary and statistical analyses of the dictionary word list as well as dictionary phraseology.

The phraseology content of a dictionary, i.e. the sentences and the examples in the context of different words in the dictionary, is indeed a form of closed corpus and could even be regarded as a specific discourse or genre, which, after an adequate data-processing treatment, adapts perfectly to corpus linguistics. The quantitative method, which makes it possible to take into account the totality of the corpus simultaneously, gives a synthetic and impartial view of the language communicated by the dictionary.

The different analyses enable the user to compare, exogenously, different bilingual or monolingual dictionaries, different periods of time etc., and endogenously, to have a global vision of the homogeneity and the consistency of the examples in the different articles, their size, variety and diversity or, to the contrary, to have a precise view of the recurrent themes conveyed by the dictionary.

The selection of the nomenclature and the examples, which appear in the dictionary, is mainly founded on the linguistic awareness of the lexicographers and on their knowledge of the audience they are addressing. The choice of words and sentences to include in the dictionary is inevitably subjective, and often reflects the lexicographer's own conception of society and personal perception of reality.

However, a corpus made out of the examples in a dictionary is not only a very culturally marked text, but also a strong reflection of a certain period of time or an époque.

A vocabulary is the expression of a society. As Georges Matoré (Matoré: 1953) already stated: "Les mots ne tombent pas du ciel, ils apparaissent à leur heure, et la date de leur naissance est intéressante dans la mesure où elle révèle une modification survenue dans l'histoire d'une civilisation."[1]

At certain intervals in the history of a society a new vocabulary appears (the old one of course being maintained in its broad outlines), which reflects new social, political, economic, aesthetic conditions etc., i.e. changes of human conditions, and, consequently other words disappear. And while such changes do not take place abruptly, and the majority of the "users" of a language might not realize what changes are taking place in the vocabulary, the effects of these changes on the operational efficiency of the bilingual dictionary are important for the dictionary user.

–Is there a way to evaluate the French distributed in a bilingual dictionary? –How does one define this portrait of the French language, this window towards the foreigner, which it indisputably constitutes? This paper proposes a method for the evaluation of the French language in a French-Swedish contemporary dictionary, applying statistical and numerical techniques.

[1] "Words do not fall down from the sky, they appear at their given time, and their date of birth is interesting as it reveals a modification/change, which has occurred in the history of a civilization."

Application to a French-Swedish Dictionary

The database explored in this paper includes the sentences and the syntagms of the French part of the dictionary and the corpus counts 159.263 tokens distributed over the 26 letters of the alphabet, here being used as a reference base. This traditional reference base allows firstly, a global view of the quantitative structure of the dictionary, and secondly, the identification of the relative importance given to different letters within the dictionary or in comparison with other dictionaries.

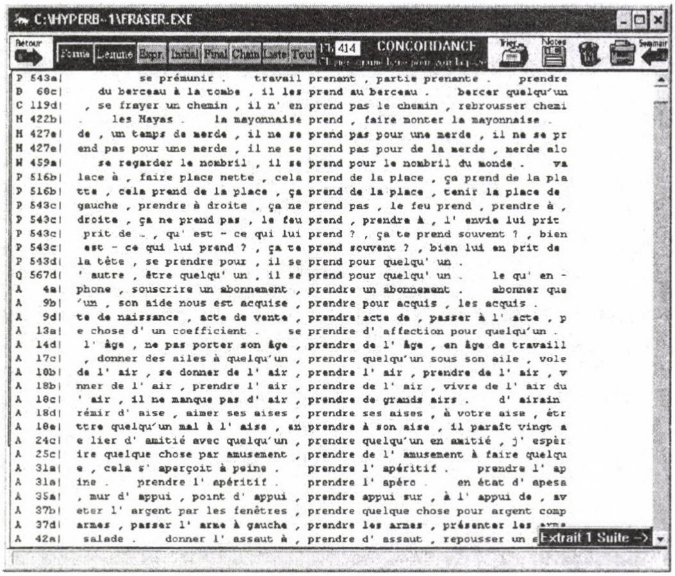
The analysis of the high frequency distribution, i.e. the relatively most significant words within the corpus, makes it possible to distinguish the particular features of the corpus.

As an example, in the dictionary, it is noteworthy that among the 100 most frequent words there is not one single noun or adjective, words that one would find on top of the list investigating any other literary, political or journalistic corpus.

rang	frq	mot	rang	frq	mot	rang	frq	mot
1	4002	à	11	1157	être	21	458	ne
2	3051	la	12	1022	une	22	437	que
3	3036	un	13	948	des	23	407	son
4	2289	le	14	945	avoir	24	396	par
5	2018	en	15	779	il	25	384	mettre
6	1991	se	16	723	pas	26	366	ça
7	1901	d'	17	667	chose	27	359	pour
8	1730	l'	18	656	quelque	28	336	comme
9	1635	faire	19	647	dans	29	333	prendre
10	1169	les	20	612	sur	30	308	et

In fact, the dictionary is a true verbal sphere and not only the rich frequency of verbs but also the importance of the pronoun reflects that reality. Verbs such as *faire*, *être*, *avoir*, *prendre*, *mettre* and *donner* on the top of the list are not only used as stative verbs for descriptions but is indeed incessant within the dictionary articles.

The list of concordance of the verb – *prendre* (take) – illustrates the large application field of this lemma which appears 333 times in our corpus:



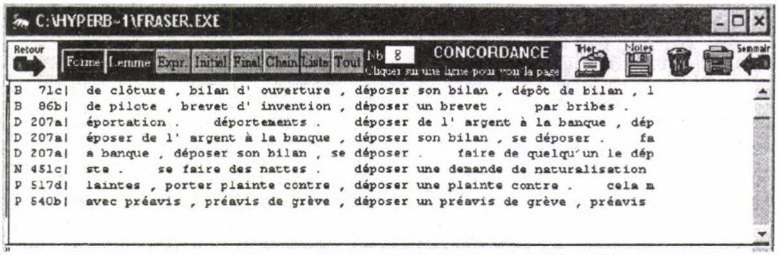
```
P 543a|          se prémunir .   travail prenant , partie prenante .   prendre
B  68c|     du berceau à la tombe , il les prend au berceau .   bercer quelqu'un
C 119d|   , se frayer un chemin , il n' en prend pas le chemin , rebrousser chemi
H 422b|  .   les Mayas .   la mayonnaise prend , faire monter la mayonnaise .
H 427a|  de , un temps de merde , il ne se prend pas pour une merde , il ne se pr
H 427e|  end pas pour une merde , il ne se prend pas pour de la merde , merde alo
N 459a|     se regarder le nombril , il se prend pour le nombril du monde .   va
P 516b|  lace à , faire place nette , cela prend de la place , ça prend de la pla
P 516b|  tte , cela prend de la place , ça prend de la place , tenir la place de
P 543c|  gauche , prendre à droite , ça ne prend pas , le feu prend , prendre à ,
P 543c|  droite , ça ne prend pas , le feu prend , prendre à , l' envie lui prit
P 543c|   prit de .. , qu' est - ce qui lui prend ? , ça te prend souvent ? , bien
P 543c|   est - ce qui lui prend ? , ça te prend souvent ? , bien lui en prit de
P 543d|  la tête , se prendre pour , il se prend pour quelqu' un .
Q 567d|  ' autre , être quelqu' un , il se prend pour quelqu' un .   le qu' en -
A   4a|  phone , souscrire un abonnement , prendre un abonnement .   abonner que
A   9b|  'un , son aide nous est acquise , prendre pour acquis , les acquis .
A   9d|  te de naissance , acte de vente , prendre acte de , passer à l' acte , p
A  13a|  e chose d' un coefficient .   se prendre d' affection pour quelqu'un .
A  14d|   l' âge , ne pas porter son âge , prendre de l' âge , en âge de travaill
A  17c|   , donner des ailes à quelqu'un , prendre quelqu'un sous son aile , vole
A  18b|  de l' air , se donner de l' air , prendre l' air , prendre de l' air , v
A  18b|  nner de l' air , prendre l' air , prendre de l' air , vivre de l' air du
A  18c|  ' air , il ne manque pas d' air , prendre de grands airs .   d' airain
A  18d|  rémir d' aise , aimer ses aises , prendre ses aises , à votre aise , étr
A  18e|  ttre quelqu'un mal à l' aise , en prendre à son aise , il paraît vingt a
A  24c|  e lier d' amitié avec quelqu'un , prendre quelqu'un en amitié , j' espèr
A  25c|  ire quelque chose par amusement , prendre de l' amusement à faire quelqu
A  31a|  e , cela s' aperçoit à peine .   prendre l' apéritif .   prendre l' ap
A  31a|  ine .   prendre l' apéritif .   prendre l' apéro .   en état d' apesa
A  35a|  , mur d' appui , point d' appui , prendre appui sur , à l' appui de , av
A  37b|  eter l' argent par les fenêtres , prendre quelque chose pour argent comp
A  37d|  armes , passer l' arme à gauche , prendre les armes , présenter les arme
A  42a|   salade .   donner l' assaut à , prendre d' assaut , repousser un e
```

Figure 1: Concordance of the lemma prendre – *take*

By investigating the lexical specificity, using the *TLF* corpus as an external reference, the characteristic and significant words or themes of the dictionary corpus are brought out in an impartial way. In the dictionary used in this analysis one easily notices the importance given to words and themes reflecting Swedish society and its tax systems with syntagms such as *annualité de l'impôt* (yearly taxation), *assiette de l'impôt* (tax base), *assujetti à l'impôt* (liable to tax), *assujettissement à l'impôt* (tax liability).

An important question for the lexicographer to revolve is under what dictionary entry a sentence, an idiom or an expression should be listed. Should an expression such as "*it's raining cats and dogs*", for example, be listed under "rain", under "cat", under "dog" or under the three entries? *Lexicotext* will not be able to resolve this delicate question. However, it will help the lexicographer, by providing an immediate and exact inventory of the different lexical compositions.

The program allows an instant hypertextual inquiry of an idiom or a set expression as for example *déposer son bilan* (go into bankruptcy) where the left column shows the different location of the occurrences, in this case under B (bilan) and D (déposer):



```
B  71c|  de clôture , bilan d' ouverture , déposer son bilan , dépôt de bilan , l
B  86b|  de pilote , brevet d' invention , déposer un brevet .    par bribes .
D 207a|  éportation .    déportements .    déposer de l' argent à la banque , dép
D 207a|  époser de l' argent à la banque , déposer son bilan , se déposer .    fa
D 207a|  a banque , déposer son bilan , se déposer .    faire de quelqu'un le dép
N 451c|  ste .    se faire des nattes .    déposer une demande de naturalisation
P 517d|  laintes , porter plainte contre , déposer une plainte contre .    cela m
P 540b|  avec préavis , préavis de grève , déposer un préavis de grève , préavis
```

Figure 2: Concordance of déposer

Moreover, exploring a large statistical database also provides the lexicographer with a wide range of possibilities of context and collocation research in different types of corpus. When the predefined corpus is extended from being a pure dictionary nomenclature corpus to other fields or genres, the program enables the user to extract, in an immediate and extensive way, idioms, expressions or even the semantic extension of a word or concept.

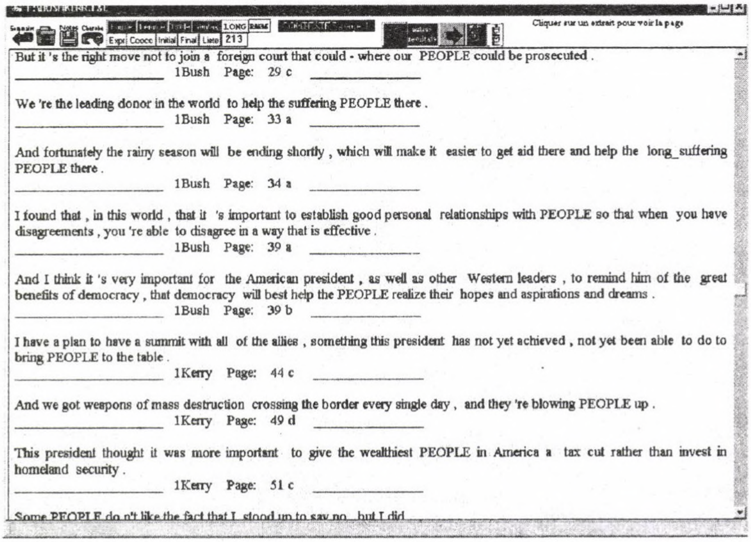
Context analyses

Lexicotext is not limited to include only dictionary corpora, but can contribute to lexicographic research and dictionary compilation by providing the ability to investigate other useful corpora. In fact, it is possible to check up the record of a word or expression in a quick and efficient way by exploring a lager general reference base. The program allows not only concordance inquiries, but also the analysis of a greater inventory than the immediate context of a word by a function called CONTEXTE. This function shows the investigated word in its natural context in fluent text, i.e. the paragraph. Although the dictionary text is of a different nature and not a continuum of fluent text, a broader and more general context inquiry can be very useful for the lexicographer.

The word is in fact far from being autonomous, it does not necessarily appear in a phrase with the same signification as it would have as an independent unit. It is important to analyse the context and not only the linguistic component individually to make a good translation. In fact, larger semantic and thematic analyses give valuable indications to lexicographers. The thematic surroundings, or topic, of a word is to be considered as a semantic macrostructure composed by stable semantic structures linked to each other. However, all semantic patterns rely on a very personal and fragile net of associations.

When simultaneously exploring two different languages, as in a bilingual dictionary, the divergence of word association can be very important. The difference between the two systems is not only linguistic, the divergence should in fact also be observed at a deeper socio-cultural level.

This tool and the context function make it possible to extract more than a concordance or a collocation from a corpus. In this example the external corpus consists of the presentations made by the candidates in last year's presidential election in the USA; George W. Bush and John Kerry. The extract shows a few contexts of the word *people*:



But it 's the right move not to join a foreign court that could - where our PEOPLE could be prosecuted .
1Bush Page: 29 c

We 're the leading donor in the world to help the suffering PEOPLE there .
1Bush Page: 33 a

And fortunately the rainy season will be ending shortly , which will make it easier to get aid there and help the long_suffering PEOPLE there .
1Bush Page: 34 a

I found that , in this world , that it 's important to establish good personal relationships with PEOPLE so that when you have disagreements , you 're able to disagree in a way that is effective .
1Bush Page: 39 a

And I think it 's very important for the American president , as well as other Western leaders , to remind him of the great benefits of democracy , that democracy will best help the PEOPLE realize their hopes and aspirations and dreams .
1Bush Page: 39 b

I have a plan to have a summit with all of the allies , something this president has not yet achieved , not yet been able to do to bring PEOPLE to the table .
1Kerry Page: 44 c

And we got weapons of mass destruction crossing the border every single day , and they 're blowing PEOPLE up .
1Kerry Page: 49 d

This president thought it was more important to give the wealthiest PEOPLE in America a tax cut rather than invest in homeland security .
1Kerry Page: 51 c

Some PEOPLE do n't like the fact that I stood up to say no , but I did

Figure 3: Contexts of the word people

Once the context extracts are assembled there is a THEME function available. It is based on a particular calculation of the specificity; since one is no longer looking for a relation between a word and a text, but for a privileged relation between the words themselves – which also is measured by the calculation of correlation, when two series are juxtaposed in the same series. However, the procedure is not reduced to two words compared with each other, but to the indefinite entity of all the words surrounding a specific lexical unit (or collocation) that one defines as being the pole. By confronting the word *people* with its entourage – in this case the remainder of the paragraph – one obtains a discontinuous file of words constituting a sort of "under-corpus" containing the words surrounding the pole word. This corpus is then compared to the whole corpus and the program provides a list containing the semantic surroundings, as below, for the word *people:*

```
Environnement thématique (ordre      | Environnement thématique (ordre
hiérarchique)                        | hiérarchique)
                                     |
Ecart Corpus Extrait Mot             | Ecart Corpus Extrait Mot
                                     |
43.30    214   216 PEOPLE            |  3.14      28     8 UNDERSTAND
 8.28      8     8 EARNING           |  3.08     128    24 SO
 6.45      2     3 UNAFFORDABLE      |  3.07      45    11 THESE
 6.45      2     3 ROLLING           |  3.07      45    11 ABLE
 6.36     63    22 AMERICAN          |  3.06      10     4 SOCIETY
 5.97      8     6 WEALTHIEST        |
 5.91    129    34 WHO               |  3.06      10     4 HANDS
 5.52      9     6 MEASURE           |  2.88      85    17 OR
 5.07      3     3 HIRE              |  2.81      11     4 INSURANCE
 4.91     92    24 THEIR             |  2.81      11     4 CREDIT
 4.86     21     9 MILITARY          |  2.81       7     3 SPECIAL
                                     |  2.81       7     3 BENEFITS
 4.26     38    12 LOT               |  2.76      55    12 SOME
 4.22      4     3 VALUE             |  2.59      12     4 THREE
 4.22      4     3 GEORGE            |  2.59      12     4 DIFFERENT
 4.22      4     3 CHECKS            |  2.59      12     4 ACROSS
 4.04      7     4 AFFORD            |  2.58      22     6 LONG
 3.97     14     6 START             |  2.52      28     7 HARD
 3.86     47    13 GOOD              |  2.50       8     3 VIEW
 3.80     11     5 CAMPAIGN          |  2.50       8     3 USING
 3.66      8     4 VALUES            |  2.50       8     3 LISTENING
 3.62      5     3 ADD               |  2.45      23     6 DAY
 3.54     16     6 COMING            |  2.41      18     5 LOVE
 3.52     51    13 LIKE              |  2.41      18     5 BORDER
 3.46   1742   226 TO                |  2.39      54    11 THAN
 3.27     32     9 YEAR              |  2.38     214    33 ARE
 3.20    202    35 BUT               |  2.25       9     3 REASONS
 3.18    146    27 CAN               |  2.25       9     3 HURT
 3.18     18     6 RIGHTS            |  2.22      14     4 FULLY
 3.17      6     3 OIL               |  2.10     101    17 OUT
 3.17      6     3 AFFECTED          |
```

Figure 4. The thematic environment of people

This tool also allows investigations and analyses of the lexical distance – or lexical connections comparing different dictionary corpora. In this case, the program takes the whole corpus made out from different dictionaries into account, to find out which dictionaries that are thematically close to each other.

Lexical distance

There are several methods to calculate lexical distance. The easiest one is to take into account the presence or absence of a word in a corpus. To be precise, when one seeks to estimate the connection between to texts, a certain word contributes to bring these two texts closer if it is common to both of them and to increase the distance if it is private and observed only in one of the dictionaries. The collection of the data is rather heavy because it is necessary to consider all the words without exception and that for each word the program must take into account all possibilities of matching of the texts two-and-two. This kind of analysis can be very useful and revealing, as the program allows the treatment of large corpora and allows an instant comparison between more than fifty dictionary corpora. In fact, the different analyses show that there is a strong chronological factor involved and that the dictionaries from the same époque tend to be semantically closer to each other and also that there is an important difference in the French language within the bilingual dictionaries, related to the target language. In fact, these analyses show that there is a lexical proximity between the French corpus in the dictionaries exploring Latin languages in the same way as there is a proximity in the Scandinavian and the Germanic spheres.

*

The possibilities by exploring a lexicostatistical tool in lexicography given by *Lexicotext* are multiple and combinatory. Applying a lexicostatistical database, containing only one dictionary corpus, allows documentary and statistical analysis of a specific dictionary, and gives precise and useful information to the lexicographer. The different versions of *Lexicotext,* containing an inventory of dictionaries from different publishers or different editions of the same dictionary, open up new and interesting possibilities for a more theoretical comparative dictionary research. Finally, the use of a large "multi-genre" corpus, containing literary, political or journalistic text is a reliable and useful tool for dictionary compilation.

Bibliography

Béjoint H., Thoiron Ph. 1996. *Les dictionnaires bilingues*, Louvain-la-Neuve, Aupelf-Uref, Editions.

Kastberg Sjöblom M. 2003. 'Les dictionnaires dans la paire français–suédois ; une approche culturelle' in A.-M. Laurian et T. Szende (eds.) *Dictionnaires bilingues et interculturalité,* Editions Peter Lang, Collection « Etudes contrastives », Berne, pp. 183-200.

Matoré G. 1953. *La méthode en lexicologie, domaine français*, Paris, Marcel Didier.

Norstedts stora svenk-franska och fransk-svenska ordbok*, le Grand Dictionnaire français-suédois et suédois-français*. 1998. Stockholm, Norstedts.

Rastier F. 1991. *Sémantique et recherches cognitives*, Paris, puf, coll. Formes sémiotiques.

Svensén B. 1987. *Handbok i lexikografi, Principer och metoder i ordboksarbetet*, Stockholm, TNC, Esselte Studium.

Czech Lexical Database – First Stage

JANA KLÍMOVÁ, KAREL OLIVA
Institute of the Czech Language,
Academy of Sciences of the Czech Republic
Letenská 4, 118 51 Praha 1, Czech Republic
{klimova,oliva}@ujc.cas.cz

KAREL PALA
Faculty of Informatics,
Masaryk University
Botanická, 602 00 Brno, Czech Republic
pala@fi.muni.cz

ABSTRACT

The aim of this paper is to present the main ideas of the project of the Czech lexical database (CLD) and to describe the current state of the art of the data sources. This project is intended to be performed jointly by the Institute of the Czech Language, Academy of Sciences of the Czech Republic, Prague (ÚJČ AV ČR) and Faculty of Informatics, Masaryk University, Brno (FI MU). Data available in different dictionaries and text corpora of both written and spoken language (existing in different formats and structures) will serve for the build up of the CLD.

In this paper the conception of CLD will be described in detail. In particular, its structure and the contents of the particular fields of the database will be presented, the unified format of data sources will be proposed and the software tools needed for this work will be listed.

The ultimate aim of the lexical database of the Czech language will be to serve as a source for the most different applications of Natural Language Processing (NLP) systems, as e.g. lemmatisation, tagging, machine translation and last but not least for the lexicographic use.

The work on this paper has been supported by the grant No 405/03/0913 of the Czech Science Foundation (Grantová agentura ČR).

1 Introduction

The idea of the Czech lexical database has been existing for several years and was inspired by several similar projects abroad, as e.g. the Celex database. The intention is to collect resources existing in electronic form and containing information about Czech words, as dictionaries and text corpora. All these data will be converted into the suggested unified XML format (see sect. 6 below) and concentrated in a database with the structure proposed in the paper. This database will be prepared for all kinds of linguistic research.

2 Data resources (in electronic form)

Texts from newspapers, magazines, fiction and specialised literature existing in electronic form are collected and transformed into a corpus at the Institute of the Czech National Corpus (at the Faculty of Arts, Charles University, Prague). Also the corpora existing in the Natural Language Processing Laboratory (at Faculty of Informatics Masaryk University Brno) containing almost 630 mil. current words are available for this project. The texts come from different sources, exist in different formats and in various qualities. After conversion into a unified format these texts are concentrated into the representative corpus (Králík, 2001) SYN2000 (containing 100 million current words) or into the Bank of CNC that is still being enlarged. Texts will be gathered also from web sites.

Several texts of spoken language, e.g. of Prague or Brno variants of the colloquial Czech or various T.V. debates, are also available. The lemmatisation of the colloquial language is more complicated than that of the standard language. However, the morphological module Ajka (see in Sedláček 2004 below) is able to handle lemmatisation of the texts in colloquial Czech.

There exists a great variety of Czech language dictionaries of different types: monolingual, bilingual, specialised, valency, phraseological and synonymical. Some of them, e.g. Dictionary of Literary Czech (Slovník spisovného jazyka českého = SSČ), Dictionary of Literary Czech Language (Slovník spisovného jazyka českého = SSJČ), Dictionary of Czech Synonyms (Pala, Všianský, 1994) Czech WordNet (Pala, Smrž, 2004) exist in electronic form.

3 Basic ideas of the conception of the Czech lexical database (CLD)

CLD will contain all kinds of linguistic information about words, cover data mainly from the 21st century, while the information about words collected in the 20th century will be included, too. It is intended to bring language resources from standard and colloquial Czech together. The data in the database will enable to compare the information about words coming from different electronic data sources.

As can be seen in classical dictionaries, the lexicographers follow several general but rather pragmatic principles in building the dictionary definitions. In other words, the techniques applied in building dictionary definitions are based on the selected general principles but we can hardly say that they form a consistent and complete theory. In this respect, it can be observed that there are considerable differences between semantic theories (see e.g. Leech 1974, Lyons 1977, Cruse 2004) and the lexicographer's practice. However, the dictionaries are still almost the only resources of the lexical data for NLP, thus we have to pay attention to them at the first place.

Though lexicographers use well-established techniques, many objections can be raised with regard to the consistency of the dictionary definitions, both from the formal and from the semantic point of view. A considerable number of the dictionary definitions are expressed just by examples though selected carefully from corpora, see for example the dictionaries like NODE (Hanks, 1998). It is useful to have a look at the types of the definitions (meaning descriptions) that can be found in dictionaries and should appear in CLD as well:

- definitions using **genus proximum** (GP) and the distinguishers (differentia specifica); these are mostly typical for nouns,
- definitions using **semantic components** or **features** (primitives), quite often with verbs: for example *kill = cause to die*,
- definitions based on the **relation of troponymy** are typical for verbs: e.g. *talk = whisper*, *cry = sob*,
- definitions using **synonymical explanations** or just one word synonyms (typical for adjectives, for example *clever = bright*, *beautiful = nice, pretty*),
- definitions based on **collocational determination** of the sense of entry (typical for adjectives: *good student*, *good mile*).
- definitions exploiting various kinds of **ad hoc descriptions** or explanations (these can occur with any PoS),
- definitions based on the **descriptions of events** or **situations** (see e.g. the following definition: *if you ask for a* ***table*** *in a restaurant, you want to have a meal there* (Cobuild 95, p. 1697).

In CLD it will be our task to treat the mentioned types of the definitions systematically and consistently, investigate the possible relations between them and describe them as formally as possible.

4 Structure of the Czech lexical database

The individual entries will include as basic units either single lemmata like *dům (house)* or standard collocations such as e.g. *vysoká škola* (*university*). Other types of collocations such as toponyms, proper nouns, acronyms (so called named entities) will be considered as well and included either as the headwords or they will be contained in the special lists linked to the respective entries. We propose that a record in the CLD have the following structure:

1. **orthographic variants** – the Rules of Czech orthography (*Pravidla českého pravopisu*, 1993) compiled in ÚJČ AV ČR containing a list of spelling variants (with the size approximately 12 000 units) will be included after revision.

2. **morphology** – all the necessary morphological information for Czech has to be accessible, i.e. information about PoS and the respective grammatical categories, as well as the inflectional paradigms (i.e. all forms of the given word). In other words, it means that the detailed information about the inflectional pattern(s) has to be offered as well. This can be done in two ways:

- the morphological module, i.e. lemmatizer and generator will be integrated into the CLD and it will produce all the necessary information dynamically on demand, using the aproppriate indices, or

- the morphological information can be pre-processed and stored in the corresponding tables from which it can be obtained when required.

In both cases the morphological module Ajka (Sedláček 2004) is able to yield the necessary information.

For example, for verbs this information typically includes 8 categories (attributes): <negation>, <person>, <number>, <tense>, <mode>, <voice>, <aspect> and <gender>. Their values would be accessed dynamically through the <inflectional paradigm index of the verb>. For other PoS it can be done in a similar way.

3. **segmentation of the headword**, it includes two kinds of information:

a) morphemic segmentation displaying the basic segmentation into prefixes, stems, suffixes, endings and even postfixes. All this information belongs to the morphology field. Morphological module Ajka offers this kind of segmentation in a basic form. Morphemic segmentation is also related to the area of the derivational morphology which is mentioned below with regard to the derivational nests,

b) hyphenation – the information how to segment words for the purposes of typesetting. This kind of segmentation is not typically based on the morphemic segmentation. The hyphenator assumed here will be the one used in TeX typesetting environment and based on the patterns (TexLive 2004).

4. **sense description** or **definition**, with <senses1...n>, where for each sense the following should be given:

- <**semantic features**> that can be associated with an entry – now it appears possible to base them on a selected ontology such as EuroWordNet Top Ontology

(Vossen, 1999). According to our knowledge there have not been systematic attempts to use semantic features for the meaning descriptions on larger scale. They seem to be more suitable for some nouns only, though in Princeton v. 2.0 WordNet and in WordNets of other languages the semantic features are associated with all nouns. In our view, it will be useful to include the feature information for all nouns in CLD if it is at hand. The final decision can be made later if it appears necessary.

It can be seen that the semantic features come from the hypero/hyponymical trees (H/H) in WordNets but it has to be examined how large parts of the trees or subtrees can be exploited – we estimate that the plausible number of the nodes (= semantic features) used may be about 5-6,

- <descriptions using **genus proximum** (hypero/hyponymy relationships)> and **<distinguishers>** (differentia specifica) (GP + d1, d2, ..., dn scheme) represent the standard type of the meaning description that will appear in CLD, and as a rule it will be given for the noun entries. In fact, the genus proximum definitions can be viewed as the subparts of the hypero/hyponymy trees where only two adjacent nodes are considered. The *<distinguishers>* represent a sort of problem: we know that individual dictionaries differ most in the way in which they deal with the distinguishers – there is no general agreement as to which distinguishers should be selected and included in the particular entries. In CLD we intend to take the distinguishers from the existing Czech dictionaries (namely SSJČ and SSČ).

- **<semantic classes of verbs>** – for verbs the genus proximum definitions may not be always appropriate and they work well only for selected classes of the verbs. The definitions then can exploit semantic classes the verbs belong to and they will be offered where possible. In this respect we are preparing a semantic classification of Czech verbs related to Levin's (Levin, 1995) though in Czech this task is going to be more complicated because of the category of aspect, thanks to which Czech verbs regularly occur in pairs (the number of the Czech verbs in the list counts presently about 3200 verbs).

On the other hand, it is also obvious that the semantic classes of verbs are closely related to the valency (verb) frames and especially to the semantic roles appearing in the valency frames. It can be observed that the particular semantic classes (for example, verbs of drinking or verbs of clothing, etc.) combine with a small number of semantic roles typical just for a given semantic class. Using corpus data we will be able to check the individual semantic classes with the corresponding semantic roles characteristic to them and in this way make the development of both the valency frames and the semantic classes objective and reliable enough.

- **<synset>** will be given for a given lexical unit (entry, lemma), possibly in the WordNet fashion or following the Dictionary of Czech Synonyms (Slovník českých synonym, SČS). The reasons for having synsets follow from the fact that the relation of synonymy (and antonymy as well) can serve as one of few relatively reliable ways of characterisation of meaning.

- **valencies of the nouns, adjectives and adverbs** will be given in CLD as well. An attempt will be made to obtain adjective valencies from the corresponding verb frames, if possible semiautomatically. It is also obvious that surface valencies can be recorded in a similar way as for verbs, and the semantic roles can be the same as for verbs as well. The same approach can be applied, in our view, to the noun and the adverb valencies.

We have set it as our task to include in CLD verb valency frames containing both the syntactic (morphological cases) and semantic information (semantic roles or selectional restrictions). The question may be raised where to place the valency frames: whether to the syntactic field of CLD or to the semantic one. It is obvious that the verb frames comprise both the syntactic information about the verb itself, i.e. which morphological (in the highly inflected languages such as Czech) cases are required by a particular verb and which semantic roles are required by the meaning of a verb. Considering the complex character of the information yielded by the valency frames we place them in the field related to the description of the meaning of verb entries.

It should be remarked that standard (both paper and electronical) dictionaries bring this kind of information only in part, for example, in the representative SSJČ the syntactic (surface) valencies are sometimes given explicitly but rather implicitly through the examples and without any detailed information about the semantic roles (it should be remarked, however, that when SSJČ was published (in 1960-61), the theory of the deep cases and semantic roles had not existed.

It should be remarked that valency frames for Czech verbs are presently being prepared for the Czech WordNet (about 1600 items) at NLP Lab at FI MU and for the ValLex Dictionary those are being developed at the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics at Charles University Prague, (approx. 2500 items, Lopatková, Žabokrtský, 2002, Hajičová, Panevová, 2002). The two approaches differ in the inventory of the semantic roles used but for the purpose of CLD an integrated solution will be seeked for.

Below we give examples of the valency frames for Czech verbs *učit* (*teach*), *pít* (*drink*) and *vyhrát* (*win*) as they are developed within the framework of the Czech WordNet.

They can be represented in the text form in the following way:

(vf2){*vyučovat:1,teach:1*} kdo1*AG(person:1) ;
co4*KNOW(subject:3).
(vf3){*pít:1,drink:1*} kdo1*AG(person:1|animal:1) ;
co4*LIQUID(beverage:1).
(vf4){*vyhrát:1,win:1*} kdo1*AG(person:1|organization:1) ;
co4*ACTIVITY(human activity:1).

The forms of the pronouns *kdo1* (*who*), *co4* (*what*) with the numbers indicating the morphological cases (nominative, genitive...) refer to the surface (syntactic) valencies. The semantic roles have two level structure, the first part is formed by a general label like AGENT, PATIENT, INSTRUMENT, etc., the second part is represented by a particular literal from Princeton WordNet 2.0, which can be found in a respective H/H tree and it yields a selectional restriction required by a given verb. In this way the notation allows us to capture the existing differences in lexical meaning.

It is considered if and in what form the other fields, as e.g. phonetics (pronunciation variants, phonetic output), genre and style characteristics, phrasemes and typical collocations and etymology will be included in the CLD.

5 Tools for CLD

For building up the CLD a number of tools is needed. In our view, at least the following ones should be considered:

Ajka – morphological module yielding automatically information about POS, inflection, grammatical categories, derivational relations of the entries,

I_PAR – editor of the Czech morphological database, cooperates with Ajka and is able to associate stems with the Ajka paradigms automatically,

SYNT and **VADIS** – syntactic parsers for Czech, able to recognize and generate Czech syntactic structures (using two separate formal grammars of Czech),

VisDic – local editor and browser for WordNet-like lexical databases and other dictionaries based on XML format,

DEB – dictionary editor and browser (Smrž, Povolný, 2003) that will serve as a main tool for the preparation of CLD. In fact, it is intended to be a lexicographer's workbench, allowing to integrate in one tool the corpus manager Manatee/Bonito, corpus editor CED, morphological module Ajka and other necessary dictionary and text resources.

Manatee/Bonito & **Word Sketches** (Rychlý, Smrž, 2004) – corpus manager and its graphical interface which now exists as a client/server tool or its new version Bonito 2 implemented as a web interface. Word Sketches engine offers the immediate contexts for the words from a corpus and also semantic clusters.

6 Electronic version of the Dictionary of Literary Czech Language – ESSJČ

The representative Dictionary of Literary Czech Language (referred to as SSJČ, 1960-1971) has been converted into XML format under the grant project GAČR 405/96/K214 (*Czech in the Age of Computers,* 1999-2001). At ÚJČ AV ČR it was scanned and processed by OCR software and then converted into MSWord format, at the NLP Lab at FI MU the data were ported into the XML format.

Presently, we work with the two variants of the XML format:
- low level format using labels of the different font types – it is more suitable for correcting the errors but not so well readable. In the example below we show it for entry *terorismus* (*terrorism*):

```
<entry>
<bold>terorismus</bold>
<ital>způsob vlády vymáhající terorem poslušnost; hrůzovláda, krutovláda,
despotismus:</ital>
<norm>vojenský t.; nesnesitelný t.; demagogie a t.; </norm>
<small>přen. expr.</small>
<norm>to je t., nedejte si to líbit</norm>
</entry>
```

- high level format reflecting the logical structure of the entries in SSJČ – it is the final XML format that will be used also for building CLD and storing other Czech dictionaries. It is more suitable for querying and standardisation. The entry for *terorismus* (see above) takes the following form:

```
<entry>
  <hw>
        <orth>terorismus</orth>
  </hw>
   <morph>
    <paradig>socialismus</paradig>
  </morph>
  <senses>
   <sense>
      <def>způsob vlády vymáhající terorem poslušnost</def>
      <def>hrůzovláda</def>
      <def>krutovláda</def>
      <def>despotismus</def>
      <eg>vojenský terorismus</eg>
      <eg>nesnesitelný terorismus</eg>
      <eg>demagogie a terorismus</eg>
      <eg>
   <usg type=style>přen.expr.</usg>
        to je terorismus, nedejte si to líbit
      </eg>
   </sense>
  </senses>
</entry>
```

The electronic version of SSJČ helped to obtain also some statistical information showing e.g. the frequency distribution of the particular PoS's contained in the ESSJČ (see Table 1).

7 Conclusion

In the paper we have presented the outline of the project of Czech lexical database and described the current state of the art of the data sources. We have shown what software tools will be used in the preparation of CLD and a proposed unified XML format, in which data will be stored. The data contained in the CLD will serve as a source for the most different applications in the field of Natural Language Processing (NLP), as e.g. lemmatisation, tagging, machine translation and last but not least also for the lexicographic use.

POS	headwords	in a paragraph	total
nouns	58777	18257	77034
adjectives	30084	11599	41683
pronouns	334		
numerals	188	147	335
verbs	31621	11004	42625
adverbs	2036	11104	13040
prepositions	83		
conjunctions	72		
particles	14		
interjections	1142		
acronyms	132		
prefixes	89		
compounds	709		
verb bases	1444		
total	126725	52111	174717

Table 1 Distributed frequency of the particular PoS's in SSJČ

References

Celex, Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen 2001, http://www.ru.nl/celex/

Cruse A.: Meaning in Language: Introduction in Semantics and Pragmatics (2nd Edition), OUP Oxford 2004

Čermák F., Klímová J., Pala K., Petkevič V.: Design of the Czech lexical database, In: Proceedings of Corpus linguistics conference, (Ed. McEnery A., Rayson P.) - University of Lancaster 2001, pp. 119-125

Fellbaum C. (ed.): WordNet: An Electronic Lexical Database, MIT Press 1998

Hajičová E.: Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study, in: Proceedings of LREC 2002, Las Palmas, 2002

Hajičová E., et al: Prague Dependency TreeBank 1.0, CD ROM, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, 2001

Hanks P. et al., New Oxford Dictionary of English, Oxford University Press 1998

Králík J.: Vyvážení zdrojů synchronního korpusu češtiny SYN2000. Slovo a slovesnost 62, 2001, č. 1, s. 38-53

Leech G.: Semantics. Harmondsworth, Penguins, 1974

Levin B.: English Verb Classes and Alternations. Chicago, The University of Chicago Press 1995

Lopatková M., Žabokrtský Z.: Valency Dictionary of Czech Verbs. ELRA 2002

Lyons J.: Semantics. Cambridge University Press, Cambridge 1977

Pala K., Všianský J.: Slovník českých synonym. Nakladatelství Lidové noviny, 2.vyd., Praha 2000

Pala K., Smrž P.: Building Czech Wordnet. Romanian Journal of Information Technology and Science, vol. 7, No 1-2, pp. 79-88, Bucharest, 2004

Pala K.: A List of the Czech Surface Valencies, 16 000 verbs, NLP Lab. Faculty of Informatics, Masaryk University (unpublished), 2000-2005

Pala K., Rychlý P., Smrž P.: DESAM – Annotated Corpus for Czech, In Proceedings of SOFSEM 97, Heidelberg, Springer Verlag, 1997. pp. 523-530

Pravidla českého pravopisu, Academia Praha 1993

Příruční slovník jazyka českého, Česká Akademie věd a umění, Praha 1935-1957

Rychlý P., Smrž P.: Manatee, Bonito and Word Sketches for Czech, Trudy meždunarodnoj konferencii ”Korpusnaja lingvistika” – 2004. Izdatel'stvo Sankt-Peterburgskogo universiteta, Sankt-Petersburg, 2004, pp. 324-334.

Sedláček R.: A Morphemic Analyser for Czech, Ph.D. Thesis, Faculty of Informatics, Masaryk University, Brno, 2004

Slovník spisovné češtiny pro školu a veřejnost (SSČ), Academia Praha 2001

Slovník spisovného jazyka českého (SSJČ), electronic version (referred to as ESSJČ), ÚJČ AV ČR Praha, FI MU Brno, 2002

Smrž P., Povolný M.: DEB – Dictionary editing and browsing, Proceedings of the EACL03 on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003), pp.49-55, Budapest, 2003

SYN2000: Ústav Českého národního korpusu FF UK Praha 2000, http://ucnk.ff.cuni.cz

ValLex: http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/

Vossen P. et al.: Final Report on EuroWordNet-2, 2D041. CD ROM, v.1, Amsterdam, University of Amsterdam,1999

Attempts and examples for the discovery of hidden information of the Concise explanatory dictionary of Hungarian (2nd edition, 2003)

ATTILA, MÁRTONFI

MTA–ELTE Research Group of Academic Dictionary of Hungarian
Budapest VI., Benczúr u. 33. H–1068
wad@ludens.elte.hu

Knowledge discovery and data mining – as its part – are trendy areas of IT, their aim is utilizing characteristically commercial databases, at least partially its approach and toolkit are applicable to lexicographical databases. The XML version of the new edition of the *Concise explanatory dictionary of Hungarian* (ÉKsz.[2]) provides not only a more complete and more modern data chart (than merely the substitution of *a)* the length in characters, *b)* the number of senses, *c)* the etymology, and *d)* the usage label given in the head of entries, which were omitted from Papp Ferenc's *Reverse-alphabetized dictionary of the Hungarian language* VégSz.), but contains also up-to-date etymological facts about the widest group of Hungarian words, and absolute frequency data based on *Hungarian National Corpus*. With some simple queries the generated relational database gives type and token frequency indices of various etymology, usage label, part-of-speech or number of senses word-groups. With the toolkit of data mining more interesting analyses could be performed to discover hidden patterns of the above parameters by means of extracting association rules, these are concerned with syllable-structure, part-of-speech category, etymology, semantic structure, etc.

1 Introduction

Knowledge discovery and data mining – as its part – are trendy areas of IT, their aim is utilizing characteristically commercial databases. However the goal (namely extracting as much hidden data and unknown patterns as possible in an automatized manner) is essentially the same as the most general goal of scientific research, therefore at least partially its approach and toolkit are applicable to lexicographical databases. (Since the size of lexicographical databases is usually smaller by orders of magnitude than monumental commercial databases occurring with the primer area of data mining, the device requirement of the operations is significantly less and the extractable information is more restricted.)

The first notable lexicographical database of the Hungarian language is Papp Ferenc's *Reverse-alphabetized dictionary of the Hungarian language* (VégSz.) (Papp 1969) and its derivative database on PC. The database which is the base of Papp's dictionary has four additional fields in comparison to the paper-version: the length in characters, the number of senses in ÉrtSz. *(Explanatory dictionary of the Hungarian language)* (Bárczi–Országh 1959–1962), the etymology based on SzófSz. *(Etymological dictionary of Hungarian)* (Bárczi 1941), and the usage label given in the head of entries in ÉrtSz. – because of typographical reasons these are omitted from the paper-version and its derivative database.

2 Conversion of data

The new edition of the *Concise explanatory dictionary of Hungarian* (ÉKsz.^{2}) (Pusztai 2003) – as an up-to-date lexicographical project should be – was first prepared as an XML document, and though its grammatical information (constituting the skeleton of VégSz.) is substantially more poor, with suitable conversions a more complete and more modern data chart can be generated. It is more modern, because ÉKsz.^{2} provides up-to-date etymological facts about the widest group of Hungarian words, and it is more complete, since apart from the part-of-speech and usage labels and the numbers of drawn senses all of the entries in this dictionary have the absolute frequency based on *Hungarian National Corpus,* furthermore the word-length in the number of phonemes or syllables can be coded. (See *Table* 1.)

Therefore it was necessary to generate a data chart from the XML document, which serves the required information in a simple form, in order to obtain the more or less hidden information from the corpus. After the conversion (because in another structure other data errors may appear) we could perform some data-cleaning operation, too. The converted and cleaned data chart contained 72,444 records, with 10 fields in each record.

Each field contains the following information: *ID* – individual ID of the record; *Lemma* – the main variant of the lemma; *Split* – indicator of that if there is | or ~ in the lemma (existence of it refers to morphological articulateness, in case of its absence the word can be articulated or unarticulated); *Syll* – the length of lemma in the number of syllables; *Phon* – the length of lemma in the number of phonemes; *Freq* – the absolute frequency based on Hungarian National Corpus; *Usg* – usage labels given in head of entries separated by spaces; *POS* – part-of-speech labels separated by spaces; *Sens* – the number of senses; *Etym* – etymology in compressed form (in fact the name of the language or language-family of origin, sometimes the mode of origin).

3 Simple queries

On the base of this database we can distinguish two kinds of frequencies: type frequency shows how often a feature appears in the dictionary, token frequency shows how often a feature appears in a text-corpus. With some simple queries the generated relational database gives token and type frequency indices of various etymology, usage label, part-of-speech or number of senses word-groups. (Since those fields, which contain

Table 1: Some records of the database

ID	*Lemma*	*Split*	*Syll*	*Phon*	*Freq*	*Usg*	*POS*	*Sens*	*Etym*
3053	asszonykerülő 'woman-hater'	☑	5	11	0		adj n	1	
3054	asszonykéz 'woman's hand'	☑	3	8	21		n	2	
3055	asszonykormány 'petticoat government'	☑	4	11	0	rare joc	n	1	
3056	asszonymunka 'woman's job'	☑	4	10	1	dial	n	2	
3057	asszonynéni 'woman-aunt'	☑	4	9	0	arch	n	1	
3058	asszonynép 'womenfolk'	☑	3	8	62	dial	n	1	
3059	asszonynév 'married name'	☑	3	8	33		n	1	
3060	asszonyos 'womanlike'	□	3	7	34		adj	2	
3061	asszonypajtás 'better half'	☑	4	11	7	fam joc	n	1	
3062	asszonyrokon 'kinswoman'	☑	4	10	1		n	1	
3063	asszonyság 'countrywoman'	□	3	8	296		n	3	
3064	asszonytárs 'fellow-woman'	□	3	9	15		n	2	
3065	asztag 'rick'	□	2	5	92		n	2	Slav.
3066	asztal 'table'	□	2	5	16,627		n	5	Slav.
3067	asztalbontás 'leaving the table'	☑	4	11	4	ref	n	1	

more usage or part-of-speech labels, in case of *n* labels the k^{th} label got $\frac{k}{1+2+\cdots+n}$ – fortunately in the case of $n = k = 1$ it is 1)[1]. Such token frequency indices – for want of a satisfactory database or corpus background – formerly could not have been calculated; the type frequencies provide the possibility for comparison with Papp's examinations based on former sources (Papp 1969, 2000). (See *Figures* 1–8.[2, 3])

[1] Values of token frequency are inaccurate given from the essence of the calculation, because there are no data about the real dispersion of part-of-speech or usage labels of each token.

[2] Words without etymology label have been coded as inner origin.

[3] All of the terminological (so the capitalized) labels has been coded as *term*.

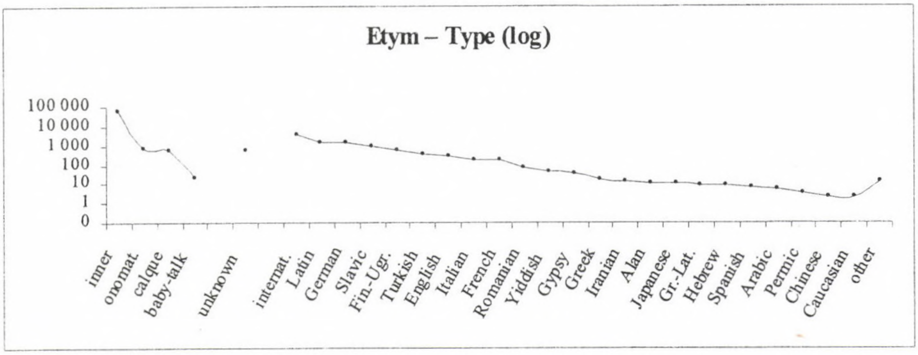


Figure 1: Type frequency of several etymology words (on log-scale)

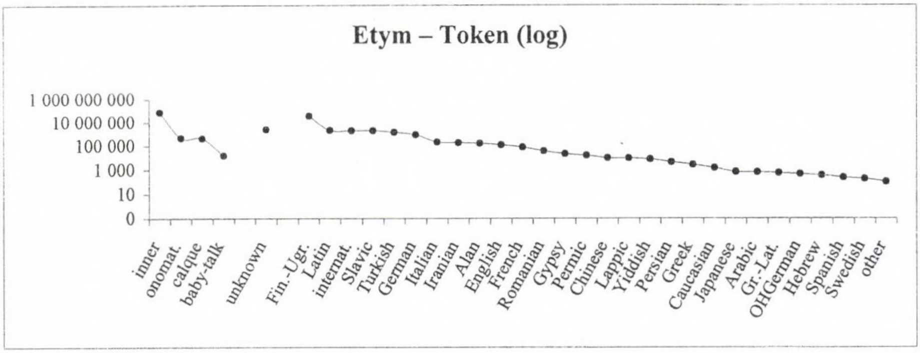


Figure 2: Token frequency of several etymology words (on log-scale)

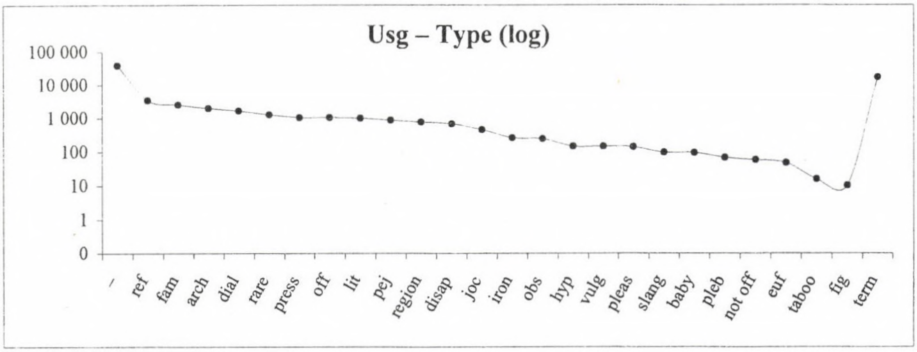


Figure 3: Type frequency of several usage labeled words (on log-scale)

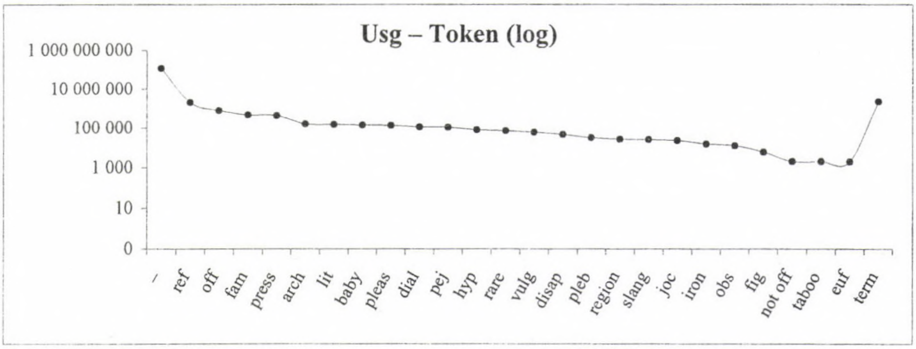


Figure 4: Token frequency of several usage labeled words (on log-scale)

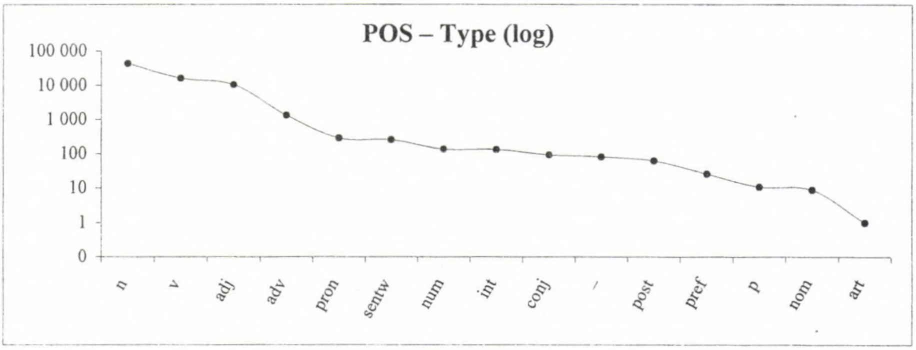


Figure 5: Type frequency of several part-of-speech labeled words (on log-scale)

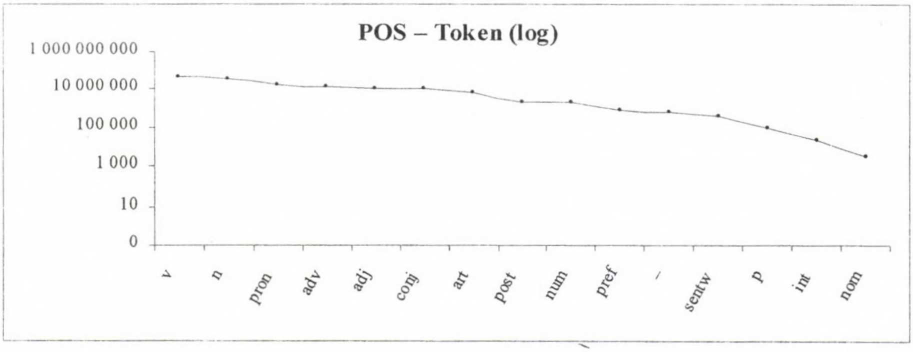


Figure 6: Token frequency of several part-of-speech labeled words (on log-scale)

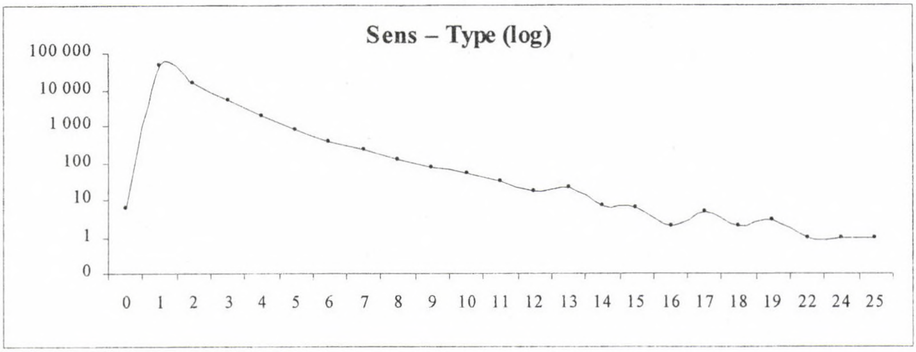


Figure 7: Type frequency of words having several senses (on log-scale)

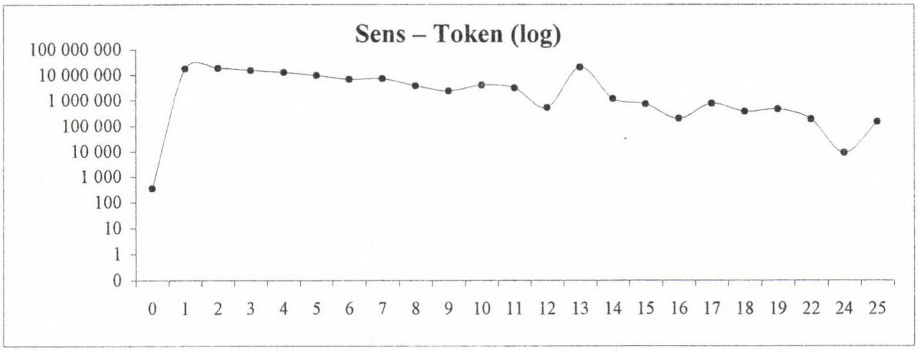


Figure 8: Token frequency of words having several senses (on log-scale)

4 Association rules

With the toolkit of data mining more interesting analyses could be performed to discover hidden patterns of the above parameters by means of extracting association rules. These rules describe the statistical connection of different values in a data table. The general form of association rules is: $X_1, \ldots, X_n \Rightarrow Y$. In exploring the hidden patterns – in case of more usage or part-of-speech labels – only the pattern in the first position was taken into consideration.

After selecting the pairs whose holder[4] is at least 10, altogether 7,015 rule-candidates have been founded, further examinations affected just 254 pairs of them, which had more than 1 percent (724) holder). Only 103 rules had at least 67 percent probability at least in one direction.

[4] Holder = the frequency of concomitance.

The *A*→*B* association rules in which the frequency of *A* is low and the frequency of *B* is high are not really meaningful, although there are many of them. The most frequent *B*s were the 'inner origin' and the 'monosemantic', occurring in 71 pairs of the above mentioned 103.

The connection between the word-length measured in the number of phonemes and measured in the number syllables is meaningless in a certain sense, because this is trivial. However if is also expressive, since it gives information about the syllable-length in Hungarian. In this topic there are the following association rules:[5]

5 phonemes ⇒ 2 syllables	(8%, 92%);
3 phonemes ⇒ 1 syllable	(1%, 87%);
7 phonemes ⇒ 3 syllables	(11%, 81%);
10 phonemes ⇒ 4 syllables	(9%, 79%);
8 phonemes ⇒ 3 syllables	(12%, 77%);
4 phonemes ⇒ 2 syllables	(2%, 73%);
13 phonemes ⇒ 5 syllables	(2%, 68%);
12 phonemes ⇒ 5 syllables	(3%, 67%).

Reversal of any rule does not reach the 50%.

Although the 'noun' has quite high frequency, too, we have to mention that the

terminological ⇒ noun	(20%, 87%)

rule because of its high holder and frequency, as well as those rules, which declare, that word with very low frequency are nouns in all likelihood:[6]

0 occurrence ⇒ noun	(5%, 77%);
1 occurrence ⇒ noun	(3%, 72%);
2 occurrences ⇒ noun	(2%, 69%);
5 occurrences ⇒ noun	(1%, 68%);
6 occurrences ⇒ noun	(1%, 67%);
4 occurrences ⇒ noun	(1%, 67%).

We can observe that the longer words are with larger probability nouns, too:

14 phonemes ⇒ noun	(1%, 73%);
6 syllables ⇒ noun	(3%, 71%);
13 phonemes ⇒ noun	(2%, 71%);
12 phonemes ⇒ noun	(4%, 70%);

[5] The first figure in the parentheses is the holder, followed by the probability of the rule in percent. The latter one means how often a rule is operative.

[6] For interpretation of rules it is important to know, that 59 percent of the lemmas are nouns.

5 syllables ⇒ noun (8%, 68%);
11 phonemes ⇒ noun (6%, 67%).

The lemmas, which have *archaic* label, are quite often nouns:

arch ⇒ noun (1%, 67%)

The etymology is connected with as noun part-of-speech, too

Slavic ⇒ noun (1%, 87%);
international ⇒ noun (1%, 87%);
German ⇒ noun (2%, 75%);
Latin ⇒ noun (2%, 70%).

because the reason of this is that, the noun is the most open (in fact the only totally open) part-of-speech category, so it is the mostly affected group by the borrowing of words.

In the scope of the lemmas, which has no usage label in the head of the entry[7] we can observe some regularities. Naturally, often there is no usage label in the head of the polysemantic lemmas:

5 senses ⇒ no usage label (1%, 97%);
4 senses ⇒ no usage label (2%, 93%);
3 senses ⇒ no usage label (6%, 88%);
2 senses ⇒ no usage label (17%, 77%).

For comparison, in the case of monosemantic lemmas:

1 sense ⇒ no usage label (27%, 42%).

An interesting, 4-holdered rule related to the absence of the usage labels, too:

verb ⇒ no usage label (16%, 71%)

– which is probably induced by the verb's dispose for polysemantism. The rule

1 syllable ⇒ no usage label (2%, 67%)

might have similar reason.

[7] These are the 56 percent of lemmas.

5 Conclusion

Above we examined, what types of information are hidden in a dictionary, and we showed some examples, how to obtain them.

References

Bárczi Géza (1941). *Magyar szófejtő szótár* (Etymological dictionary of Hungarian). Egyetemi Nyomda, Budapest.

Bárczi Géza–Országh László (eds.) (1959–1962). *A magyar nyelv értelmező szótára I–VII.* (Explanatory dictionary of the Hungarian language I–VII). Akadémiai Kiadó, Budapest.

Papp Ferenc (ed.) (1969). *A magyar nyelv szóvégmutató szótára* (Reverse-alphabetized dictionary of the Hungarian language). Akadémiai Kiadó, Budapest.

Papp Ferenc (2000). *A debreceni thészaurusz* (Thesaurus of Debrecen). Linguistica. Series C. Relationes 11. MTA Nyelvtudományi Intézet, Budapest.

Pusztai Ferenc (ed.) (2003^2). *Magyar értelmező kéziszótár* (Concise explanatory dictionary of Hungarian). Akadémiai Kiadó, Budapest.

Adaptation of a computerized dictionary for language learning: The "*Trésor de la Langue Française informatisé*" and French language

CHRYSTA PELISSIER
ATILF & LIRDEF
17 quai du Port Neuf F-34500 Béziers
chrysta.pelissier@iutbeziers.univ-montp2.fr

CLAIRE BECKER
ATILF & CRAPEL
44, avenue de la Libération F-54063 Nancy cedex
claire.becker@atilf.fr

Abstract:

The objective of this article is to present the first results of a research related to the design of a linguistic resource dedicated to help pupils in their learning process. We intend to show how a dictionary for language specialists can be adapted in order to help children to learn French language.

By "linguistic resource" we mean any document (paper and computerized), which presents us with information related to language. Among these documents, we can quote language dictionaries, encyclopaedias, lexicons and glossaries, etc. The use of these documents is recommended for the preparation of the baccalaureate, the exam taken by pupils before leaving high school. Indeed, French teachers are asked by the French Department of Education to integrate dictionaries, encyclopaedias and databases in their courses, as well as documents related to the press [Ministère de l'Education Nationale 2002].

This paper first presents the main computerized dictionary called the TLFi (*Trésor de la langue française informatisé*, which can be translated as "Computerized Treasury of the French Language") created by the research laboratory ATILF (Analyse et Traitement Informatique de la Langue Française). Then, we will present various uses of this dictionary for French learning in France. We will specifically focus on three didactic situations during which the dictionary is used.

1 General presentation of The TLFi

A dictionary, in its electronic form, is a textual database to be used in any natural language processing system. The size and the contents of the existing dictionaries vary a lot according to the target of the attended public and the cost of their collected resources. The *Trésor de la langue française* (or TLFi) [on the Internet at http://atilf.atilf.fr/tlf.htm] is the most important electronic dictionary on French language. It first existed as a paper version and groups the vocabulary of the 19th and 20th centuries, in sixteen volumes. The first volume was published in 1971 and the last one in 1993. It contains about 100 000 head words with their etymology and history, that means 270 000 definitions, 430 000 examples with their source, the majority of them are extracted from the database Frantext. Frantext is a textual database, which gives access to more than 4 000 texts (the access to Frantext depends on a subscription). This resource is used by linguists, researchers and others specialists of French literature. Frantext gives the possibility to consult parts of text (which contain a word or an expression) and to consult a list of authors or a list of books (which were written by a particular author at a particular time). The computerized version of the dictionary, the TLFi, contains the same data as in the paper version; with its 350 million characters, its articles are structured according to the notion of textual objects. Thanks to its software Stella, it can be seen as a finely lexical structured database.

1.1. The specific data

Its originality is based, firstly, on its wordlist, which is rich of about 100 000 entries, present either in the funds or in dictionaries in the ATILF laboratory. The TLF was a pioneer in the treatment of morphemes or in the treatment of structures of specific vocabularies. Then we can say that its originality lies in the richness of the number of examples (about 430 000) and syntagms (about 165 000), quoted throughout its 16 volumes. Besides, its list of meta-textual objects such as headwords, definitions, indications of domains, semantic and stylistic indicators, and examples with their sources, fixed phrases is exceptional (about 40 different meta-textual objects). The data are proposed in different sections: synchrony, etymology, history, pronunciation, and bibliography. One of the main advantages of a computerized dictionary is to consider it as a knowledge database in which one can extract any items contained in any textual object in eliminating noise in the requests. To allow this, the whole dictionary has been tagged into an XML document, with special delimiters for each type of its textual objects. Tags in the TLFi introduce possibilities to do queries.

1.2. Different level of queries

Three levels of queries are possible depending on the user's need. The first level is called "Simple visualization of an article". You can read an article dedicated to a specific headword by three means. Firstly, you have the possibility to write the word with mistakes if you do not know the right spelling of the word. That is very useful for the users who do not remember the right French accents (acute, grave or circumflex) for instance. All kinds of mistakes are allowed as long as the right pronunciation is correct. Secondly, you can use the possibility of seeing the list of the main articles contained in the TLFi; this allows the user to discover unknown words, just as if he was turning the pages of the paper version of the dictionary. Thirdly, the user can find an article thanks to selecting sounds and not alphabetical characters. At that level of consulting, you read the dictionary article by article, yet with easy ways of searching a word. The second level is called "aided requests". At that level you have the possibility of using the dictionary as a textual knowledge database and to make queries throughout the sixteen volumes in one click of mouse. One can make requests on graphic forms, on inflected forms as well, on sequences of words, etc. The third level is called "complex requests". The user, at that level, can make requests using regular expressions, lists of words or even more complex requests crossing criteria and playing with embedded structure or related textual objects. One can extract all the conjugated forms of the French verb "mourir" contained in the core of an example taken from Balzac. It is possible to make lists of words and to use them in requests. For example, one can extract all the words ending with suffix *–able*, and then extract from this list all words, which are not adjectives. The fine structure contained in the TLFi, allied to a very friendly user's interface, with help on line, allows pertinent results when making very complex requests.

1.3. Hyper-navigation

Hyper-navigation throughout all the databases interconnected under Stella is possible. For example, when consulting the TLFi and by simply clicking on any word in an article, the user can navigate between another article of the TLFi, Frantext if he needs more examples, a lexical database which gives information on the grammatical category of the word, the Academy dictionaries (8e and 9e editions), and the historical database for French language, called DDL. This function can help the user to construct his own knowledge. He can then navigate like he wants, to go deeper in a particular notion or a word in different resources.

2 Various uses of the TLFi

At the beginning, the TLF in its paper version was intended to a public of linguists and language specialists. The computerization of its data, the friendly interface with its help on line and the capacities of the software Stella give a new life to the TLF. The idea is to propose the TLFi as a tool for different pupils (pupils of primary, secondary and high schools). Thus, a group of researchers in the ATILF laboratory gathered together in September 2003 and created the team called EDUC'ATILF (www.atilf.fr/pedagogie). The objective of this team is now to show how important resources like the TLFi and Frantext can be adapted to different types of pupils. A particular methodology has then been used for our research [Pélissier & al. 2004]. First, we identify the knowledge concerning French language learning in primary schools. Then, we precisely describe the resource with its data and their organisation. In addition, we characterize some didactic activities using the TLFi. Indeed, we only have made a study of Frantext for high schools pupils (15-18 years old) since some texts in this database are studied within the framework of the French baccalaureate. We have set out three different uses of the two main linguistic resources of the ATILF laboratory: Frantext and the TLFi. These uses will be described later on in this article. First, we will explain the use of the TLFi for primary school pupils, then we will cope with the use of the TLFi in secondary schools and finally, we will deal with the new tool made up of Frantext and the TLFi for teachers and pupils in high schools.

2.1 Uses in primary schools

In France, the primary school begins for 5 year-old-children and ends when they are 12. For this public, some dictionaries are proposed.

2.1.1 The position of dictionaries in learning French language

Since 1972, French educational official texts give dictionaries a real place in the classroom. Nowadays, teachers advise parents to buy a dictionary for their children. This is the reason why these books nowadays have a place in French family and school. Indeed, 8 students out of 10 own a dictionary and about 30 000 dictionaries are sold every year [Gross 1989].

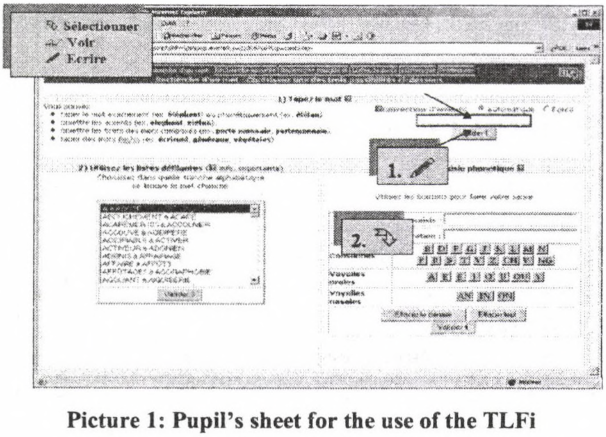
2.1.2 Different dictionaries for children

A lot of dictionaries for children are available in the market. Pupils, their parents and teachers encounter some difficulties in choosing the most adapted dictionary since it exists a lot of paper dictionaries, electronic dictionaries or language and encyclopaedic ones, associated to different cycles of primary school. The repartition of dictionaries is made according to the age and the classroom of the child ([Lehmann 2000]; [Buzon 1983]).

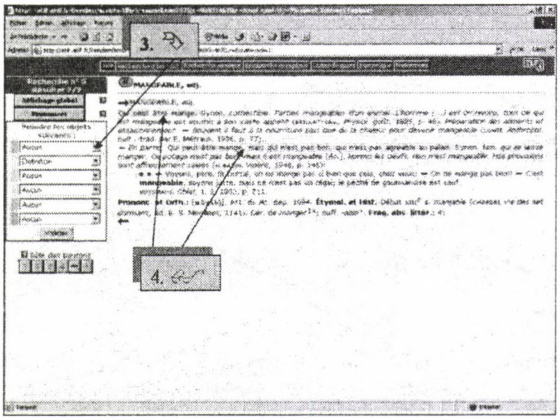
2.1.3 Problems of the dictionaries in use

The majority of electronics dictionaries generally propose only one possibility of access words. For example, in *Mon Premier dictionnaire Super Génial* [Nathan 98], only the selection of the first letter of the unknown word is given. But, for 5-year-old child, it is impossible to isolate, identify and select the first letter of word. Thus, the majority of electronic dictionaries does not offer this possibility of accessing to words when spelling is unknown by the user. Besides, another friendly way to read the computerized TLFi consists in putting in evidence one or several textual objects by colouring them. Its interest for linguistic research is obvious for it is a powerful tool to help the user who wants to study grammatical classes (verbs, adjectives and adverbs for instance), syntactical classes (in studying constructions), etymological classes (verbs borrowed from English), stylistic classes (ironical or metaphorical uses) or morphological classes (words ending with a suffix or beginning with a prefix). The use of the TLFi can also be foreseen for learning and teaching French language. We are now thinking of the pedagogic possibilities given by that computerized dictionary such as it is available at the moment. We also work on the modifications we could make in order to encourage its use in education. In order to use the TLFi as it is available nowadays, we have defined a methodology whose aim is to grant pupils an access to this important linguistic resource. In collaboration with a teacher we have prepared and tried out a set of pedagogical activities for children in the CE1 class (7-year-old children).

To illustrate our topic, here is an extract of a didactic activity made for CE1 pupils and presented to them afterwards.



Picture 1: Pupil's sheet for the use of the TLFi



Picture 2: Example of didactic activity using TLFi for CE1 children

In this activity, children use the TLFi to discover that a single word can be associated to one or more definitions. The first step is to use the hyper-navigation system to select a particular word in the text. Then, thanks to the colouring process, children can mark the different definitions associated to this word. Eventually, he will be able to determine whether this word is associated to a single or several definitions. The results show that young children who are learning to read are able (alone or in pairs) to use the TLFi in order to write a word in capital letters, to tell its grammatical category, to find a phrase containing this particular word or to say whether it is monosemic or polysemic. As regards the modifications we could implement in order to get the TLFi more adapted to young learners, we could put examples coming from another textual database containing youth literature instead of literary examples from Frantext. Secondly, we intend to present the information in a slightly different way. For instance, the syntagms could be found at the beginning of a new paragraph, instead of being where it is not easy to detect. Thirdly, we intend to give online specific helps and commentaries to guide the young learner.

2.2 Uses in Secondary schools

The same methodology has been adapted to create didactic activities for secondary school pupils (children between 11 and 15 years old). Official texts edited by the National Department of Education insist on the fact that children have to master both French language and the use of the Internet. These two items are verified through an exam taken at the end of secondary school. As we have seen before, the TLFi can mix these two abilities and then meet the requirements of official syllabus, of teachers' and pupils' needs. Two experiments have been led in

2004 with two groups of 12-year-old children. These experiments have been made in order to define whether the graphical user interface of the TLFi allows usage in autonomy by children of this age without being helped by a technical sheet or by the teacher. Children had to look for the word "*oiseau lyre*" in two dictionaries (the TLFi and the TV5 dictionary available on the site http://dictionnaire.tv5.org/) in order to compare the definitions. They also had to give their impressions concerning the handling of these two tools. We can point out that pupils have found that the TLFi is very powerful and offers a big amount of information in its definitions; however, on the practical point of view, they had preferred a more easy to handle electronic dictionary (more rapid and less austere) to the TLFi. Furthermore, we have created a pedagogical activity for 14-year-old pupils based on the construction of French compound nouns and their plural. Pupils had to find the right spelling of compound nouns, to understand the sense of affixes and their behaviour. The last exercise for them was to identify grammatical rules and to write them down. However, the experiment has not taken place yet.

2.3 Uses for High school students

2.3.1 General overview

The environment LyText (Lycée + Textes) is now a part of the regional project called "e-Lorraine". The purpose of "e-Lorraine" is to give French teachers and high-school pupils a linguistic resource adapted to their needs. The resource, named LyText, will be available on the Internet, on the "e-Lorraine" website (www.e-lorraine.net). LyText has two main goals: helping pupils to prepare the French baccalaureate and giving a tool for the teachers, helping them to prepare their lessons as well as the exams taken by the pupils.

2.3.2 Various types of modules

Since it is dedicated to language teaching and learning, LyText is made up of three modules called "Texts visualization", "Texts preparation" and "Training". The module "Texts visualization" shows extracts of books associated to an informational model. These extracts can be the subject of a particular work in class such as a preparation for the oral test or the constitution of the corpus for the written test. In these two tests, the pupil has to answer one (oral exam) or several (written exam) questions about one (or several) extract. The pupil has to understand the text to answer these questions. He also has to interpret it. So as to facilitate his work, this module will propose him, for each text, to visualize an informational model made up of various pieces of information given by the text. For example, these elements are related to the textual structure, the lexicon or the stylistic devices in the text. The module "Texts preparation" makes it possible for the teacher to choose books belonging to the same period (a century, a particular date), to the same literary movement (romanticism, classicism, humanism, realism...) or to select books having the same genre (poetry, plays, autobiography...) or belonging to the same literary topic [Bouty, 1992]. He can then select the particular extracts on which he wants his pupils to work. Thus, the teacher will be able to define new groups of texts or to select new extracts of books studied for the oral test. This module also gives the teacher the possibility to modify (to remove or to add) information associated to each text in the informational model. For instance, with LyText, the teacher will be able to remove the presentation of a stylistic device like the zeugmas or the metaphors, to modify the lexical fields found in the text (to remove a word or a group of words belonging to a particular field). Eventually, the module "Training" gives the pupil the possibility to train himself to determine which pieces of information are relevant for his own analysis of the text. These pieces of information are those of the informational model presented in the module "Texts visualization". Thus, the pupil can choose to work on one or more extracts of books (which he already knows or not), to train himself to determine one or more types of information (lexical fields, stylistic devices, connectors...) that he chooses to seek inside this (or these) text(s).

2.3.3 Use of the TLFi in LyText

In LyText, the TLFi is used on two levels: the former is the user level, the latter is the designer level.

On the level user (in the module "Texts visualization"), the informational model gives the definition taken from the TLFi, for certain words of the text, which seems to be the most adapted. This definition helps the pupil understand the word in its context [Fayol, 2003]. In the module "Texts preparation", the teacher can determine the words of each extract, which he thinks they might pose a problem in understanding. Then, he can choose to give for each of these words one or more definitions extracted from the TLFi and/or various other information such as the etymology of the word. Finally, in the module "Training", the system will present the text and the possibility offered to the pupil to click on each word (or group of words). Then, he can pick the definition extracted from the TLFi, which seems to be the most adapted. The LyText environment can help the pupil in his decision-making. The system will be able to propose assistance to the pupil like brighter information in the TLFi (the field to which the definition belongs, the indicators of time [old, out-of-date], the register of language [popular, standard...], etc.

On the designer level, the concern is to use the TLFi within the framework of automatic processing. Two levels of processing are currently under development in the module "Texts visualization". First, it is a question of automatically determining in each text the definitions taken from the TLFi, which seem to be the most adapted according to the context (for certain words defined by the teachers like difficult to understand) [Véronis & Ide, 1990; Pélissier & Jacquey 2004]. Finally, in the module "Texts visualization", the informational model offers to visualize the various lexical fields of the text. We are currently determining a system that would automatically extract lexical fields (love, sadness, death...) using all pieces of information contained in each input of TLFi (especially indicators and fields).

In February 2005, we have experimented LyText in a French classroom in Lunéville (54), France. Two teachers have accepted to use LyText with their pupils. On the one hand, pupils studying literature have used LyText as a means to revise a text; on the other hand, for pupils studying science and technique, LyText has been used as a tool allowing the discovery of a new text, which eventually has led to a text commentary. We can point out that the construction of different interfaces for these two uses of LyText is compulsory. Indeed, the approach of a text commentary does not imply the same functionalities than for the revision of text analysis for the oral test.

Conclusion

Designed and produced first as a paper dictionary, the TLFi is nowadays a computerized dictionary in which each user, linguist, and researcher or not, can find different pieces of information. Concerning primary schools, children can find information to understand words and expressions. In secondary schools, pupils could understand orthographic problems when looking for historical information. Then, for high schools pupils as we have seen, the TLFi is not quite adapted: the TLFi is included in the LyText environment. The latter totally uses the capacity of the TLFi. Indeed, to show the most appropriate definition of TLFi according to the context, the system must use all information (date, indicator of language and meta-language, etc.). This work is currently in progress.

References

Bernard, P., Bernet, C., Dendien, J., Pierrel, J.-M., Souvay, G., Tucsnak, Z., 2001, "Un serveur de ressources informatisées via le Web", Actes de TALN-2001, Tours, Juillet 2001, pages 333-338.

Bernard, P., Dendien, J., Lecomte, J., Pierrel, J.-M., 2002. "Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella", Actes des 8^{e} Journées Internationales d'Analyse Statistique des Données Textuelles JADT 2002, Saint-Malo 2002, pages 137-149.

Bernard, P., Lecomte, J., Dendien, J., Pierrel, J.-M., 2002. "Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: TLFi, Frantext and the software Stella". Actes de LREC-2002, Las Palmas (Canaries).

Bouty, M. (1992). *Dictionnaire des oeuvres et des thèmes de la littérature française*. Hachette-Education, 2002, Paris.

Buzon, C., 1983, "Au sujet de quelques dictionnaires monolingues français en usage à l'école élémentaire", Christian Buzon, Etudes de Linguistique Appliquée, n°49, janvier-mars 1983, Didier Erudition, pp 147-173.

CNRS (1976-1993), Trésor de la langue française, dictionnaire de la langue du 19e et du 20e siècle, CNRS, Gallimard, Paris.

Fayol, M., Gaonac'h, D. (2003). "La compréhension, une approche de psychologie cognitive". In Gaonac'h, D., Fayol, M., coord., *Aider les élèves à comprendre, du texte au multimédia,* Profession Enseignant, Hachette, 2003, pp 5-72.

Gross, G., 1986, "Reconnaissance des emplois à l'aide d'un dictionnaire électronique", Etudes de Linguistique Appliquée, Janvier-Juin 1992, n°85 - 86, Didier Erudition, pp 89-97.

Gross, G. 1989, "Le dictionnaire et l'enseignement de la langue maternelle", Gaston Gross, Ein internationales Handbuch zur Lexicographie / An International encyclopedia of Lexicography / Encyclopédie internationale de lexicographie, Hausmann Franz Josef, Reichmann Oskar, Wiegang Herbert Ernst et Zgusta Ladislav Hrsg., (1989-1991), « Wörterbücher / Dictionaries / Dictionnaires », volume 1, article n°22, Berlin / New York, Walter de Gruyter, 1989, pp 174-180.

Lehmann, A., 2000, "Les dictionnaires pour enfants : diversité et uniformisation", Le Français Aujourd'hui – Construire les compétences lexicales, Alise Lehmann, revue trimestrielle, n° 131, revue de l'Association Française des Enseignants de Français, septembre 2000, pp 87-98.

Ministère de l'Education Nationale, 2002, *Programmes Français, classe de seconde,* Collection Lycée, voie générale et technologique, CRDP, 2002, Paris.

Nathan 1998, « Mon premier dictionnaire Super Génial Nathan », premiers apprentissages, plateformes Mac/PC, dès 3 ans, Pack « Nathan Benjamin Super Génial », Nathan, 1998.

Pélissier, C., 2002 : "Analyse de Mon Premier Dictionnaire", revue ALSIC (Apprentissage des Langues et Systèmes d'Information et de Communication), rubrique Analyse de logiciel, décembre 2002, volume 5, numéro 2, pp 269-286. Internet : http://alsic.u-strasbg.fr/Num09/pelissier/alsic_n09-log1.htm.

Pélissier, C., Jadelot, C., Pierrel, J.-M., 2004, "Méthodologie liée à l'Utilisation de Grandes Ressources Linguistiques dans le Cadre de l'Apprentissage : le cas du TLFi en Français au Cycle 3". EURALEX 2004, Lorient.

Pélissier, C., Jacquey, E., 2004, "Contribution d'un dictionnaire de référence informatisé dans un environnement didactique du français", Actes de la journée *TAL et apprentissage des langues,* Grenoble.

Pierrel, J.-M., Dendien, J., Bernard, P., 2004, "Le TLFi ou Trésor de la langue française informatisé". EURALEX 2004, Lorient.

Pruvost, J., 2001, "Les dictionnaires d'apprentissage monolingues et langue française (1856-1999) : problèmes et méthodes", Jean Pruvost, Les dictionnaires de la langue française, sous la direction de Jean Pruvost, Honoré Champion, Paris, 2001, pp 67-96.

Véronis, J., Ide, N., 1990, "Word sense disambiguation with very large neural networks extracted from machine readable dictionaries", Proceedings of the 14th International Conference on Computational Linguistics (COLING'90)

The dictionary of Polish of the 16th century and the computer: from paper to (structured) file.

TADEUSZ PIOTROWSKI [1], KRZYSZTOF SZAFRAN[2]

[1] English Department, Opole University, Kopernika 11, 45-051 Opole,
tadpiotr@plusnet.pl
[2] Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warszawa,
kszafran@mimuw.edu.pl

Abstract

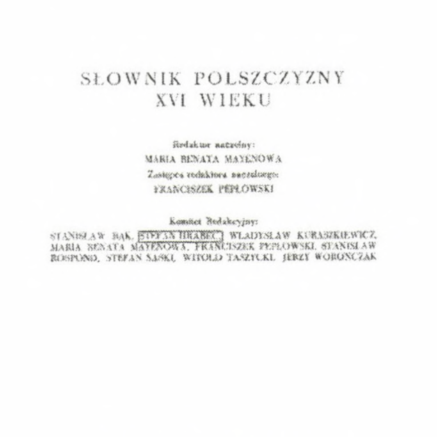
The paper discusses the dictionary of Polish of the 16th century (*Słownik polszczyzny XVI wieku*) that has been compiled since the 1960's, describes the dictionary, perhaps the most ambitious dictionary project in Poland, and shows the stages in which it goes from the paper edition to the electronic one, i.e. from hot type to cold type. The paper focuses then on the project, of the authors, to digitize the whole dictionary, to convert it into an electronic version that would conform to TEI, by the use of typesetting codes. It shows samples of the dictionary at the successive stages and discusses the difficulties that the project faces.

Introduction: the dictionary

In Poland the 1940's was the period in which the compilation of historical period dictionaries of Polish started: the dictionary of Old Polish, a dictionary of Polish in the sixteenth century, there were also plans of a dictionary of the 17th and the first half of the 18th c. The first-mentioned dictionary, *Słownik staropolski*; edited by Urbańczyk, on which work had actually begun at the start of the twentieth century, was completed in the 2000's. The last-mentioned one (Siekierska 1999-) has just started to appear. This paper will be concerned with a dictionary of Polish in the sixteenth century, *Słownik polszczyzny XVI wieku* (henceforth *SXVI*; Mayenowa & Pepłowski 1966—); work on it started in 1949, and it has reached its half-way point.

SXVI is perhaps the most ambitious dictionary being produced in Poland; it is also one of the largest. This is because of the importance of the 16th century in the development of Polish and Polish literature, and because of the methods used in the compilation. The period is usually called the Golden Age of Polish literature, and this is related to the history of the language: it was then that standard Polish evolved, which in that century rapidly developed into a vehicle of sophisticated literature, which, in the texts of Jan Kochanowski, attained quality that has been rarely matched in the following five hundred years. That was also a period of growth of the lexicon and of lexical exuberance, the number of lexical items is several times as high as that in the preceding period.

Figure 1 Dictionary of Polish of the 16th c. Title page.



POLSKA AKADEMIA NAUK
INSTYTUT BADAŃ LITERACKICH

SŁOWNIK POLSZCZYZNY
XVI WIEKU

Redaktor naczelny:
MARIA RENATA MAYENOWA
Zastępca redaktora naczelnego:
FRANCISZEK PEPŁOWSKI

Komitet Redakcyjny:
STANISŁAW BĄK, STEFAN HRABEC, WŁADYSŁAW KURASZKIEWICZ, MARIA RENATA MAYENOWA, FRANCISZEK PEPŁOWSKI, STANISŁAW ROSPOND, STEFAN SAWICKI, WITOLD TASZYCKI, JERZY WOROŃCZAK

WROCŁAW · WARSZAWA · KRAKÓW · GDAŃSK
ZAKŁAD NARODOWY IMIENIA OSSOLIŃSKICH
WYDAWNICTWO POLSKIEJ AKADEMII NAUK

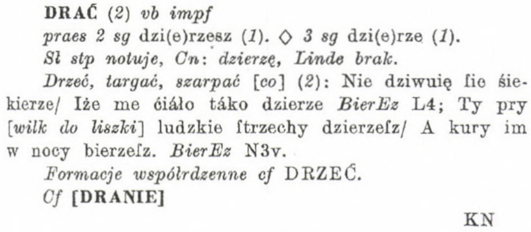
The publisher of the dictionary is the Institute for Research into Literature, Polish Academy of Sciences (Instytut Badań Literackich Polskiej Akademii Nauk, IBL PAN), its begetter and first editor was Maria Renata Mayenowa, after her death the work has been carried on by Franciszek Pepłowski. Initially the work was done at several academic centres, now two are still operational, the main dictionary office is in Toruń, and the smaller one is in Wrocław. The first volume of *SXVI* appeared in 1966, thirty-one volumes were published until 2004, reaching P (entry *przemieść*). The number of copies that are printed is usually around 1,000.

The dictionary, when finished, will have from 50 to 60 volumes. It is not strictly a historical dictionary but actually a synchronic dictionary of one period. Initially the editors attempted to include all words found in texts that form the canon (selected by experts) from the 16th century. This undertaking proved to be unfeasible: the 35 most frequent words have more than 15,000 citation slips each, therefore the attempted completeness has been abandoned and the dictionary now uses a very broad reading programme The editors collected 8,000,000 citation slips, which were sorted alphabetically and are stored in one place. It is believed that they would yield about 100,000 entries.

An individual entry includes variants, definitions as well as synonyms to particular senses, shows not only established lexical items but also recurrent word combinations, classified semantically and syntactically, naturally it includes copious citations, therefore one large-size volume of the dictionary, with 500-600 pages, has only about 1,150 entries. High-frequency items, for example prepositions, can cover several dozen pages (*przed* has 40 pages), those entries have an intricate mutlilevel structure, up to six levels, and their own table of contents, which, however, only identifies the headings, and does not point to relevant page numbers.

The synchronic character of the dictionary can be also seen in the fact that the senses are not arranged chronologically and that the dates of citations are not given very consistently. On the other hand, spelling in the citations has not been modernized, therefore a number of characters are used that are no longer used in modern typesetting. The dictionary is the first in Poland to show the frequency of words and their senses in texts. The extremely short entry (from a hot-typeset volume), in Fig. 2 shows not only the basic structure of the entry but also some of the special characters found in citations.

Figure 2 Example: A short entry.



DRAĆ (*2*) *vb impf*
praes 2 sg dzi(e)rzesz (*1*). ◇ *3 sg* dzi(e)rze (*1*).
Sł stp notuje, Cn: dzierzę, Linde brak.
Drzeć, targać, szarpać [*co*] (*2*): Nie dziwuię ſie śiekierze/ Iże me ćiáło táko dzierze *BierEz* L4; Ty pry [*wilk do liszki*] ludzkie ſtrzechy dzierzeſz/ A kury im w nocy bierzeſz. *BierEz* N3v.
Formacje współrdzenne cf DRZEĆ.
Cf **[DRANIE]**

KN

From hot type to cold type.

Obviously, the dictionary was first set in hot type, to be exact, probably the first twenty one volumes, until as late as 1995, were typeset and printed traditionally, in hot metal, from typescript prepared by the editorial staff. Unfortunately we do not have too much information on how actually this was going on (how the archaic characters were represented, for example). The text in those volumes is available only on paper.

Later, between 1995 and 2002, i.e. for volumes from XXII to XXX, all work was carried out by the staff: compilation, editing of the text and typesetting. The printing shop received the dictionary as ready-to-print files. For typesetting a computer system, developed in Poland, was used, Cyfroset 2, which came with its own editing application, Mini Cyfroset 1.0; the software used its own character coding. Unfortunately, at that time the printed paper version was treated as the final product, and the computer files were treated as technological waste, which can be discarded after the proper job had been done. Therefore we have only files for volumes XXIV and XXVII–XXX, and the text in the remaining four volumes is available only on paper. As far as we know, Cyfroset is still used, version number 6 (http://www.cyfronex.neostrada.pl/produkty_dtp.htm).

Starting from volume XXXI (i.e. from the year 2002) the dictionary is typeset by a more recent phototypesetting system, called KOMBI (http://www.3n.com.pl/), also developed in Poland; one of its advantages is certainly the low cost but also the flexibility and willingness of the producer to customize the system to the unique needs of the project. The KOMBI system supports Unicode (utf-16 character coding system), and uses a kind of text markup, which, however, unfortunately does not conform to any standard. As can be expected, the tags are used only to mark typographical distinctions, their set is fixed, that is, new tags cannot be added, in the mark-up two types of tag are used: tag and subtag. Below there is an example.

Figure 3 KOMBI mark-up: start of entry *przed*
<LP> (1)

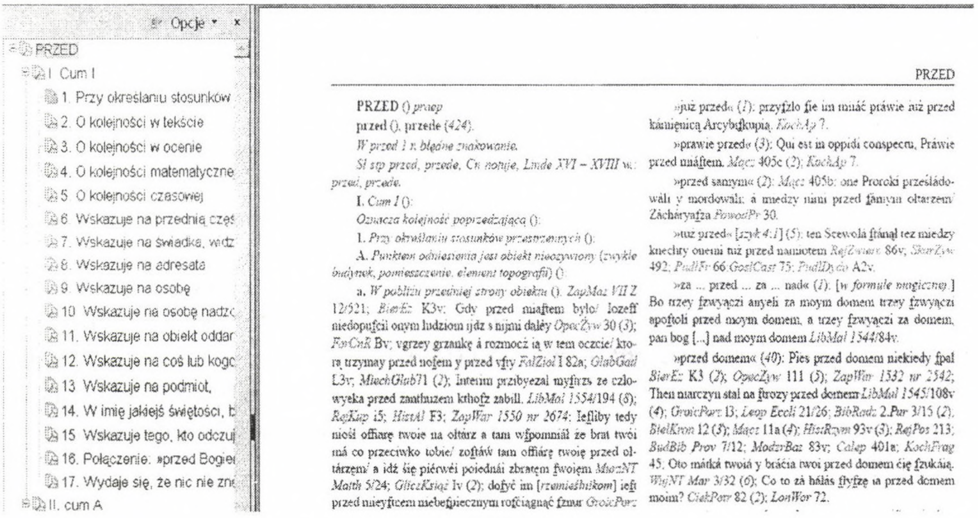
<S/-F/-I/B>PRZED<S/-F/-I/-B>
<LP> (1)
(<S/-F/I/-B><S/-F/-I/-B>) (2)
<S/-F/I/-B>praep (3)
<A/HL/J:3/S> (4)
<S/-F/-I/B>przed<S/-F/-I/-B> (3)
(<S/-F/I/-B><S/-F/-I/-B>),
<S/-F/-I/B>przede<S/-F/-I/-B>
(<S/-F/I/-B>424<S/-F/-I/-B>).

As with most typographical home-made tags, there are some problems when using the text for processing.

1. Sometimes there is no distinction between a starting and an ending tag. In example 2 <LP>, which is used like that, means the running header.
2. The number of starting and ending tags can be different in a file. In example 2 there is an uneven number of –B's (which means bold).
3. Sometimes the ranges of subtags overlap.
4. For some types of tag (e.g. new paragraph) there are no ending tags (or they are not explicit).

However, even with these problems, KOMBI files can be fairly easily converted to other formats, for example into Adobe Systems' Portable Digital Format (PDF), or, with some more difficulty, into a structured standard mark-up format, like SGML/XML. For this particular paper perl and emacs were used by Krzysztof Szafran for string manipulation, and LaTeX for actual conversion into PDF. Below there is an example of a PDF file (entry *przed*).

Figure 4 PDF file: entry przed



PRZED
I Cum I
1. Przy określaniu stosunków
2. O kolejności w tekście
3. O kolejności w ocenie
4. O kolejności matematyczne
5. O kolejności czasowej
6. Wskazuje na przednią częś
7. Wskazuje na świadka, widz
8. Wskazuje na adresata
9. Wskazuje na osobę
10. Wskazuje na osobę nadzc
11. Wskazuje na obiekt oddar
12. Wskazuje na coś lub kogc
13. Wskazuje na podmiot,
14. W imię jakiejś świętości, b
15. Wskazuje tego, kto odczuj
16. Połączenie: »przed Bogie
17. Wydaje się, że nic nie zn
II. cum A

PRZED

PRZED () *praep*

przed (), przede (*424*).

W przed 1 r. błędne znakowanie.

Sł stp przed, przede, Cn notuje, Linde XVI – XVIII w.: przed, przede.

I. *Cum I* ():

Oznacza kolejność poprzedzającą ():

1. *Przy określaniu stosunków przestrzennych* ():

A. *Punktem odniesienia jest obiekt nieożywiony (zwykle budynek, pomieszczenie, element topografii)* ():

a. *W pobliżu przedniej strony obiektu* (). *ZapMaz VII Z 12/521*; *BierEz* K3v: Gdy przed miaſtem było/ Iozeff niedopuſcił onym ludziom idź s nimi daléy *OpecŻyw* 30 (*3*); *FalCnR* Bv; vgrzey grzankę á rozmocz ią w tem oczcie/ ktorą trzymay przed noſem y przed vſty *FalZioł* I 82a; *GlabGad* L3v; *MiechGlab*71 (*2*); Interim przibyezal myſirzs ze czlowyeka przed zamthuzem kthoſz zabill. *LibMal 1554*/194 (*8*); *RejKup* i5; *HistAl* F3; *ZapWar 1550 nr 2674*: Ieſliby tedy nioſł offiarę twoie na ołtarz a tam wſpomniał że brat twói má co przeciwko tobie/ zoſtáw tam offiárę twoię przed ołtárzem/ a idź śię pierwéi poiednái zbratem ſwoięm *MurzNT Matth 5/24*; *GliczKsiąż* Iv (*2*); doſyć im [*rzemieśhnikom*] ieſt przed mieyſcem niebeſpiecznym roſciągnąć ſznur *GroicPorz*

»już przed« (*1*): przyſzło ſie im mináć práwie iuż przed kámienicą Arcybiſkupią. *KochAp* 7.

»prawie przed« (*3*): Qui est in oppidi conspectu, Práwie przed miáſtem. *Mącz* 405c (*2*); *KochAp* 7.

»przed samym« (*2*): *Mącz* 405b: one Proroki prześládowáli y mordowáli: á miedzy nimi przed ſámym ołtarzem/ Záchárynſza *PowodPr* 30.

»tuż przed« [*szyk 4:1*] (*5*): ten Scewolá ſtánął tez miedzy knechty onemi tuż przed namiotem *RejZwierz* 86v; *SkarŻyw* 492; *PudlFr* 66;*GostCast* 75; *PudlDyc* A2v.

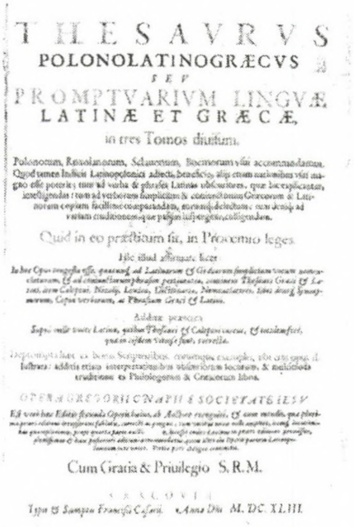
»za ... przed ... za ... nad« (*1*): [*w formule magicznej*.] Bo trzey ſzwyączi anyeli za moym domem trzey ſzwyączi apoſtoli przed moym domem, a trzey ſzwyączi za domem, pan bog [...] nad moym domem *LibMal 1544*/84v.

»przed domem« (*40*): Pies przed domem niekiedy ſpal *BierEz* K3 (*2*); *OpecŻyw* 111 (*5*); *ZapWar 1532 nr 2542*; Then marczyn stal na ſtrozy przed domem *LibMal 1545*/108v (*4*); *GroicPorz* I3; *Leop Eccli 21/26*; *BibRadz 2.Par* 3/15 (*2*); *BielKron* 12 (*3*); *Mącz* 11a (*4*); *HistRzym* 93v (*3*); *RejPos* 213; *BudBib Prov* 7/12; *ModrzBaz* 83v; *Calep* 401a; *KochFrag* 45: Oto mátká twoiá y bráćia twoi przed domem ćię ſzukáią. *WujNT Mar* 3/32 (*6*); Co to zá hálás ſlyſzę ia przed domem moim? *CiekPotr* 82 (*2*); *LorWor* 72.

Digitization of the dictionary

It is more and more strongly felt that historic and historical dictionaries of Polish should be available not only on paper but also in electronic form. It was this conviction that drove some people to undertake the actual work of digitization, Tadeusz Piotrowski and Krzysztof Szafran (with other scholars) submitted a grant proposal to digitize the major dictionaries of Polish (cf. Piotrowski 2005), which included also *SXVI*, for which Szafran was to be chiefly responsible. The project was perhaps felt to be too ambitious, and the proposal was not successful. Fortunately, a university library (Warsaw University) decided to cooperate with linguists in this respect, to be precise, with Professor Włodzimierz Gruszczyński, the present editor of the dictionary of the 17th century, and to scan some of the historic dictionaries. They started with the dictionary of Knapski (*Thesaurus Polono-Latino-Graecus*; 1643-1644), the most important Renaissance/Baroque dictionary, with perhaps as many as 40,000 entries (its resources are being used in *SXVI*; cf. abbreviation *Cn* in Figure 2). They plan to include other valuable historic dictionaries, especially those that are still used by researchers.

Figure 5 Knapius: Thesaurus



THESAVRVS
POLONOLATINOGRÆCVS
SEV
PROMPTVARIVM LINGVÆ
LATINÆ ET GRÆCÆ,
in tres Tomos diuisum.

Polonorum, [illegible] Sclauorum, [illegible] vsui accommodatum. Quod tamen Indicis Latinopolonici adiecti beneficio, aliis etiam nationibus vsui magno esse poterit; [illegible]

Quid in eo præstitum sit, in Proœmio leges.

[illegible]

Cum Gratia & Priuilegio S. R. M.

CRACOVIÆ

Typis & Sumptu Francisci Cæsarii. *Anno Dni* M. DC. XLIII.

Apart from those dictionaries, in which form IS content, so to say, and which would be extremely difficult to convert into text, there are also modern ones, especially those that have not been completed, and which in future should be available not only on paper but also on computer. This way not only more interested people could use the dictionaries, especially when they would be available on the Internet, but also text would be fully usable. Among those dictionaries two historical dictionaries are first of all of interest: the dictionary of the 17th century, which has just started appearing, therefore it can be produced both on paper and on computer, and *SXVI*, which reached its half-way point, therefore it is still possible at least to produce the volumes that are to be compiled in electronic form.

Some work has already been done, and at present we have the following items available in a digitized form: the editorial instruction, lists of various resources used in compilation: list of abbreviations, with more than eight hundred items, in three separate lists: list of the texts, list of additional texts, list of other abbreviations (mainly grammatical). The editorial instruction is an interesting document, interesting for historians of lexicography and for researchers into the history of Polish. What we have is most probably the final version of a number of successive documents, as the year of publication given is 1976 (and the first volume appeared in 1966). It is a large typewritten document, with 328 pages. The quality is very low, in the text traditional attributes of the text are used (underscoring, and the like), which makes it almost impossible to be interpreted adequately by an OCR application. Therefore it was scanned as a graphic image, and was converted into a number of formats, DjVu being one of the most convenient one because of its small size. One volume was converted into a PDF format, which shows that conversion is certainly feasible.

There are some suggestions that slips should be scanned, both for needs of their safety and because of their value. Some researchers (for example Piotr Żmigrodzki, in print) suggest that the citations are the most valuable element in the dictionary, and that the metalinguistic descriptions, for example syntactic, on which much effort has been expended, do not quite fulfil the needs of the contemporary researcher. However, the quantity and quality of the slips: low-quality paper, barely legible typescript, inclusion of important handwritten notes, as well as the fact that they are constantly in use, and are transferred between the two dictionary centres, make this undertaking not only difficult but also financially prohibitive. Perhaps it would be cheaper and more convenient to compile a corpus of Renaissance texts instead (a very imperfect text collection of this sort is available at http://monika.univ.gda.pl/~literat/books.htm).

The main task now is to study the structure of the dictionary, and to represent it formally as a DTD. This is what Dr Szafran is doing at the time being. There are two approaches possible: either to treat the structure superficially, to describe only the divisions of structure and main types of structural units that can be found in the dictionary. The structure will be treated as a framework into which data are put, and no regard is paid to the validity of the type of data and purposefulness. That would be the easier solution, and more flexible, as that would allow one to easily accommodate what there already is in the dictionary, notwithstanding any individual differences between particular volumes or entries. In the other approach there would be some effort at understanding the decisions of lexicographers to describe the linguistic data by including it in some type of structural division, or by giving it some status in the entry. Hopefully, one result would be that the dictionary could then be made more consistent, both in the volumes already produced (when digitized) and in the volumes that are to be written (especially if some sort of editing software would be produced that would allow for description of data only in a permitted way). However, that might just as well be unrealistic: dictionary-making is an art and the individual decisions of the editors might not be compatible over long stretches of the dictionary. Moreover, it is uncertain whether that sort of analysis is possible only at the level of structure, without re-doing the linguistic analysis of Renaissance Polish texts. The choice of the approach is crucial now in the development of DTD, and Dr Szafran is busy studying the highly complex structure of the whole dictionary.

There are many difficulties with digitizing the dictionary. While Dr Szafran has the full cooperation of the editors, the chief problem with the authorities is the misinterpretation of the possible use, and possible commercial exploitation, of the results of digitization, which mirrors the current disputes over the allowed use of various digital resources. While it is understandable that a dictionary of contemporary Polish can be exploited commercially, one has some doubts whether that sort of thinking can be extended to a dictionary accessible to specialists only. This, however, certainly sheds some interesting light on some mental problems with use of digital resources and with digitization.

List of references

Knapski Grzegorz : *Thesaurus Polonolatinograecus seu Promptuarium Latinae et Graecae...* Ed. 2, Kraków : Fr. Cezary 1643-1644 (*Thesaurus* Grzegorza Knapskiego: reprodukcja cyfrowa, Warszawa: Bilbioteka Uniwersytecka 2004; editors W. Gruszczyński and Marek Kunicki-Goldfinger)

Mayenowa, Maria Renata & Franciszek Pepłowski, editors-in-chief. 1966—. *Słownik polszczyzny XVI wieku* [Dictionary of Polish in the 16th c.]. Wrocław: Ossolineum. (From 1995 published by Instytut Badań Literackich, PAN, Warszawa.).

Piotrowski, Tadeusz. 2001."Lexicography in Poland: From Early Beginnings — 1997". *Towards a History of Linguistics in Poland. From the Early Beginning to the End of the Twentieth Century.* Red. E. F. K. Koerner, Aleksander Szwedek. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 101-122 (also in 1998. "Lexicography in Poland: From Early Beginnings — 1997", *Historiographica Linguistica XXV*, 1/2: 1-24)

Piotrowski, Tadeusz. 2005. "Digitization of Polish Historic(Al) Dictionaries", *Преглед НЦД* (Review of the National Center for Digitization) *6:* 4: 95–102 [also available at http://www.komunikacija.org.yu/komunikacija/casopisi/ncd/6/index_e]

Siekierska, Krystyna, et al., eds. 1999-. *Słownik języka polskiego XVII i 1. połowy XVIII wieku*. Kraków : Wydaw. IJP PAN

Szafran, Krzysztof. 2004. "Ku elektronicznej wersji Słownika polszczyzny XVI wieku", *Bulletin de la Société Polonaise de Linguistique*. *LX*: 89-97

Urbańczyk, Stanisław, editor-in-chief. 1953—2004. *Słownik staropolski* [Dictionary of Old Polish]. Wrocław: Ossolineum. (From 1991, Fasc. 64, published by Instytut Języka Polskiego PAN, Kraków.)

Developing Dictionary Databases as Lexical Data Bases[1]

F. Sáenz and A. Vaquero

Departamento de Sistemas Informáticos y Programación,
Facultad de Informática, Universidad Complutense de Madrid,
E-28040 Madrid, Spain
{fernan, vaquero}@sip.ucm.es

Abstract

We propose to apply classical development methodologies to the design and implementation of Lexical Databases (LDB), which embody conceptual and linguistic knowledge. We represent the conceptual knowledge as an ontology, and the linguistic knowledge, which depends on each language, in lexicons. Our approach is based on a single language-independent ontology. Besides, we study some conceptual and linguistic requirements; in particular, meaning classifications in the ontology, focusing on taxonomies. We have followed a classical software development methodology for implementing lexical information systems in order to reach robust, maintainable, and integrateable relational databases (RDB) for storing the conceptual and linguistic knowledge. The result is a methodology to develop information systems for building and querying LDB (SV 02). Based on this methodology, we have developed software tools for authoring and consulting different kinds of linguistic resources: monolingual, bilingual and multilingual dictionaries. Conventionally, dictionaries are conceived for human use and lexical databases are conceived for natural language processing (NLP) applications. Our methodology leads to friendly usable dictionaries, but structurally prepared to be easily embedded in computer applications, as we show along the paper.

1 Introduction

Due to the immaturity of the knowledge representation topic, lack of standardization is broadly felt as a very undesirable state into the community around language resources (LREC 02). For instance, standard terminology for a common reference ontology is yet a goal to be reached. No doubt about what lexicon means, but ontology is differently understood in the computational linguistic literature. For instance, WordNet is mentioned as an ontology (USC 96), CYC is provided with a formal ontology (PRI 01), etc. Here, ontology, in a LDB, is the set of con-

[1] This work has been partially supported by the Spanish CICYT project number DPI2002-02924.

cepts in the domain of the base and the relationships that hold among them, without including linguistic knowledge, and common to all of the languages supported in the base.

Weak attention has been paid on topics about development methodologies for building the software systems which manage LDBs, and dictionaries in particular. We claim that the software engineering methodology subject is necessary in order to develop, reuse and integrate the diverse available linguistic information resources. Really, a more or less automated incorporation of different lexical databases into a common information system, perhaps distributed, requires compatible software architectures and sound data management from the different databases to be integrated. The database subject have already done a long way reaching a strong standardization, and supplying models and methods suitable to develop robust information systems. We apply RDB design methodologies to develop LDB consisting of ontologies and lexicons. The conceptual knowledge is represented as an ontology, and the linguistic knowledge, depending on each language, is stored in its lexicon.

Subjects about electronic dictionaries for diverse natural language processing applications have been extensively studied (ZOC 03), (WIL 90), (WIL96), as well as LDB (MIL 95), world knowledge bases (LEN 90), ontologies in general (ONT), ontologies for computational linguistics (NIR), and the like. But there are no references on how these information systems have been developed and upgraded along their life. Moreover, tools for managing ontology-based linguistic information systems have been described (MOR 02), but there is no a declared software engineering approach for the development of these tools.

We follow the classical RDB design based on the conceptual, logical, and physical models for building LDB, and software engineering techniques based on UML for building LDB interfaces (which are not described in this paper).

2 Conceptual and Linguistic Requirements

Conceptual and linguistic knowledge incorporated in computing systems devoted to NLP are relevant in the definition of a conceptual model for LDBs. Regardless of the language, the knowledge in the discourse universe is conventionally divided in two classes: conceptual and linguistic. Terms and sentences refer to concepts, but they have particular structural and morphological features in each language. All of this information is not available in any dictionary, electronic or not, although it is the objective in the most exigent ontology-based linguistic Knowledge Bases, such as MikroKosmos (MIK). (SV 02) provides more linguistic requirements we are interested in.

2.1 Lexical Databases

For a given language, we have a set of terms, meanings and categories holding certain relationships among them. Conventional LDB, such as WordNet (MIL 95), have term classification through synonymy (grouped in the so-called synsets). LDBs based on ontological semantics go beyond by playing the role of meaning taxonomy and supporting more complex semantic relationships (NIR 95). All of the relationships (meronymy, holonymy, hypernymy, hyponymy, and so on) represented in the more complete lexical databases, such as WordNet or EuroWordNet (EWN), are also represented in ontology-based databases, such as MikroKosmos; but in this case, all of the concepts and their relationships are present in the ontology, while each lexicon has the terms for each language and their linguistic arguments, as well as the links with the concepts into the ontology. The mapping between ontology and lexicon is the key for successfully coordinate all of the lexical and semantic relationships. This approach does full separation between ontology and lexicon. If we now think of several languages, the same ontology applies for each one of the lexicons.

Any other approaches has been adopted. Each one of them leads to a more or less complex LDB structure. We claim for the approach ontology-lexicons as the most appropriated to reach a simple, robust and controlled LDB structure, prepared to be reused in different applications and integrated with another ones with the same structure.

The architecture ontology-lexicons is criticized in (POL 03), given that each language has its own lexical semantics. Then, strictly speaking, there is no one single ontology independent of the considered languages. In favor of our position, we argument that the fact of the nonexistence of one single ontology common to diverse languages is independent of assuming one imposed undesirable a priori hierarchy, which is considered in (POL 03) as unavoidable considering the common ontology approach. But in our methodology, the hierarchy (taxonomy) is incrementally created when building the LDB. For a monolingual database (French in the case of the DiCo LDB), there is only one ontology; thus, there is no problem. However, certain problems could arise in multilingual LDB, because the boundary between ontology and lexicon does not appear clearly always. There are many ways to face up these problems considering other approaches different from ours, when the ontological semantics is distributed among the different languages at multiple levels. For instance, in the Papillon project (MAN 03), the different languages are linked to a common dictionary of meanings (axies in French). In the EuroWordnet project, the different WordNets (one for each considered language) are linked by two levels of common concepts, and the resulting structure is not appropriated for the multilingual applications. In MILE (ABB 02), SIMPLE templates play the role of ontologies; so the resulting LDB structure is more complex than that resulting from the approach ontology-lexicons.

We adhere to the criterium from (MAH 95) conceiving ontology as a language-neutral body of concepts. In this case, the problems can be solved putting in each specific lexicon the own lexical-semantic information required, which is not present in the common ontology (VIE 98); so the ontology is the conceptual model of the domain and each lexicon is linked to the same ontology. From this approach, the system design to develop LDB is enhanced in robustness, because an architecture with two abstraction levels is reached.

From this approach we apply very carefully the RDB techniques to reach a methodology assuring a sound and simple structure of the LDB, and a controlled way for building any particular LDB through an administration interface. This work is indeed previous to the formal definition of an interlingua (FAR 04). We are far from reaching this goal, but there are a lot of NLP applications, not only monolingual ones, that do not need formally and completely represent the text meaning. We claim for reaching an interlingua in the future from LDB conceived from the ontology-lexicons approach and developed with our methodology.

Another central idea in this work is to develop for each group of applications one LDB, the most appropriated one. Certain applications are more exigent of linguistic resources than other ones. Why to use the same LDB for no matter what application?. This vision contemplates, besides our methodology to build different LDB, building subsets of LDB already build as 'views' of the DB; in this case the LDB has to have been developed from the ontology-lexicons approach. We claim for this way in order to integrate different LDB.

2.2 Our LDB for Dictionaries

In this approach, relationships among terms from different languages come from considering jointly the involved Ontology-Lexicon schemes, as we will see later when considering the bilingual dictionary. In the dictionary here considered, the ontology only consists of one relationship which gives tree-structure to the conceptual taxonomy. A taxonomy is a natural structure for meaning classification. Each node in the taxonomy corresponds to a category. In principle, every category in the taxonomy can have meanings, regardless of its taxonomy level. It must be noted that every category in the taxonomy contains at least the term which names the category, so that all categories are non-empty. On the other hand, the creation of new categories as belonging to several predefined ones should be avoided, in order to reach a compact relationship as the taxonomy structuring backbone. We have developed dictionaries without overlapped classifications (RK 02), and only permitting tree-structured taxonomies. Since a meaning can belong to different categories, the extensional definition of categories is hold (SV 02).

When consulting or building dictionaries, there are a number of advantages in classifying meanings as taxonomies. First of all, meaning taxonomy is a useful facility for an electronic dictionary, because meaning classification embodies additional semantics, which provides more information to the user than usually provided. As long as we know, this kind of facilities (meaning classification), normally used in conceptual modeling through ontologies (MCG 00), has not been implemented before into dictionaries.

3 Conceptual Model of the Terminological Database (TDB)

There are different TDBs built for different purposes. Some of them have incorporated the ontology structure, and so, they could possibly be used for the pedagogical goals proposed above. But there are a lot of difficulties when intending to do this, not being the less the fact that these very large databases are yet complete or almost complete. So only the tools for building terminological databases are needed. Moreover, the development of this kind of tools must be made taking into account the pedagogical goals which have not been the case of the LDB already built.

Our work in developing the tools is based on a sound conceptual model for the terminological database which shall eventually hold the terms, definitions, meanings, and semantic categories. Since it is intended to deal with two or more languages (bilingual or multilingual dictionaries), we need to represent instances of terms, textual definitions, and textual semantic categories for each language, but, as meanings are not language dependent, we'll use unique representations for them.

The entity-relationship model is used to describe the conceptual model we propose, shown in figure 1. In this figure, entity sets are represented with rectangles, attributes with ellipses, and relationship sets with diamonds connecting entity sets with undirected lines (many to many mapping cardinality). Undirected lines also connect attributes to entity sets. Relationship set and entity set names label each diamond and box, respectively.

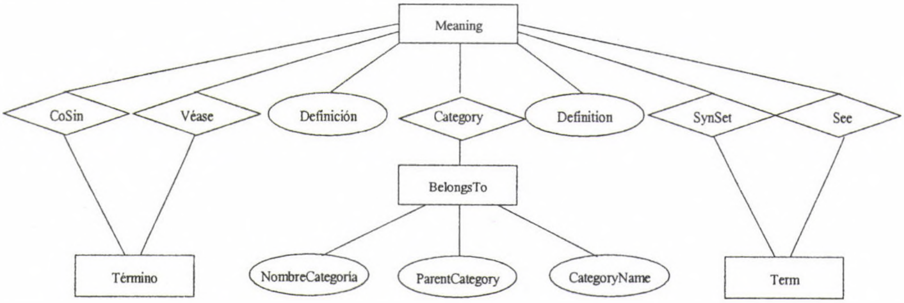


Figure 1. Entity-Relationship Model for an English-Spanish TDB

For the sake of clarity and conciseness, in this figure we show an instance of a multilingual terminological database for only Spanish and English languages, although we have extended for multilingual support (SV 02). The entity set Meaning is the central entity set other entity sets rest on. In fact, this is the entity set which is language independent. The relationship set SynSet denotes the English synonym set. The entity set Term represents all the English terms that compose the terminological database. The relationship set between Meaning and Term is many to many since a synonym set contains several terms, and a term may be contained in several synonym sets (obviously, with different meanings.)

The relationship set See denotes the set of English terms related under a given meaning. This relationship which connects Meaning and Term is many to many because a meaning may refer to several English terms, and one term may be polysemic. The entity set Category denotes the category each meaning belongs to. The relationship set BelongsTo between Category and Meaning is many to many since many meanings are in a category, and a meaning could be in several categories (this situation is expected to be reduced to the minimum since the goal is to keep the classification as disjoint as possible). This relationship set embodies the fact that our classification is not lexical (there is not a direct relationship between Category and Term) but semantic (we relate meanings to categories, i.e., we categorise meanings.) The entity set Category has three attributes: CategoryName, NombreCategoría, and ParentCategory. The first two correspond to the textual name of the category in each considered language, English and Spanish, respectively. The last attribute, ParentCategory, represents the links in the taxonomy by relating a category with its parent. Since each entity Category has a monovalued attribute for parent, this means that we restrict taxonomies to trees. If we change this attribute by a multivalued attribute (or, alternatively, we connect the entity set Category with itself via a relationship set named ParentCategory), we allow a taxonomy graph instead of a tree. Meaning has two attributes: Definition and Definición, which correspond to the textual definition in the same considered languages. The remaining entity and relationship sets (CoSin, Véase, Término) are homologous to the ones in the other language (SynSet, See, Term.)

The logical and physical models for the development of any terminological database following the principles above expressed have to be based on this conceptual model.

4 Functionalities of the Tools

4.1 The User Tool

We have developed a user tool, a query interface which allows us to easily recover the information about both English and Spanish terms as well as their relationships from the so-called terminological database. This database holds the terms, categories, their attributes, and the relationships. The interface allows the user to navigate the semantic categories, also allowing to retrieve the relevant information of any term (definition, other related terms, translation, synonyms, ...) as shown in (SV 02).

The Start window of this tool allows the user to select the base language (i.e., the source language for translations and for representing dialogues) among the available languages by pressing its button (from now on, we consider a bilingual dictionary so that it is unnecessary to select the source language or the target language.)

This action pops up the Semantic Category window; its left pane shows the semantic categories structured as a tree, and the right pane, all the words under the highlighted semantic category. The total number of terms is showed on top of the right pane. The nodes in the tree can be clicked in order to expand or contract semantic categories subtrees. A text box is used for term lookups so that the closest word to the substring typed is shown in the right pane. Pressing Enter or double-clicking the highlighted word yields to the Query window. This window shows the relevant information about the selected term: its definition, comments; the list of semantic categories it belongs to (the one corresponding to the shown definition is highlighted), the synonym set and the list of related terms. It also displays a navigation history. It is possible to select another semantic category in this window, which results in updating all the relevant information. Direct access to the terms in both the synonym and related terms windows is allowed by double-clicking.

The Semantic Category window has a control box with buttons to activate the return to the Start window, navigate backwards, translate the selected word, print, and exit the interface. The Translate button offers one of the main functionalities of this interface, i.e., the translation from the (source) base language to the target language and, when pushed, it pops up the Translation window. This window shows a first field for the term in the first language, and a second field for the term in the second language. There are also navigation buttons for searching other terms in the same semantic category under an alphabetical order. It is possible to translate from the first or from the second language by using two buttons which express the two possible translation directions. Also, the Go to buttons allow us to go to the Semantic Category window for the selected term. This completes the overall description of the functionalities of the user tool.

4.2 The Author Tool

The author tool allows the author to add new terms to the terminological database, and all the relevant information, such as its definition, semantic categories, meanings, synonym sets, and related terms. We have developed a Spanish user interface for this tool (easily rewritable for allowing to customise the use of any other language), and it consists mainly of one Author window. It has several management areas which are explained next.

Semantic Category Management Area This area is intended for managing all the operations related to semantic categories. It has several controls: a hierarchical view of the semantic categories (with expand/collapse functionality), text fields for the semantic category names (English and Spanish), and the buttons Add Category, Delete Category, and Modify Category. The insertion point when adding a new semantic category is the highlighted semantic category, and the Spanish and English texts for the semantic category name must be typed in the aforementioned text fields.

Meaning Management Area The area for meaning management consists of two lists for the meanings in both languages and the buttons Add, Delete, and Modify for addition, deletion, and modification of meanings, as well as buttons for edition (Copy and Paste buttons.) These lists shows the meanings in the form Term -> Definition for the highlighted category, so that one can see several meanings for the same term. Moreover, when a pair Term -> Definition is selected, the corresponding Term -> Definition translation is automatically highlighted; there is a one-to-one mapping between meaning representation in all the languages. It should also be noted that meanings, which are language independent, are shown with the *best* representation we have in a given language, i.e., a pair Term -> Definition, since there are no other pair Term -> Definition2 with the same meaning (note that is the same term in both pairs.)

Synonyms and Related Terms Management Area This area has four lists for the synonyms, and related terms in both languages which correspond to the highlighted meaning in the Meaning Management area.

Database Control Area This area contains a button which is used to obtain a report about consistency of the database. Consistency detection reports about lack of textual definitions for terms, and other inconsistencies (circular references) and omissions (lack of related terms via relationships See and SynSet). This is quite important when authoring dictionaries, since a dictionary cannot be consistently built at each step, but it is constructively built from terms to relationships between terms (polysemy, synonymy.)

5. Conclusions

Continuing with the refinement of our development methodology of information systems for lexical databases, we have followed an elaborated and well sound design method. The design is based on the ontological semantics approach, and we have signaled the advantages of this approach in face of the non-ontological one. The design has been tested and used to complete the development of certain information systems to build and consult monolingual, bilingual and multilingual dictionaries.

Of course, the advantages of applying software engineering principles and methods to information systems for lexical databases are evident. Moreover, by using the resulting tools, the LDB authoring is a friendly simple task, and the inserted information has to accomplish certain constraints (consistency, non recurrence, ...) controlled by the system, helping the authoring process (avoiding violation of hard constraints and reporting the violation of soft constraints). Besides, the integration of diverse LDB built with these tools is assured by the migration tools developed for this purpose. In addition, the resulting dictionaries are friendly usable and supply very useful semantic information to the reader.

References

(ABB 02) Atkins S., Bel N., Bertagna F., Bouillon P., et al (2002) “From Resources to Applications. Designing TheMultilingual ISLE Lexical Entry”. In Proceedings of LREC 2002, Las Palmas, Canary Islands, Spain.

(EWN) http://www.uva.nl/EuroWordNet.html

(FAR 04) D. Farwell “Intermediate Representation”. Seventh Interlingua Workshop AMTA’04: Determining Interlingua Utility for Machine translation. Washington, DC, October, 2004.

(LEN 90) D.B. Lenat, and R.V. Guha, “Building Large Knowledge-Based Systems”, Reading, Massachussets, Addison-Wesley, 1990.

(LREC 02) Workshop on “International Standards of Terminology and Language Resources Management”, Las Palmas de Gran Canaria, June, 2002.

(MAH 95) K. Mahesh, and S. Nirenburg, “A situated ontology for practical NLP”. IJCAI’95. Montreal, August 19-21.

(MAN 03) M. Mangeot-Lerebours, G. Sérasset, M. Lafourcade. “Construction collaborative d’une base lexicale multilingue. Le projet Papillon”. TAL, Vol. 44 – 2. 2003

(MCG 00) Deborah L. McGuinness. "Conceptual Modeling for Distributed Ontology Environments", Proc. of The 8th Int. Conf. on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000), Darmstadt, Germany, August 14-18, 2000.

(MIK) MikroKosmos, http://crl.nmsu.edu/Research/Projects/mikro/index.html

(MIL 95) G. Miller, "WordNet: A Lexical Data Base for English", Communications of the ACM, Vol. 38, 11, 1995.

(MOR 02) A. Moreno, and C. Pérez, "Reusing the Mikrokosmos Ontology for Concept-based Multilingual Terminology Databases", Proc. of LREC, 2002.

(NIR 95) S. Nirenburg, V. Raskin, and B. Onyshkevich, "Apologiae Ontologiae", Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, Center for Computational Linguistics, Catholic University, Leuven, Belgium, pp. 106-114, 1995.

(NIR) S.Nirenburg and V.Raskin, "Ontological Semantics", In http://crl.nmsu.edu/Staff.pages/Technical/sergei/book.html

(ONT) http://www.ontology.org/main/papers/iccs-dlm.html

(POL 03) A. Polguère, "Etiquetage sémantique des lexies dans la base de données DiCo". TAL, Vol. 4 – 2. 2003.

(PRI 01) U. Priss, "Ontologies and Context". Midwest Artificial Intelligence And Cognitive Science Conference. Oxford, OH, USA, 2001

(RK 02) C. Raguenaud and J. Kennedy, "Multiple Overlapping Classifications: Issues and Solutions". 14th International Conference on Scientific and Statistical Database Management (SSDBM'02). Edingburgh, Scotland, 2002.

(SV 02) Sáenz, F. & Vaquero, A. "Towards a Development Methodology for managing Linguistic Knowledge Bases". Proceedings ES'2002. Springer-Verlag, 2002. pp 453 – 466.

(USC 96) M. Uschold and M. Gruninger, "Ontologies: principles, methods, and applications". Knowledge Engineering Review, Vol. 11, 2. 1996, pp 93-155.

(VIE 98) E. Viegas, "Multilingual Computational Semantic Lexicons in Action: The WYSINNWYG Approach to NLP". Int. Conference on Computational Linguistics, ACL. Montreal, 1998.

(WIL 90) Y.A. Wilks, D.C. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M.Slator, "Providing machine tractable dictionary tools". Machine Translation, 5, 1990, pp. 99-151.

(WIL 96) Y. Wilks, B. M. Slator, and L.M. Guthrie, "Electric words: Dictionaries, Computers and Meanings". MIT Press. Cambridge, 1996.

(ZOC 03) M. Zock and J. Carroll "Les dictionaires électroniques". TAL, Vol. 44, 2. 2003.

Dynamic Metalanguage Customisation with the Dictionary Application TshwaneLex

GILLES-MAURICE DE SCHRYVER

Department of African Languages and Cultures, Ghent University
Rozier 44, 9000 Gent, Belgium
gillesmaurice.deschryver@UGent.be
&
TshwaneDJe Human Language Technology
P.O. Box 299, Wapadrand 0050, Tshwane (Pretoria), South Africa
gillesmaurice.deschryver@tshwanedje.com

DAVID JOFFE

TshwaneDJe Human Language Technology
P.O. Box 299, Wapadrand 0050, Tshwane (Pretoria), South Africa
david.joffe@tshwanedje.com

In the present contribution, the point of departure is the dynamic metalanguage customisation that is realised in real time on the Web for reference works produced with the lexicography software TshwaneLex. This unique feature is absent from even the best electronic dictionaries currently on the market. It is shown that, to achieve this type of customisation, functionality beyond straightforward XML had to be implemented in TshwaneLex. This extra functionality has been made available to the dictionary compilers through a user-friendly editor dialog as part of the fully customisable and built-in DTD. Once set up, the language and format of all metalanguage can not only be easily changed at any point during compilation, but dictionaries can also be customised for particular target users or particular dictionaries (e.g. pocket versus unabridged editions) when outputting for print, while truly instantaneous tailoring is effectively made possible for electronic and online dictionaries.

The dictionary writing system TshwaneLex

TshwaneLex is professional off-the-shelf lexicography software written in the C++ programming language and built using wxWidgets (an Open Source application development library). The stand-alone version requires a PC with Microsoft Windows 98 / Me / 2000 / XP. For full Unicode support, Windows 2000 or XP is recommended. Storage space and memory requirements are dependent on the size of the dictionary project.

TshwaneLex is currently being used as the lexicographic backbone for several projects at Oxford University Press, Macmillan and Van Dale Lexicografie, among others. Several dozen (smaller) dictionary projects around the world and for a multitude of different languages also make use of the software.

Over the past few years, TshwaneLex has been covered in a number of publications. A general overview and a first elaboration on some computational aspects of TshwaneLex may be found in Joffe *et al.* (2003a), respectively Joffe *et al.* (2003b). Secondly, a lexicographic perspective and an in-depth study of a real-world online lexicographic application are offered in Joffe & De Schryver (2004), respectively De Schryver & Joffe (2004). Thirdly, the fully customisable and built-in DTD editor of TshwaneLex, as well as more advanced DTD aspects are reported on in Joffe & De Schryver (2005), respectively De Schryver & Joffe (2005). Readers are invited to consult those publications for background information, if they so wish, before proceeding with the current discussion.

Beyond straightforward XML in TshwaneLex & Problem statement

The advanced dictionary-compilation-specific functionality built into TshwaneLex, such as Linked View (whereby implicit links between the two sides of a bidirectional bilingual dictionary are automatically made visible for the lexicographer), Automatic Reversal (whereby single articles or even an entire (semi-)bilingual dictionary may be reversed by the software), or Cross-reference Tracking (whereby cross-reference integrity is ensured at all times by means of the automatic updating of target homonym and sense numbers whenever these change), strongly differentiate TshwaneLex from any ordinary generic XML editor.

Even so, the TshwaneLex DTD system is ‘modelled’ after the XML DTD system, meaning that most of the major components of XML DTDs such as elements, attributes, attribute types, child relations, etc. have been implemented. In some cases, however, ‘special extras’ beyond straightforward XML had to be put into place, precisely to add features that make TshwaneLex more powerful as a dictionary editing environment. The present contribution deals with one such extra, namely the need to be able to dynamically customise the metalanguage, and this throughout compilation, at output stage, as well as during electronic and online use.

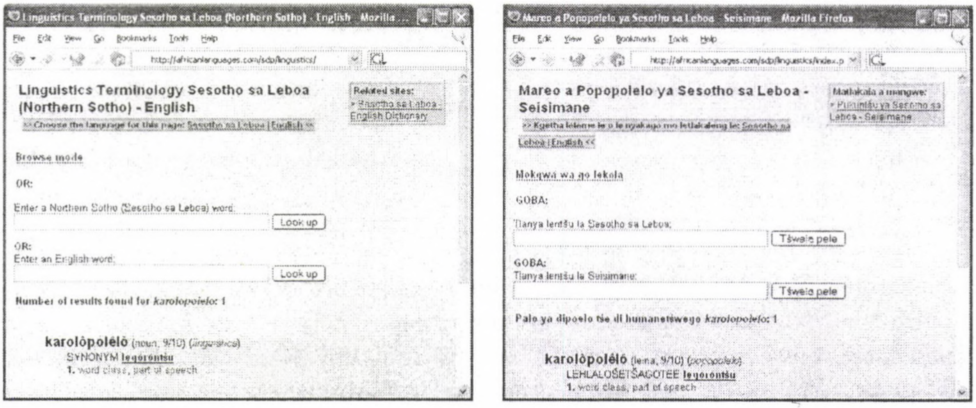
To begin with, and referring to the nature of the metalanguage in bilingual paper dictionaries, Honselaar states:

> Naturally, the meta-language is [in] the native language of the target group. So, in an English-Swedish dictionary for English speakers, comments will be in English. If a set of dictionaries X-Y and Y-X is meant for speakers of both X and Y, the meta-language may consist of words and abbreviations that are common to both languages. A neutral medium such as Latin may also be used. (Honselaar 2003: 324)

This is indeed how lexicographers *used* to go about it, and opting for one of the options is still the case when publishers intend to print only one set of dictionaries to cover all markets simultaneously. In an electronic environment, this need not be the case anymore, of course. In order to illustrate this, two screenshots are shown in Figure 1 reflecting typical instances of the customisation of the output-language in the *Linguistics Terminology Sesotho sa Leboa (Northern Sotho) – English* (Taljard & De Schryver 2003), an online dictionary produced with TshwaneLex.

When using the online dictionary with the interface in English, looking up a word like **karolopolelo** will – apart from the English translation equivalent 'word class, part of speech' – also return the POS tag (noun), label (linguistics) and the cross-reference marker text (SYNONYM) in English. For users who use the Sesotho sa Leboa interface, however, this same information will be *customised* for them, and POS tags, labels and cross-reference marker texts are all displayed in Sesotho sa Leboa (here respectively as **leina**, **popopolelo** and **LEHLALOŠETŠAGOTEE**).

According to De Schryver (2003: 12) the terminology list shown in Figure 1 contained a world's first for any Web dictionary as, at the time, no other Web


Linguistics Terminology Sesotho sa Leboa (Northern Sotho) - English - Mozilla ...
File Edit View Go Bookmarks Tools Help
http://africanlanguages.com/sdp/linguistics/
Look up
Look up
Mareo a Popopolelo ya Sesotho sa Leboa - Seisimane - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://africanlanguages.com/sdp/linguistics/index.p
Tšwela pele
Tšwela pele

Linguistics Terminology Sesotho sa Leboa (Northern Sotho) - English

>> Choose the language for this page: Sesotho sa Leboa | English <<

Related sites:
› Sesotho sa Leboa - English Dictionary

Browse mode

OR:

Enter a Northern Sotho (Sesotho sa Leboa) word:

OR:
Enter an English word:

Number of results found for *karolopolelo*: 1

karolôpolêlô (noun, 9/10) (*linguistics*)
SYNONYM **legorontšu**
1. word class, part of speech

Mareo a Popopolelo ya Sesotho sa Leboa - Seisimane

>> Kgetha leleme le o le nyakago mo letlakaleng le: Sesotho sa Leboa | English <<

Matlakala a mangwe:
› Pukuntšu ya Sesotho sa Leboa - Seisimane

Mokgwa wa go lekola

GOBA:

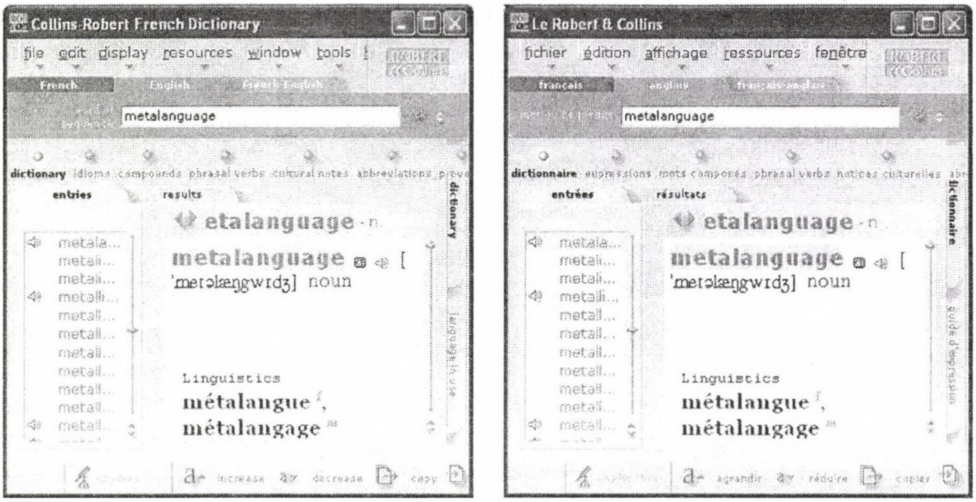
Tlanya lentšu la Sesotho sa Leboa:

GOBA:
Tlanya lentšu la Seisimane:

Palo ya dipoelo tše di humanetšwego *karolopolelo*: 1

karolòpolêlô (leina, 9/10) (*popopolelo*)
LEHLALOŠETŠAGOTEE **legorontšu**
1. word class, part of speech

Figure 1. Looking up in an online dictionary produced with TshwaneLex, with the interface in English (left) versus Sesotho sa Leboa (right).


Collins-Robert French Dictionary
file edit display resources window tools
French
English
French-English
dictionary idioms compounds phrasal verbs cultural notes abbreviations
entries
results
dictionary
language in use
increase
decrease
copy
Le Robert & Collins
fichier édition affichage ressources fenêtre
français
anglais
français-anglais
dictionnaire expressions mots composés phrasal verbs notices culturelles
entrées
résultats
dictionnaire
guide d'expression
agrandir
réduire
copier

metalanguage

etalanguage - n

metalanguage ['metəlæŋgwɪdʒ] noun

Linguistics
métalangue f,
métalangage m

Figure 2. Looking up in the English – French side of *Le Grand Robert & Collins Électronique* (2003), with the interface in English (left) versus French (right).

dictionary dynamically customised the output-language of POS tags, usage labels and cross-reference marker texts depending on the interface-language chosen. Unfortunately, up to this day, even the better commercial bilingual electronic dictionaries do indeed not achieve this, as is illustrated in Figure 2 for *Le Grand Robert & Collins Électronique* (2003), a bidirectional bilingual French – English / English – French dictionary.

Although the entire interface text is either presented in English or in French depending on the option the user chose, the metalanguage itself is *not* customised. In the example from the screenshots in Figure 2, where the word 'metalanguage' is being looked up, even when consulting the electronic dictionary in a French environment, the POS is still indicated in English as 'noun', and the label is still indicated in English as 'Linguistics', instead of '**nom**' and '**Linguistique**' respectively.

Also note that, while the small 'f' and 'm' in superscript (following the translation equivalents) might stand for both 'feminine noun / **nom féminen**' and 'masculine noun / **nom masculin**' respectively, these abbreviations clearly have not been conceptualised as being part of the metalanguage. If one clicks on 'f' one obtains a new window with two options: 'F - nm' and 'F - abr'. The first leads to "**F, f** … nom masculin …", the second to "**F** … abréviation … franc … Fahrenheit … frère". The options for 'm' lead to: (1) the letter M, m, (2) me / m', (3) the abbreviation for metre, (4) m' (cross-referred to (2)), and (5) the abbreviation for mister. In other words, if one does not already know what these abbreviations stand for, one receives *no* guidance at all, with the first option for 'f' as 'nom masculin' definitely confusing. One is thus forced to go into the help files attached to this electronic dictionary, to the sub-section 'Symbols and abbreviations / Symboles et abréviations'. This links in with Atkins' statement:

> All metalanguage should be in the user's mother tongue (L1). This will obviously involve reduplication of effort at the compiling stage, but in an online dictionary should not result in redundant information at the point of use. (Atkins 1996: 525)

While it is true that presenting just one language to the user should not result in redundant information being offered, it is *not* true that preparing this kind of information involves a reduplication of effort. By and large, the metalanguage of a dictionary is predictable, with POS tags, labels, cross-reference marker texts, and the like, all belonging to closed sets. In a truly modern dictionary compilation program all these metalanguage elements should therefore be selectable from lists (and should thus never be typed in when compiling articles). If one now designs the software in such a way that each of those lists can have as many (customisable) 'linked variant / alternative lists' as one wants, then the metalanguage of an entire dictionary can be swapped from one language to another, or from a long form to an abbreviated form, etc., with just one instruction. This is precisely how TshwaneLex was designed.

On attribute lists in TshwaneLex

XML DTDs are too limited when it comes to the handling of 'closed lists' for practical lexicographic use. It was strongly felt that these were required in TshwaneLex, however, as should be clear from the problem statement above.

In TshwaneLex, lists are stored in a single, central place at the beginning of the XML file. Example [1] shows a section of the PyaSsaL file in this regard, PyaSsaL (Mojela *et al.* 2004) being the in-progress and PanSALB-sponsored monolingual dictionary for Sesotho sa Leboa compiled with TshwaneLex at one of South Africa's eleven National Lexicography Units:

```
[1]  <dtdlist id="2" name="Part of speech">
       <dtdlistitem id="8" name="noun"/>
       <dtdlistitem id="9" name="pl noun"/>
       <dtdlistitem id="10" name="verb"/>
       <dtdlistitem id="11" name="adjective"/>
     ...
       <labelset name="Sesotho sa Leboa">
         <label listitemid="8" name="leina ka botee"/>
         <label listitemid="9" name="leina ka bontši"/>
         <label listitemid="10" name="lediri"/>
```

```
    <label listitemid="11" name="lehlaodi"/>
...
  </labelset>
  <labelset name="Sesotho sa Leboa (abbreviated)">
    <label listitemid="8" name="l.bot."/>
    <label listitemid="9" name="l.bon."/>
    <label listitemid="10" name="ldr."/>
    <label listitemid="11" name="lhl."/>
...
  </labelset>
...
</dtdlist>
<dtdlist id="3" name="Noun class">
  <dtdlistitem id="18" name="0   0/6"/>
  <dtdlistitem id="19" name="1   1/-"/>
  <dtdlistitem id="20" name="1   1/2"/>
  <dtdlistitem id="21" name="1a  1a/2a"/>
  <dtdlistitem id="22" name="2   1/2 p"/>
  <dtdlistitem id="23" name="2a  1a/2a p"/>
...
  <labelset name="Sesotho sa Leboa">
    <label listitemid="18" name="%b0%b/6"/>
    <label listitemid="19" name="%b1%b/-"/>
    <label listitemid="20" name="%b1%b/2"/>
    <label listitemid="21" name="%b1a%b/2a"/>
    <label listitemid="22" name="1/%b2%b"/>
    <label listitemid="23" name="1a/%b2a%b"/>
...
  </labelset>
...
</dtdlist>
```

Each item in the list is given a unique ID. Internally, when one selects a list item on an attribute, it stores, for that attribute in the document, a list of the list item IDs that are selected, rather than the text of the selected items. So for the article **lengwalo**[1] in PyaSsaL, one might have:

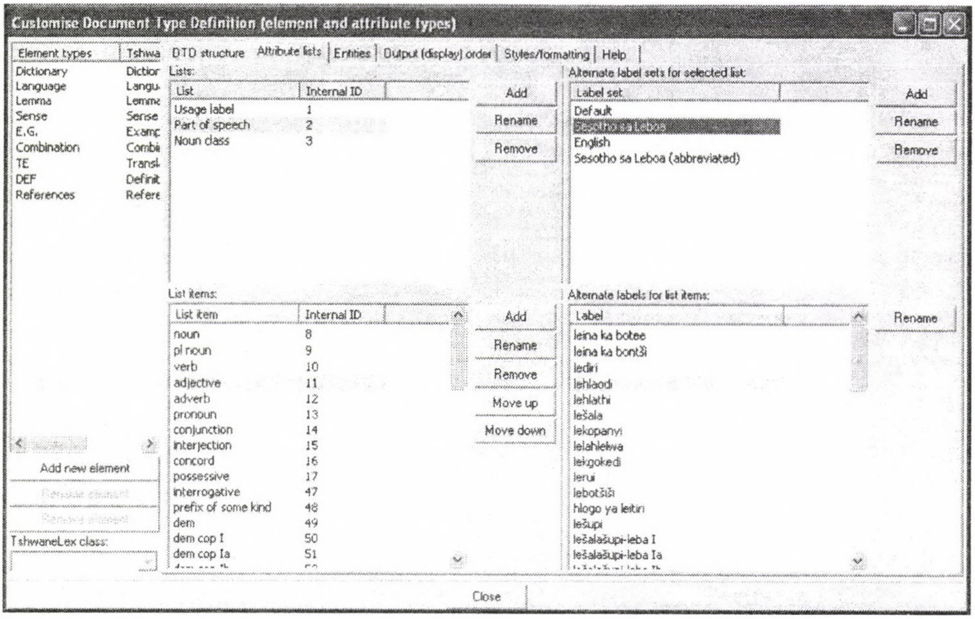
[2]

```
<Lemma id="3028" LemmaSign="lengwalo"
HomonymNumber="1" Pronunciation="lengwalô"
PartOfSpeech="8" NounClass="28">
  <Sense id="3029">
    <References id="23888"/>
    <DEF id="3030" Definition="pampiri yeo go
    ngwadilwego atrese le ditaba goba melaetša go
    yona; gantši e phuthelwa ka gare ga omfolopo ya
    romelwa, gantši ka poso"/>
    <E.G. id="3031" Example="Maphutha o amogela ~
    la go tšwa go kgoro ya thuto"/>
  </Sense>
  <Sense id="3032">
    <DEF id="3033" Definition="pampiri ya bohlatse
    yeo e bontšhago gore motho o na le ditshwanelo
    goba dithuto tše di itšego"/>
    <E.G. id="3034" Example="Ke tla hwetša khuetšo
    le nna ka hwetša ~ la matriki"/>
  </Sense>
</Lemma>
```

As may be seen from example [2], for 'PartOfSpeech' TshwaneLex stores, internally, '8' rather than '**leina ka botee**' 'singular noun'. This effectively makes it possible to change the text corresponding to '8' in just one place, or to define a substitute text label such as an abbreviated form '**l.bot.**', or even to create a translated version in another language. Likewise, in example [2] the 'NounClass' is stored as '28' rather than '**5**/6', which again means that such a notation may be changed throughout the entire dictionary to, say, '**le**-/ma-' in one go. Clearly, it is thanks to features such as these (singular and plural cross-reference types are handled in a similar way) that the entire metalanguage may easily be customised in dictionaries compiled with TshwaneLex. This further also allows labels to be Unicode text and to consist of any character(s), unlike the 'enumerated list' XML attribute type.

Of course, to do this, and to provide a self-explanatory interface for this – see in this regard Figure 3, which shows one of the tabs of the DTD editor dialog – TshwaneLex is taking care of a number of aspects 'behind the scenes' that an ordinary generic XML editor does not do. Further note that there are two different list types. In the first the lexicographer can only select one item from the list ('one of'), in the second zero or more items may be selected (the latter, again not possible with the XML DTD 'enumerated list' type). For the second type, a difference is also made between 'sorted' and 'unsorted'. For the sorted type, the order of the output of selected list items will always be the same as the order of the items defined centrally for the list. For the unsorted type, the order of the output of selected list items will be the same as the order in which they are selected by the lexicographer. Lastly, also note that any field can be converted from a free text field to a closed list (and vice versa) at any time in TshwaneLex.

Once an attribute list and its alternates have been set up in the DTD (cf. Figure 3), lexicographers may immediately use those to compile their articles, as may be seen from the F2 sub-window in Figure 4. Swapping to an alternate label set for attribute lists (or cross-reference texts) may easily be done under the F4 sub-window, as shown in Figure 5. Changes take immediate effect throughout the entire dictionary database, as seen in the preview, as well as when exporting data.



Customise Document Type Definition (element and attribute types)

Element types	Tshwa
Dictionary	Dictior
Language	Langu
Lemma	Lemma
Sense	Sense
E.G.	Examp
Combination	Combi
TE	Transl
DEF	Definit
References	Refere

Lists:

List	Internal ID
Usage label	1
Part of speech	2
Noun class	3

Alternate label sets for selected list:

Label set
Default
Sesotho sa Leboa
English
Sesotho sa Leboa (abbreviated)

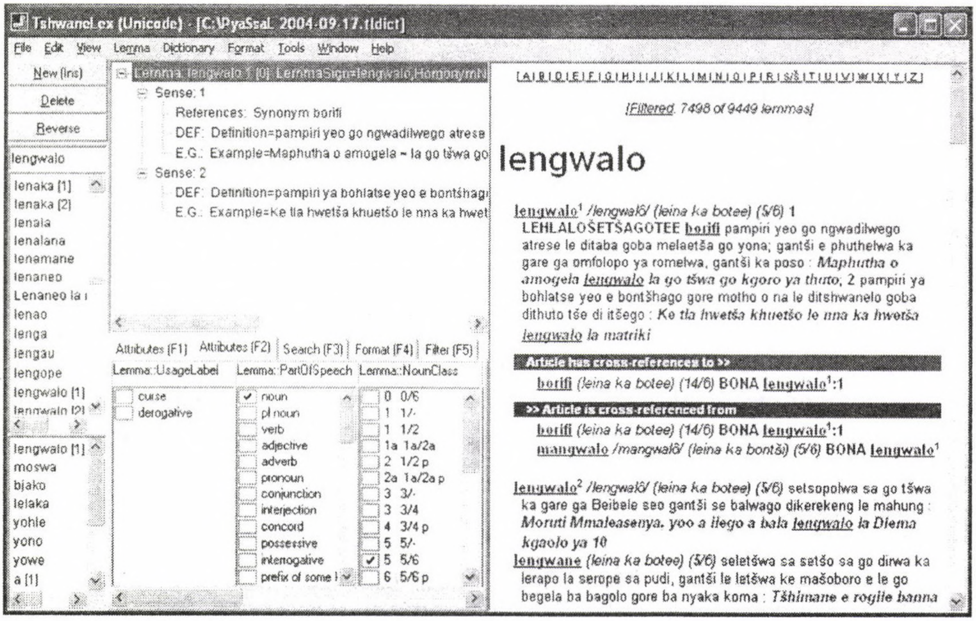
List items:

List item	Internal ID
noun	8
pl noun	9
verb	10
adjective	11
adverb	12
pronoun	13
conjunction	14
interjection	15
concord	16
possessive	17
interrogative	47
prefix of some kind	48
dem	49
dem cop I	50
dem cop Ia	51

Alternate labels for list items:

Label
leina ka botee
leina ka bontši
lediri
lehlaodi
lehlathi
lešala
lekopanyi
lelahlelwa
lekgokedi
lerui
lebotšiši
hlogo ya leitiri
lešupi
lešalašupi-leba I
lešalašupi-leba Ia

Figure 3. TshwaneLex ‘attribute lists’ editor dialog.



TshwaneLex (Unicode) - [C:\PyaSsaL 2004-09-17.tldict]

- Lemma: lengwalo 1 [0]; LemmaSign=lengwalo,Homonym...
  - Sense: 1
    - References: Synonym borifi
    - DEF: Definition=pampiri yeo go ngwadilwego atrese
    - E.G.: Example=Maphutha o amogela ~ la go tšwa go
  - Sense: 2
    - DEF: Definition=pampiri ya bohlatse yeo e bontšhag
    - E.G.: Example=Ke tla hwetša khuetšo le nna ka hwet

Lemma::UsageLabel	Lemma::PartOfSpeech	Lemma::NounClass
□ curse	☑ noun	□ 0 0/6
□ derogative	□ pl noun	□ 1 1/-
	□ verb	□ 1 1/2
	□ adjective	□ 1a 1a/2a
	□ adverb	□ 2 1/2 p
	□ pronoun	□ 2a 1a/2a p
	□ conjunction	□ 3 3/-
	□ interjection	□ 3 3/4
	□ concord	□ 4 3/4 p
	□ possessive	□ 5 5/-
	□ interrogative	☑ 5 5/6
	□ prefix of some k	□ 6 5/6 p

|A|B|D|E|F|G|H|I|J|K|L|M|N|O|P|R|S/Š|T|U|V|W|X|Y|Z|

[Filtered. 7498 of 9449 lemmas]

lengwalo

lengwalo[1] /lengwalô/ (leina ka botee) (5/6) 1 LEHLALOŠETŠAGOTEE borifi pampiri yeo go ngwadilwego atrese le ditaba goba melaetša go yona; gantši e phuthelwa ka gare ga omfolopo ya romelwa, gantši ka poso : *Maphutha o amogela lengwalo la go tšwa go kgoro ya thuto*; 2 pampiri ya bohlatse yeo e bontšhago gore motho o na le ditshwanelo goba dithuto tše di itšego : *Ke tla hwetša khuetšo le nna ka hwetša lengwalo la matriki*

Article has cross-references to >>

borifi *(leina ka botee) (14/6)* BONA lengwalo[1]:1

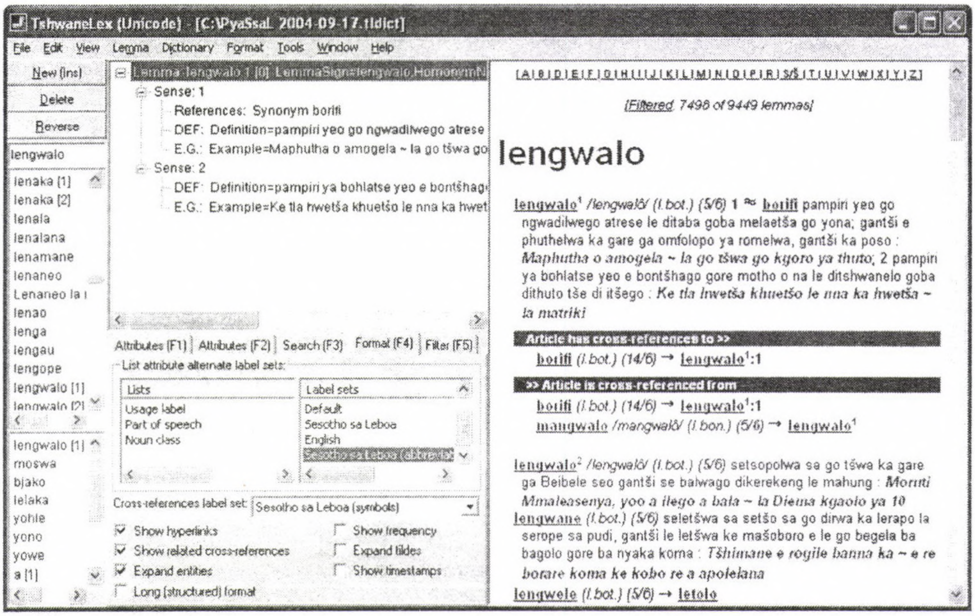
>> Article is cross-referenced from

borifi *(leina ka botee) (14/6)* BONA lengwalo[1]:1
mangwalo */mangwalô/ (leina ka bontši) (5/6)* BONA lengwalo[1]

lengwalo[2] */lengwalô/ (leina ka botee) (5/6)* setsopolwa sa go tšwa ka gare ga Beibele seo gantši se balwago dikerekeng le mahung : *Moruti Mmaleasenya. yoo a ilego a bala lengwalo la Diema kgaolo ya 10*

lengwane *(leina ka botee) (5/6)* seletšwa sa setšo sa go dirwa ka lerapo la serope sa pudi, gantši le letšwa ke mašoboro e le go begela ba bagolo gore ba nyaka koma : *Tšhimane e rogile banna*

Figure 4. Selecting list items (under F2) while compiling in TshwaneLex.



Lemma: lengwalo 1 [0] LemmaSign=lengwalo,HomonymN

Sense: 1
- References: Synonym borifi
- DEF: Definition=pampiri yeo go ngwadilwego atrese
- E.G.: Example=Maphutha o amogela ~ la go tšwa go

Sense: 2
- DEF: Definition=pampiri ya bohlatse yeo e bontšhag
- E.G.: Example=Ke tla hwetša khuetšo le nna ka hwet

List attribute alternate label sets:

Lists	Label sets
Usage label	Default
Part of speech	Sesotho sa Leboa
Noun class	English
	Sesotho sa Leboa (abbreviat

Cross-references label set: Sesotho sa Leboa (symbols)

☑ Show hyperlinks
☑ Show related cross-references
☑ Expand entities
☐ Long (structured) format
☐ Show frequency
☐ Expand tildes
☐ Show timestamps

(Filtered. 7498 of 9449 lemmas)

lengwalo

lengwalo^1 */lengwalô/ (l.bot.) (5/6)* **1** ≈ borifi pampiri yeo go ngwadilwego atrese le ditaba goba melaetša go yona; gantši e phuthelwa ka gare ga omfolopo ya romelwa, gantši ka poso : *Maphutha o amogela ~ la go tšwa go kgoro ya thuto*; **2** pampiri ya bohlatse yeo e bontšhago gore motho o na le ditshwanelo goba dithuto tše di itšego : *Ke tla hwetša khuetšo le nna ka hwetša ~ la matriki*

Article has cross-references to >>

borifi *(l.bot.) (14/6)* → lengwalo^1:1

>> Article is cross-referenced from

borifi *(l.bot.) (14/6)* → lengwalo^1:1
mangwalo */mangwalô/ (l.bon.) (5/6)* → lengwalo^1

lengwalo^2 */lengwalô/ (l.bot.) (5/6)* setsopolwa sa go tšwa ka gare ga Beibele seo gantši se balwago dikerekeng le mahung : *Moruti Mmaleasenya, yoo a ilego a bala ~ la Diema kgaolo ya 10*
lengwane *(l.bot.) (5/6)* seletšwa sa setšo sa go dirwa ka lerapo la serope sa pudi, gantši le letšwa ke mašoboro e le go begela ba bagolo gore ba nyaka koma : *Tšhimane e rogile banna ka ~ e re borare koma ke kobo re a apolelana*
lengwele *(l.bot.) (5/6)* → letolo

Figure 5. Varying the metalanguage (under F4) while compiling in TshwaneLex.

Dynamic metalanguage customisation with TshwaneLex

The attribute-list system described above provides a powerful yet still easy to use method for customising the metalanguage. This functionality is furthermore carried through to other areas in the system, such as cross-reference type labels, as can be seen in Figure 5. Additionally, still other mechanisms for customisation are available for situations where labels may be embedded within other fields, as for example the French 'f' and 'm' gender labels. These may be defined as so-called XML entities (e.g. '&f;' and '&m;') which are replaced in the output with labels configured in a single, central place. This not only allows electronic or online dictionary software to easily customise the language of these labels, but also allows the software to be *aware* that these labels are part of the metalanguage, and to thus provide a more meaningful response should the user click on them. Clearly, with TshwaneLex, a truly powerful set of tools to fully customise the language of the metalanguage is put into the hands of the lexicographer for the very first time.

References

Atkins, B.T. Sue. 1996. 'Bilingual Dictionaries: Past, Present and Future', in Martin Gellerstam *et al.* (eds.). 1996. *Euralex '96 Proceedings I-II, Papers* *submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden*: 515–546. Gothenburg: Department of Swedish, Göteborg University.

De Schryver, Gilles-Maurice. 2003. 'Online Dictionaries on the Internet: An Overview for the African Languages'. *Lexikos* 13: 1–20.

De Schryver, Gilles-Maurice and David Joffe. 2004. 'On How Electronic Dictionaries are Really Used', in Geoffrey Williams and Sandra Vessier (eds.). 2004: 187–196.

De Schryver, Gilles-Maurice and David Joffe. 2005. 'One Database, Many Dictionaries – Varying Co(n)text with the Dictionary Application TshwaneLex', in *ASIALEX 2005 Proceedings*.

Honselaar, Wim. 2003. 'Examples of design and production criteria for bilingual dictionaries', in Piet van Sterkenburg (ed.). 2003. *A Practical Guide to Lexicography*: 323–332. Amsterdam: John Benjamins Publishing Company.

Joffe, David and Gilles-Maurice de Schryver. 2004. 'TshwaneLex – A State-of-the-Art Dictionary Compilation Program', in Geoffrey Williams and Sandra Vessier (eds.). 2004: 99–104.

Joffe, David and Gilles-Maurice de Schryver. 2005. 'Representing and Describing Words Flexibly with the Dictionary Application TshwaneLex', in *ASIALEX 2005 Proceedings*.

Joffe, David, Gilles-Maurice de Schryver and D.J. Prinsloo. 2003a. 'Introducing TshwaneLex – A New Computer Program for the Compilation of Dictionaries', in Gilles-Maurice de Schryver (ed.). 2003. *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*: 97–104. Pretoria: (SF)^2 Press.

Joffe, David, Gilles-Maurice de Schryver and D.J. Prinsloo. 2003b. 'Computational features of the dictionary application "TshwaneLex" '. *Southern African Linguistics and Applied Language Studies* 21/4: 239–250.

Le Grand Robert & Collins Électronique. 2003. Dictionnaires Le Robert / VUEF.

Mojela, M.V. (Editor-in-Chief), **M.P. Mogodi, M.C. Mphahlele and M.R. Selokela** (Compilers). 2004. *Pukuntšutlhaloši ya Sesotho sa Leboa ka Inthanete* [Explanatory Sesotho sa Leboa Dictionary on the Internet]. Available from: http://africanlanguages.com/psl/

Taljard, Elsabé and Gilles-Maurice de Schryver. 2003. *Online Linguistics Terminology Sesotho sa Leboa (Northern Sotho) – English*. Available from: http://africanlanguages.com/sdp/linguistics/

TshwaneLex. 2002-2005. Available from: http://tshwanedje.com/tshwanelex/

Williams, Geoffrey and Sandra Vessier (eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Phonolapsological equivalence and similarity in the English lexicon – automatic derivation of a Phonetic Difficulty Index (PDI) from a lexical database

WŁODZIMIERZ SOBKOWIAK

School of English, Adam Mickiewicz University
al.Niepodleglosci 4, 61-874 Poznań, Poland
sobkow@amu.edu.pl

The rationale, functionality and current structural implementation are described of an algorithm assigning phonetic difficulty tags to English words in a lexical database. The difficulty is defined with respect to commonly observed pronunciation problems of Polish learners of English as a foreign language (EFL). The resulting PDI index contains, for each wordform, both quantitative difficulty level information, with the range of 0-10, and qualitative difficulty tags in the form of strings of 57 PDI codes, each for one specific Polglish pronunciation problem. From these indices, phonolapsological equivalence and similarity classes can be derived, i.e. sets of words of identical/similar PDI level and/or code. These constructs are argued to have a number of potential applications in phonolexicography, EFL teaching and beyond.

Keywords: phonetic difficulty, phonolapsology, phonolexicography, Polglish, pronunciation

1. Background

In a series of contributions I have developed a computer-implemented algorithm generating the EFL Phonetic Difficulty Index (PDI) for English words as acquired by Polish learners. Briefly, "the idea of the index is that it is a global numerical measure of the phonetic difficulty of the given English lexical item for Polish learners. The measure combines (a) the most salient grapho-phonemic difficulties such learners are known to have reading English, i.e. mostly spelling pronunciation, (b) some commonest phonemic L1-interference problems known from the literature and my own teaching experience, finally (c) some of the notorious developmental L2-interference pronunciation errors observed in all learners of English regardless of their L1 background" (Sobkowiak 1999:214; see Appendix 1 below). The PDI can be used as a tool in a wide variety of pronunciation-related research projects and practical applications.

In my 1998 paper, in which I speculated on whether and how "EFL MRDs can teach pronunciation", I envisaged the role of the PDI as follows:

> "A phonetic difficulty rating tagged on each MRD entry would allow the exercise module algorithms to weigh the entries so that the more difficult are used more frequently, for example. Such rating can be produced semi-automatically, given the known rule-governed phonetic problems of Polish learners of English" (Sobkowiak 1998:274).

In my 1999 book, I devoted an 8-page-long section to the PDI. In it I presented the rationale and design of the index in detail, emphasizing: (a) the need for a phonetically graded EFL lexicographic resource, (b) its pedagogical (mainly lexicographic) applications, (c) the desirability of phonetic L1-sensitivity, and (d) the role of the phonetic evaluation of defining vocabularies.

This last thread was continued in my Euralex 2002 contribution (Sobkowiak & Kuczyński 2002), where I investigated the phonetic difficulty of defining vocabularies in two EFL dictionaries: LDOCE and CIDE, concluding that they "are, after all, significantly phonetically easier than the frequency-matched portions of the reference lexicon, here treated as chance level" (p. 498).

In 2000 I used the PDI to evaluate phonetic keywords in five EFL dictionaries, to find that they vary in PDI between 1.05 (Oxford; easiest) and 1.27 (Longman; hardest), with 1.10 being the mean PDI for English monosyllables at large. As it turned out in my 2002 study (Sobkowiak 2002), also some phonetic keyword wall charts contain words with rather high PDIs, such as 1.68.

In an empirical study of 208 English philology students, which I conducted in February 2000 (Sobkowiak, unpublished) I 'field-tested' my PDI, collecting their subjective phonetic difficulty ratings for 20 frequency- and PDI-stratified words. These ratings turned out to correlate very highly with the PDI scores (r=0.684), thus lending empirical credibility to the otherwise intuitively conceived metric.

At the 4th national conference on teaching foreign pronunciation in Poland (Sobkowiak 2004a) I introduced: (a) the genesis, rationale and bibliography of PDI, (b) its current structural implementation, (c) its existing and potential functional applications, mostly phonolexicographic, pedagogical, psycholinguistic and those related to NLP.

At the 35th Poznań Linguistic Meeting, in May 2004, I described the use of PDI in the ongoing Colorado Literacy Tutor joint project, which IFA UAM has undertaken together with the Colorado Center for Spoken Language Research. PDI was being used there to phonetically evaluate and grade TIMIT sentences submitted to Polish learners of English for reading aloud in the process of training an EFL speech recognizer (Sobkowiak, in press).

Continuing my Euralex thread of phonolexicographic papers, I presented one on phonetically controlled dictionary definitions to the 11th Euralex congress in Lorient, France, in July 2004 (Sobkowiak 2004b). In that contribution I used the PDI to investigate the phonetic difficulty of definitions (rather than only defining vocabulary) in *Macmillan English Dictionary for Advanced Learners* (MEDAL).

2. Phonolapsological equivalence and similarity in the English lexicon

Finally, in my paper prepared for the 5th national conference on teaching foreign pronunciation in Poland (Soczewka, May 2005; Sobkowiak, forthcoming) I looked at the patterns of lexical co-occurrence of qualitative PDI codes denoting specific grapho-phonetic difficulties of Polglish learners. Concluding that paper, I speculated:

> Consider the following perspective: lexical items in my database differ in terms of their PDI in two respects: quantitative and qualitative. First, some words are PDI-harder from others; second, most words, possibly with the same numerical difficulty index, display difficulty code subsets different from all other words. It would be interesting to see how words are located on the PDI-code similarity/equivalence scale. In particular: is it possible to have pairs of phonemically different words which would be identical in terms of their PDI codes? Such words would not only exhibit the same phonetic difficulty level, but they would also be qualitatively phonolapsologically identical. Would such pairs be interesting/useful in any phonetic sense? Some preliminary answers to such questions appear to be affirmative. Consider, for example, the following two words: *lightning-conductors* and

> *scandalmongers*. Despite their obvious phonemic non-equivalence, they are identical in terms of PDI, both quantitatively (PDI=7) and qualitatively: <aEHJN13>. One can thus expect identical pronouncing problems to arise for the learner in processing these words (both for human-producer and machine-recognizer, by the way). This kind of strong qualitative phonolapsological equivalence holding for word-pairs and subsets in the English lexicon may also have other, equally interesting phonetic ramifications.

In what follows I will develop this last theme. First, I will briefly present my lexical database (section 2.1.), then I will show and discuss examples of phonolapsological lexical identity (section 2.2.) and – very briefly, for lack of space – similarity (section 2.3.), finally some ideas and speculations on the possible applications of these measures and indexes are offered (section 2.4.)

2.1. The PDI lexical database

The algorithm assigning PDI numerical difficulty tags and qualitative difficulty codes was run over the machine-readable *Oxford Advanced Learner's Dictionary of Current English* (OALDCE) word-list (see Mitton 1986 and 1992). It generated the PDI range between 0 and 10, with a mean of 2.45, and standard deviation 1.5. The list currently counts 85430 (unlemmatised) records and 25264 lemmas. On top of the global numerical rating of phonetic difficulty, the PDI algorithm assigns 57 qualitative difficulty codes taken from the list reproduced in Appendix 1. The following, for example, are all the four lemmas with PDI=9 in the database, with codes on the right. Notice that, while the PDI value stands at 9 for these words, their particular phonolapsological codes differ, i.e. while they are predicted to be more or less equally hard phonetically (to Polish learners), the sources of these difficulties vary. In other words: the two words are quantitatively, but not qualitatively, phonolapsologically equivalent. These are of course empirical, testable claims.

Table 1. Examples of words with PDI=9, with their difficulty codes

word	phonetic transcription	PDI	PDI code
entourage	,0ntU'rAZ	9	bgsGNQT13
misbehaviour	,mIsbI'h4v6R	9	bgACORV13
undervaluation	,Vnd@,v&ljU'4Sn	9	EJQSTX123
undistinguishable	,IndI'stINgwIS@bl	9	vEHJQTX23

By looking up the phonetic difficulty list in Appendix 1 the reader will be able to ascertain exactly which potential Polglish pronunciation problems have been recognized for each of the four words. The list has no pretense to being definitive, of course; it is simply an interim, doubtless personally biased, codification of general professional wisdom and experience among EFL teachers in Poland (but with some empirical backing, as mentioned above; see Sobkowiak unpublished).

2.2. Phonolapsological identity

In Sobkowiak (forthcoming; see quote above) the *lightning-conductors / scandalmongers* pair illustrated the case of complete phonolapsological equivalence, i.e. both quantitative and qualitative identity. Two or more phonemically nonequivalent (heterophonic) words in my lexical database can exhibit the same PDI code subset, thus constituting what I choose to call a (qualitative) PhonoLapsological Identity Class (PLIC). There are exactly 3484 such PLICs in the database, as seen in Table 2.

Table 2. Some PDI and PhonoLapsological Identity Class (PLIC) statistics

	words	#PLICs	#words in PLICs (% all words at this PDI level)	largest PLIC		examples *
PDI				**PDI code**	**#words**	
10	4					
9	24					
8	165	18	39 (24%)	bdJKNQU1	4	*surgeons*
7	684	99	277 (41%)	JNQTY13	10	*confirmations, concentrations, conversations*
6	2334	377	1336 (57%)	JNTY13	65	*operations, internationals, competitions*
5	5630	785	4044 (72%)	JTX13	100	*operation, information, international*
4	11357	1025	8958 (79%)	aJN1	302	*taxpayers, afternoons, straightforward*
3	18114	833	14611 (81%)	JN1	1613	*overs, members, problems*
2	20806	306	16430 (79%)	JN	1324	*families, systems, provide*

1	19047	41	14276 (75%)	N	3549	*and, is, did*
0	7265					
sum	85430	3484**	59971 (70%)		6967	

* 3 words belonging to top-frequency lemmas from Kilgarriff's BNC word-list (Kilgarriff 1997 and http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bnc-readme.html) were taken in each case, except PDI=8 (*surgeons* is the only word at all attested in the list).

** The number of PDI-code types, including unique ones, but excluding PDI=0, is 7104.

Analyzing the table, notice first that there are (accidentally) no PLICs at PDI level 9 and 10, and (systematically) at PDI level 0. The word PDI frequency distribution is normal, slightly positively skewed, with the mode at PDI=2. The PLIC PDI distribution is also normal, with mode at PDI=4. On average, 70% words in the database belong to some PLIC, while 30% are PDI-code-wise unique, but the actual proportion varies with the PDI level: roughly, the lower the level of difficulty, the larger the PLICs (as would be expected a priori). This is also true of the size-wise top PLICs at each PDI level.

The four PDI=10 words are: *agent provocateur*, *agents provocateurs* (abAJKQT123), *authoritarians* (fBCJMNQ123), *undervaluations* (EJNQSTY123). As can be seen, the first two words exhibit the same PDI code, but being homophones they do not qualify as forming a PLIC, hence there are no PLICs at this PDI level. All heterographic homophones were treated likewise, e.g. *baloney* and *boloney*, i.e. only heterophones were allowed into PLICs. There are no homophones among the 24 PDI=9 words, but all of them have unique PDI codes, so again there are no PLICs at this level.

The situation at PDI level 8 is different. All 18 PLICs with their 39 words are listed in Appendix 2. It will be seen that many PLICs are phonolapsologically not very interesting in that they are composed of phonemically closely similar, morphologically related, words, e.g. PLICs 10-18. In such cases there is no advantage in explicitly capturing lapsological equivalence via PLIC membership, compared to simply observing that phonomorphologically similar words tend to be also phonolapsologically equivalent. However, there are also more interesting PLICs, such as 1, 3 or 4 where phonomorphological similarity is more tenuous. Finally, such PLICs as 9 and 12 are most interesting of all in that, because their members are phonomorphologically even less similar, they would not normally be treated together in phonolapsologically relevant contexts such as EFL phonetic instruction, phoniatrics or speech processing, for example. It is in such cases that PDI-treated lexical database can genuinely help. More examples of PLICs containing highly heterophonic words at PDI level 7 are provided in Appendix 3. The level of phonomorphological similarity is on average even lower here than in the case of PLICs with PDI=8.

2.3. Phonolapsological similarity

For lack of space, only a short mention can be made here of classes of words in the lexicon which exhibit less than complete qualitative (heterophonic) phonolapsological equivalence (identity), i.e. those where the PDI codes do not ideally match. PhonoLapsological Similarity Quotient (PLSQ) is defined as the logical product (AND) divided by the logical sum (OR) of two PDI codes. Its range is 0-1 inclusive, and its values are discrete because the PDI values themselves are so, too. Phonolapsological identity discussed in section 2.2. is of course a boundary case with PLSQ = 1. The case of PLSQ=0 is not interesting in this context, as are, indeed, all low PLSQs (the cut-off point remains arbitrary, of course).

Words of the same PLSQ (<1, but >>0) form a PhonoLapsological Affinity Class (PLAC). Notice that a PLAC, unlike a PLIC, is composed of a number of distinct pairs of words (or PLICs) with the same PLSQ, but different PDI codes. Consider some PLAC examples below with high PLSQ. The words are listed in Table 3 with their PDI codes, full stop denoting mismatch.

Table 3. Examples of some PhonoLapsological Affinity Classes with top PLSQ

PLSQ	PDI code product/sum	PLAC word pair	PDI codes
0.889	8/9	undervaluation	EJQSTX123
		unconstitutional	EJQ.TX123
0.875	7/8	unperturbed	bEJKNQ13
		overburdened	b.JKNQ13
0.857	6/7	genre-paintings	aHJNQT1
		washing-machines	aHJN.T1
0.833	5/6	wonder	.AEJQ1
		tongue-twister	aAEJQ1
0.800	4/5	joi de vivre	a.JT3
		multinational	aEJT3
0.778	7/9	uncoloured	.bgEJNQ13
		sun-parlours	abgEJNQ1.
0.750	3/4	zoology	.JU1
		whichever	AJU1
	6/8	watercolour	abgAEJ.1
		troublemaker	ag.AEJY1

2.4. Possible applications of phonolapsological equivalence and similarity

A machine-readable English lexicon tagged with PDI codes and algorithmically processed for phonolapsological similarity can be an excellent resource for a number of applications. Electronic pedagogical dictionaries can fish for phonolapsological equivalences between words of a given phonetic difficulty for presentational purposes, as well as for interactive task construction. (Pronunciation) material developers can benefit from a PDI-treated lexicon as an important tool of the trade. Both teachers and learners might use it as an auxiliary phonolexical didactic resource. Psycholinguists engaged in empirical studies with some phonolexical involvement might find it useful for, e.g., stimulus selection (word, phrase or sentence, as the case may be). On a more theoretical level, dividing the (interlanguage) lexicon into phonolapsological equivalence classes makes strong empirical claims about the psychophonetic behaviour of words in a variety of processing tasks, not only in the area of phonetic lapsology. While some preliminary evidence exists for the empirical validity of the PDI measures in its current implementation, as mentioned above (Sobkowiak, unpublished), much remains to be done along these lines.

To take a more concrete example of how PDI can be used in practice: writing in the EFL lexicographic context, in my 1999 book I envisaged the following possible enhancement to electronic dictionaries:

> "For instance, if the difficulty index contains, on top of the global numerical measure, a code of the actual phonetic difficulty/ies present in the wordform [...], it will allow the user to investigate it directly through listing words with this same difficulty present, for example: 'If *radio*, which I am now having on screen is pronunciation-wise difficult for Poles because they tend to reduce the second-syllable vowel to a glide /j/ and the whole word to a bisyllable, give me more words with this phonetic problem in them'. An exemplary answer to such a query (listed in the order of frequency): *ratio, enthusiasm, polio, rodeo, appreciate, studio, embryo, associate, foliage* [...]. This same information can then be used in generating exercises of the type: 'Which of these words are like *radio* in that...' (multiple choice), or 'Which pronunciation of *radiate* is correct...?' (binary choice; with text-to-speech synthesised bisyllabic Polglish-like pronunciation)" (Sobkowiak 1999:254-5).

Should we now want to follow up on some of the 'radio-like' lexical yield, say *appreciate*, we will fish out four lemmas in the same PLIC, with the PDI code=

<JST>: *associate, officiate, propitiate, substantiate* and a number of words in some PLAC with *appreciate*, such as: *associated, propitiating, recreational, deviationist* (PLSQ=0.75), or *asphyxiation, negotiators* (PLSQ=0.6), or indeed – should we wish to go that low in phonolapsological similarity – *gigolos, womanish, Welshwoman, veracious, uppishness, volition, vacation* (PLSQ=0.5), etc. Another 'radio-like' word, *foliage*, in turn, belongs to a <sNU> PLIC containing 64 lemmas, frequency-topped by: *image, manage, village*. There is practically no limit to such serendipitous exploration.

Recently I sketched the applicational potential of the PDI metric as follows: "outside of the narrowly defined EFL arena, the potential of the PDI appears to be the greatest in (a) phonetics and phonology, (b) (meta)lexicography and lexicology, (c) lexical psycholinguistics, (d) contrastive and corpus linguistics, (e) natural language programming, especially automatic speech recognition (ASR), speech synthesis (TTS), and their applications in machine translation (MT), robotics, data mining, abstracting, and the like" (Sobkowiak 2004a). There is little I could add to this list here, except that a thorough exposition of how PDI, PLSQ, PLIC and PLAC can be used in these spheres would require much more space than I could afford in this contribution.

3. Conclusion

Google yields 95 hits to the query "phonetic difficulty" (save those coming from my own web page). The themes range from laboratory phonetics and theoretical phonology (not surprising), through stuttering research, markedness theory, denture fitting to golfing! There is practically nothing on the role of phonolexical difficulty metrics in EFL or FLT generally. A query of 'lapsology' (let alone 'phonolapsology') reaches only 9 pages, of which only the ad for (James 1998) is relevant in our context. Clearly, there is a need for a construct like the PDI, which is now seven years old, as well as for its derivatives: measures of cross-lexical phonolapsological similarity such as PLSQ and its derivatives in turn: PLIC and PLAC equivalence classes. A tool like this will equip both researchers and practitioners with new abilities in both applied linguistic theory and praxis.

Bibliography

James,C. 1998. *Errors in language learning and use*. London: Longman.

Kilgarriff,A. 1997. "Putting frequencies in the dictionary". *International Journal of Lexicography 10.2.*135-55.

Mitton,R. 1986. "A partial dictionary of English in computer usable form". *Literary and Linguistic Computing* 1.214-15.

Mitton,R. 1992. "A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English", bundled with the software.

Sobkowiak,W. 1998. "Can EFL MRDs teach pronunciation?". In T.Fontenelle et al. (eds). *Euralex'98 proceedings.* Liège: University of Liège, English and Dutch Departments. 271-77.

Sobkowiak,W. 1999. *Pronunciation in EFL machine-readable dictionaries.* Poznań: Motivex.

Sobkowiak,W. 2000. "Phonetic keywords in learner's dictionaries". In U.Heid et al. (eds). *Euralex'2000 proceedings.* Stuttgart: IMS. 237-46.

Sobkowiak,W. 2002. "Phonetic transcription wallcharts in EFL". In W.Sobkowiak and E.Waniek-Klimczak (eds). *Dydaktyka fonetyki języka obcego. Proceedings of the Soczewka Conference on teaching foreign pronunciation, 11-13.5.2001.* Płock: Wydawnictwo PWSZ. 161-172.

Sobkowiak,W. and M.Kuczyński. 2002. "Phonetics and ideology of defining vocabularies". In A.Braasch and C.Povlsen (eds). *Euralex'2002 proceedings.* Copenhagen: Center for Sprogteknologi. 495-502.

Sobkowiak,W. 2004a. "Phonetic Difficulty Index". In W.Sobkowiak & E.Waniek-Klimczak (eds). *Dydaktyka fonetyki języka obcego.* Zeszyt Naukowy Instytutu Neofilologii Państwowej Wyższej Szkoły Zawodowej w Koninie nr 3. Konin: Wydawnictwo PWSZ w Koninie. 102-107.

Sobkowiak,W. 2004b. "Phonetically controlled definitions?". In G.Williams & S.Vessier (eds). *Euralex'2004 proceedings.* Lorient, France, 6-10 July 2004. Lorient: Université de Bretagne Sud. 911-916.

Sobkowiak,W. (in press). "Automatic phonetic annotation of corpora for EFL purposes". Paper presented at the Workshop in assessing the potential of corpora at the 35th Poznań Linguistic Meeting, May 20, 2004.

Sobkowiak,W. (forthcoming). "PDI revisited: lexical cooccurrence of phonetic difficulty codes". Paper submitted to the 5th Phonetics in FLT Conference, Soczewka, 25-27 April 2005.

Sobkowiak,W. (unpublished). "Subjective phonetic difficulty of English words to Polish learners: does frequency matter?". Available at http://elex.amu.edu.pl/ ~sobkow/diffind2.doc.

Appendix 1: PDI codes used in the lexical database with their incidence (in brackets)

(a) mostly spelling and morphology

a -- compound	(11148)
b -- <ur> in word	(3145)
c -- <ei> in word	(623)
d -- <eo> in word	(427)
e -- <ow> in word	(1609)
f -- <au> in word	(991)
g -- <ou> in word	(3992)
h -- <aw> in word	(582)
i -- <lk_> in head	(132)
j -- <mb_> in head	(117)
k -- <mn_> in head	(36)
l -- <alm_> in head	(32)
m -- <gm_> in head	(9)
n -- <stle_> in head	(83)
o -- word=<mn>	(2)
p -- word=<ps>	(65)
q -- word=<al>C; C<>l	(129)
r -- <gh_> or <ght_> in head	(534)
s -- <age_> in head and not /eɪdʒ_/	(381)
t -- <ate_> in head and not /eɪt_/	(238)
u -- <ative_> in head and not /eɪtɪv_/	(163)
v -- <able_> in head and not /eɪbl_/	(510)
w -- <ey_> in head and not /eɪ_/	(278)

(b) mostly pronunciation

A -- linking /r/	(4787)
B -- /eə/	(1129)
C -- /ɪə/	(3337)
D -- /ʊə/	(851)
E -- /ʌ/	(8394)
F -- /tʃt_/ or /dʒd_/	(518)
G -- interconsonantal /ʊ/, but not <oo>	(1419)
H -- velar nasal	(10044)
I -- /ŋ/+V with no /g/	(141)
J -- short schwa	(32192)
K -- long schwa	(3639)
L -- voiced apico-dental	(724)
M -- voiceless apico-dental	(1803)
N -- final voiced obstruent	(31427)
O -- pre-voiced /dɪs/ or /mɪs/	(790)
P -- /əʊ/CCV	(784)
Q -- vowel nasalization	(7612)
R -- voiced obstruent + /s/ or /s/ + voiced obstruent	(594)
S -- /ueɪ/ or /ieɪ/	(496)
T -- post-alveolar fricatives	(7132)
U -- post-alveolar affricates	(7631)

V -- glottal fricative /h/	(4267)
W -- stop geminates	(125)
X – word-final syllabic sonorants	(3862)
Y – non-word-final syllabic sonorants	(2893)

(c) others

1 -- british<>american	(31710)
2 -- more than 5 syllables	(750)
3 -- secondary stress	(10351)
4 -- <ical_> in trisyllabic-plus adjectives -- stress	(141)
5 -- <ic_> in bisyllabic-plus adjectives -- stress	(477)
7 -- <ary_>/<ory_>/<ery_> in bisyllabic-plus heads	(717)
8 -- contraction of pronoun with verb, e.g. <you've>	(38)
9 -- proper noun	(2589)
0 -- abbreviation, incl. acronym	(363)

Appendix 2. Complete list of 18 PLICs with PDI=8 ordered alphabetically by PDI code

word	PDI code
equalitarians	BCJNQ123
parliamentarians	BCJNQ123
nonagenarians	BCJNU123
octogenarians	BCJNU123
undercharged	EFJNQU13
underprivileged	EFJNQU13
unconstitutional	EJQTX123
undenominational	EJQTX123
popularizations	GJNTY123
regularizations	GJNTY123
anthropological	JMUX1234
methodological	JMUX1234
ornithological	JMUX1234
differentiations	JNQSTY23
reconciliations	JNQSTY23
airing-cupboard	aBEHJN13
airing-cupboards	aBEHJN13
Solicitor-Generals	aJNQU123
quartermaster-generals	aJNQU123
court-martialed	abgNTY13
court-martials	abgNTY13
unburdened	bEKNQY13
unburdens	bEKNQY13
burgeoned	bdJKNQU1
burgeons	bdJKNQU1
sturgeons	bdJKNQU1
surgeons	bdJKNQU1
undernourished	bgEJQT13
undernourishment	bgEJQT13
discouraged	bgsEFNU1
encouraged	bgsEFNU1
harbourage	bgsJNUV1
harbourages	bgsJNUV1
authoritarian	fBCJM123
authoritarianism	fBCJM123
disadvantage	sJNOQU13
disadvantages	sJNOQU13
undergraduate	tDEJQU13

undergraduates	tDEJQU13

Appendix 3. Selected PLICs with highly heterophonic words with PDI=7

word	**PDI code**
Godmanchester	AJQU139
Chalfont St Peter	AJQU139
unappreciated	EJNST23
mispronunciations	EJNST23
controversial	JKQTX13
circumvention	JKQTX13
fortune-teller	aAJQU13
sergeant-major	aAJQU13
bug-hunters	aEJNQV1
dunderhead	aEJNQV1
forasmuch as	aEJNU13
Governor-Generals	aEJNU13
gentleman-at-arms	aJNQU13
cross-questioned	aJNQU13
major-generals	aJNQU13
sergeant-majors	aJNQU13
Serjeant-at-arms	aJNQU13
churchgoer	abAJKU1
curtain-lecture	abAJKU1
word-picture	abAJKU1
time-exposure	abAJT13
made-to-measure	abAJT13
cross-purposes	abJKN13
topsy-turvydoms	abJKN13
good-humoured	abgJNV1
half-hours	abgJNV1
double-dealers	agEJN13
counter-productive	agEJN13
entrepreneurs	bJKNQ13
overburdened	bJKNQ13

Boundedness and boundability: Estonian transitive change of state verbs in LFG

Anne Tamm
MTA Nyelvtudományi Intézet
Benczúr u. 33.
Pf. 701/518,
H-1399
anne@nytud.hu

This article proposes a way to represent verbal aspect in a Lexical Functional Grammar (LFG) lexicon. The focus is on the modeling of the contribution of transitive verbs in the aspectual composition of an aspectually bounded clause. The article considers Estonian transitive change of state verbs in their interaction with the aspectual case marking of objects. More specifically, clausal aspect is modeled in terms of the unification of the boundedness features in the functional structure. The lexical entries for transitive verbs are provided with specified or underspecified boundedness features in the proposed LFG lexicon. Indications about the specific types of boundedness belong to the functional specifications in the verb entries. As one option, these specifications can have the form of defining equations. In this case, boundedness is specified in the lexical entry of the verb, the verb is bounded. Alternatively, boundedness is underspecified; in that case, the functional specifications have the form of existential constraints. The general well-formedness conditions of LFG secure the sensitivity of aspectual case to verb classification.

1. Introduction

This article proposes a computational aspectual verb classification for Estonian transitive change of state verbs. This classification is meant to accommodate the systematic compatibility of verb classes with certain clausal aspectual object case marking patterns. The Lexical Functional Grammar (LFG) methodology is applied here. Clausal aspect is specified in terms of boundedness. I call a clause or a sentence bounded if it describes an event where clear boundaries are attained. Clausal boundedness is encoded in the form of features at the LFG's syntactic level of f-structure. This article studies those aspect-related attributes and values that transitive verbs contribute to the f-structure. More specifically, the focus is on change of state verbs that are either (inherently, lexically) bounded, that is, perfective, or boundable, that is, imperfective. The lexical entries for transitive verbs are provided with specified or underspecified boundedness features in the proposed LFG lexicon.

2. The problem

The problems of fixing the Estonian verbal aspect in a computational lexicon are manyfold. On the one hand, some verbs regularly allow for variation in their aspectual behavior, as the following examples with the verb *kirjutama* 'write' (1) and (2) demonstrate. Relating the morphological genitive (that is, accusative-total) case marking to definiteness and the partitive case to indefiniteness is, considering the creation verb in these examples, not plausible.

(1) *Mari kirjutas raamatut.*
M.nom write.3.sg.past book.part
'Mari was writing a/the book.'

(2) *Mari kirjutas raamatu.*
M.nom write.3.sg.past book.gen
'Mari wrote a book.'

While sentence (1) with a partitive object is unbounded (in other, comparable terminology, progressive, atelic, or imperfective), sentence (2) with the morphologically genitive accusative-total object is bounded (telic or perfective). Several accounts of the interaction between verbal aspect and clausal aspect (Verkuyl 1993) relate the variation in the aspectual behavior of these verbs to the quantificational properties of the object NP. The examples above show that the expected variation in the aspectual value of the sentence is not paralleled by the difference in the object NP properties: the quantification of the object NP *raamatu(t)* 'book.gen/part' remains constant. The issue of composition is more complicated since the prediction of most theories is that sentences with bare plural nouns are not quantized, they are unbounded. This prediction is not borne out, since sentence (3) can have a quantized, bounded, interpretation regardless of the bare plural (partitive-marked) object.

(3) *Mari* *kirjutas* *raamatuid.*
M.nom write.3.sg.past book.part
'Mari did some book-writing.'

It is clear that instead of the quantification of the object NP, another fact related to the object NP is decisive for boundedness, namely, the object case. The object case alternates between partitive as in the unbounded sentence (1), and the morphological genitive (accusative, total), as in the bounded sentence (2). However, example (3) also shows that there is no ground to relate partitive case marking to unboundedness either. Rather, the three types of case marking seem to bring out three different aspectual possibilities of the verb *kirjutama* 'write': fully bounded (total-accusative), not bounded (partitive), and indefinitely bounded "to some extent" (partitive of divisible objects).

It is also clear that there is no one-to-one correspondence between verbs and sentence aspect, on the one hand. On the other hand, we can observe that verbs cannot fully determine the case marking either. Case marking and aspect or properties of the event are, however, related: verbs determine the possible range of case marking and aspectual possibilities. The proof to the claim that verbs still do determine the relevant basis of case marking and the aspectual nature of a sentence is the existence of verbs (e.g., *vaatama* 'watch') that can have only partitive objects and that do not have any bounded interpretation with the partitive plural object.

(4) *Vaatasin* *etendust/etendusi/*etenduse.*
Watch.1.past.sg play.part/part.pl/*play.gen
'I was watching a/the play/(the) plays.'

An optimal description of verbs must, therefore, reflect the match between verbs, aspect, and case in a flexible way to accommodate the verb's aspectual possibilities with different types of case, on the one hand, and, on the other hand, to fix the sensitivity of case to verb classification.

However, next to the facts of partitive plural objects in bounded sentences, another objection to presuming a strong partitive-unboundedness correlation must be presented before proposing the aspectual features for Estonian transitive verbs. Partitive objects appear in bounded sentences with a class of event verbs; in (5), I illustrate the psych-verb *ehmatama* 'frighten'.

(5) *Mari* *ehmatas* *Jürit.*
M.nom frighten.3.sg.past George.part
'Mari frightened George.'

Psych-verbs resemble degree achievement verbs, such as *laiendama* 'widen' (6) and *pikendama* 'lengthen', which can also have bounded readings in sentences with partitive objects.

(6)*Firma* *laiendas* *kahe tunniga* *teed.*
Firm.nom widen.3.sg.past in two hours road.part
'The firm widened the road in two hours.' (in the sense of to some extent)

The difference between psych-verbs and degree achievement verbs is in the possibility of the accusative-total case on the objects of degree achievement verbs (7) and the impossibility of it on the objects of psych-verbs (8).

(7) *Firma* *laiendas* *tee* *kahe* *tunniga.*
Firm.nom widen.3.sg.past road.gen in two hours
'The firm widened the road in two hours.' (to the full extent)

(8) #*Mari* *ehmatas* *Jüri.*
M.nom frighten.3.sg.past George.gen
'Intended to mean: Mari frightened George.'

In sum, sentences are bounded in two different ways, maximally and minimally. The events described by the verb *kirjutama* 'write', *laiendama* 'widen', and *ehmatama* 'frighten' as illustrated above are, depending on the case-marking on the object, the following: unbounded (1), maximally bounded (2, 7), bounded to some extent, that is, minimally bounded (3, 5, 6).

3. Proposal

I account for the differences in the aspectual interpretation of the above discussed verbs in terms of morphologically constrained mapping from the LFG's f(unctional)-structure to the s(emantic)-structure. Both verbs and case encode grammatically relevant aspectual features.

3.1. The possibilities of the Lexical Functional Grammar framework

Importantly for this account, the LFG framework allows locating pieces of aspectual information and information about grammatical relations in many (discontinuous) constituents that may appear in several configurations in surface constituent structure syntax (c-structure). Simultaneously, it allows locating them at one place at the other syntactic level, the functional structure (f-structure). This effect is achieved by means of constraints that pertain to relations between the levels of representation. My account relies on parts of several previous analyses and methods, basically Tamm (2004). I apply an analysis where the constructive case model is used to account for encoding sentential aspect on dependents (Constructive case in LFG as in Nordlinger and Sadler 2004).

3.2. Boundability and boundedness

The goal here is to provide an account of how the information from lexical entries specifies structures of syntactic representation. The proposal is that lexical entries provide partial but basic information about clausal aspect at the f-structural level of syntactic description. I discuss here change of state verbs and I propose to represent them with a boundedness attribute (B) that is either valueless or with the value (MINimal). Lexical entries encode boundability or boundedness, respectively. Boundedness is also the term for the aspectual features in the f-structure feature matrix, where the B attribute can have the value of MINimal or MAXimal.

Many earlier Estonian accounts suggest treating Estonian aspect in terms of boundedness and verbal boundability, based on the intuition that transitive verbs are either boundable or not. The characterization of change of state verbs is roughly as follows in (9).

(9)

Boundable verbs:	*kirjutama* 'write'
Bounded verbs:	*ehmatama* 'frighten'
Variable verbs:	*laiendama* 'widen'

Sentences or clauses are bounded in two different ways: some verbs are bounded lexically and others, compositionally, e.g., by case-marked objects. Some verbs may have an aspectual feature in their entry; other verbs specify the attribute only, but the value must be provided in the clause. In that way, verbs provide partial aspectual specification in a clause. The maximal boundedness value, MAX, can be specified only compositionally for these verbs, by means of unification. Once an attribute is provided with a value lexically, it cannot be specified by another element in syntax. Elements in syntax can add information but not change attribute-value pairs. Combinations of attributes and values associated with verbs yield distinct verb classes.

In my classification, if a verb is called bounded, then its boundedness feature is specified. Indications about the boundedness of the verb belong to the functional specifications in the verb entries and in the respective terminal node of the c(onstituent)-structure. These specifications have the form of defining equations as in the verb entry of *ehmatama* 'fighten'(10).

(10) **ehmatama**, *V* : (↑PRED)= 'FRIGTHEN <(↑SUBJ), (↑OBJ)>'
(↑B)=MIN

(11)
$$\left[\begin{array}{ll} \text{PRED} & \text{'FRIGHTEN<SUBJ, OBJ>'} \\ \text{B} & \text{MIN} \end{array}\right]$$

In this case, boundedness is specified in the lexical entry of the verb and clausal aspect is determined by the verb. As a result of the mapping from c(onstituent)-structure to f(unctional)-structure, the f-structure is constrained to contain the specified boundedness feature, that is, an attribute with a "fixed" value (11). Having a fully specified feature (a defining equation) as part of its lexical entry, such as (↑B)=MIN, means for the verb that its boundedness is lexicalized, that it is an inherently perfective, bounded verb. Since clausal aspect is modeled in terms of the unification of boundedness features in the f-structure, the failure in unification explains the restrictions on case marking patterns in the model where case contributes different values. This means that these verbs are not boundable any more and the range of aspectual case marking possibilities is restricted.

If verbs are boundable, their boundedness feature is underspecified. They can be bounded, and the range of case marking possibilities is open. Indications about the boundability of the verb also belong to the functional specifications in the verb entry and are present at the terminal verb node of the c-structure. These specifications have the form of existential constraints in LFG as in (12).

(12) **kirjutama**, *V* : (↑PRED)= 'WRITE <(↑SUBJ), (↑OBJ)>'
(↑B)

In this case, clausal boundedness is not determined by the verb (by the lexical entry of the verb) but only as the result of aspectual composition, modeled as the unification of features in the clausal f-structure (13). As a result of the mapping from constituent structure to f(unctional)-structure, the f-structure is constrained to contain only the attribute part of the boundedness feature, that is, an attribute without any value.

(13)
$$\left[\begin{array}{l} \text{PRED 'WRITE <SUBJ, OBJ>'} \\ \text{B} \end{array}\right]$$

Having an existential constraint (↑B) means that the attribute B must be present in the f-structure feature matrix that corresponds to the verb in c-structure. As clausal aspect is modeled in terms of the unification of boundedness features in the functional structure, the possibility of the unification with features with different values explains the wider range of case marking patterns. In my model, the "underspecified" features become fully specified by the features of case-marked objects.

There are verbs (such as the degree achievement verb *laiendama* 'widen' in (14)) that can have two aspectual possibilities; in that case, the entry uses disjunction and both types of functional specifications, an existential constraint and a defining equation.

(14) **laiendama,** *V* : (↑PRED)= 'WIDEN <(↑SUBJ), (↑OBJ)>'
(↑B) V (↑B)=MIN

As a result of the mapping from constituent structure to f(unctional)-structure, two f-structure matrices are possible, (15) and (16).

(15)

$$\begin{bmatrix} \text{PRED 'WIDEN <SUBJ, OBJ>'} & \\ \text{B} & \end{bmatrix}$$

(16)

$$\begin{bmatrix} \text{PRED 'WIDEN<SUBJ, OBJ>'} & \\ \text{B} & \text{MIN} \end{bmatrix}$$

The next question is: given the incomplete f-structure, how will the values be obtained? Before discussing the verbs' contribution to the sentence and the interaction with case-marked objects, I present the features associated with the three types of case markers.

3.3. Inside-out constraints for features associated with case-marked objects

Accusative-total case is the case that encodes maximal boundedness; it appears in sentences that denote an event with clear boundaries and that cannot be continued. The lexical entry of the accusative case contains a defining equation, an inside-out constraint for the maximal boundedness feature, (B↑)=MAX. The entry for the accusative case is presented in (18). An accusative-total case-marked nominal specifies the f-structure information in (19).

(18) ACC: $(\uparrow \text{CASE}) = \text{ACC}$ (total)
$(\text{OBJ} \uparrow)$
$((\text{OBJ} \uparrow)\ \text{B}) = \text{MAX}$

(19)

$$fx \begin{bmatrix} \text{B} & \text{MAX} \\ \text{OBJ} & fy \begin{bmatrix} \text{CASE} & \text{ACC} \end{bmatrix} \end{bmatrix}$$

The indication $(\text{OBJ} \uparrow)$ is the inside-out designator. By virtue of this designator the information associated with the accusative case "constructs" the f-structure of a higher f-structure (f_x). The higher f-structure contains an object to which the immediate f-structure containing the case-marked nominal (f_y) belongs. The association between the nominal and its grammatical function is established by virtue of the case marker attached to it. I leave the semantic constraints that constrain the mapping between the f-structure and c-structure aside.

Partitive is the default case; it encodes only the constraint that the sentence is not maximally bounded (20). A constraint equation captures this constraint on the f-structures.

(20) PART1:$(\uparrow \text{CASE}) = \text{PART}$
$(\text{OBJ} \uparrow)$
$((\text{OBJ} \uparrow)\ \text{B}) =_c \neg\ \text{MAX}$

Partitive object NPs specify the information in the f-structure feature matrix as in (21). If the f-structure matrix contained a B attribute with a MAX value, the structure would be ill-formed.

(21)

$$f_x \begin{bmatrix} \text{OBJ} & f_y \begin{bmatrix} \text{CASE} & \text{PART} \end{bmatrix} \end{bmatrix}$$

Singular mass noun and plural count noun partitive objects are specified as follows (leaving out the semantic restrictions) in (22), and they specify the information in the f-structure feature matrix as in (23).

(22) PART 2:
(↑CASE) = PART
(OBJ↑)
((GF↑)B) = MIN

(23)

$$f_x \begin{bmatrix} \text{B} & \text{MIN} \\ \text{OBJ} & f_y \begin{bmatrix} \text{CASE} & \text{PART} \end{bmatrix} \end{bmatrix}$$

The general well-formedness conditions of LFG secure the sensitivity of aspectual case to verb classification and v.v. For instance, the sentence in (8) is ruled out by such principles. The sentence is ill-formed as a result of a feature clash between the features specified by the accusative-total case, (B↑)=MAX, and the verb *ehmatama* 'frighten', (↑B)=MIN. Partitive 1 and partitive 2 marked objects and the bounded verb form well-formed minimally bounded sentences, since the verb entry constrains the f-structures to have a "minimally bounded" feature, exactly as the semantically restricted partitive 2, and the features are unifiable, and the entry for partitive 1 fixes that the structure should not contain a "maximally bounded" feature, which it does not. The same options explain the behavior of the verb *laiendama* 'widen', but this verb allows for maximal boundedness as well; the unification of the features and attributes of the accusative-total case is successful with an extra existential constraint in the functional specifications of this degree achievement verb. The two types of bounded sentences formed by the verb *kirjutama* 'write', which has an entry with an existential constraint, are also explained: the "minimal" and "maximal" values of the attribute are provided by the case-marked objects partitive plural and accusative-total, respectively.

4. Conclusion

This article proposes a computational aspectual verb classification for Estonian transitive change of state verbs. This classification accommodates the systematic compatibility of verb classes with certain clausal aspectual object case marking patterns. I apply the Lexical Functional Grammar (LFG) methodology. Clausal aspect is understood in terms of boundedness. A clause or a sentence is bounded if it describes an event with clear boundaries. Clausal boundedness is encoded in the form of features at the LFG's syntactic level of f(unctional)-structures. This article studies those aspect-related attributes and values that transitive change of state verbs contribute to the f-structure. The lexical entries for transitive verbs are provided with specified or underspecified boundedness features in the proposed LFG lexicon.

In this framework, if a verb is called bounded, then its functional specifications contain a boundedness feature. This means that these verbs are not boundable any more and the range of aspectual case marking possibilities is restricted. If verbs are boundable, their boundedness feature is underspecified. As clausal aspect is modeled in terms of the unification of boundedness features in the f-structure, the possibility of the unification of features with different values explains the wider range of case marking patterns. In my model, the features become fully specified in the process of the unification with the features of case-marked objects.

Verbs fall into aspectual classes, distinguished from each other according to the pattern of the attributes and values in the functional specifications of the verbs' lexical entries. This verb classification is suitable for accounting for the interaction between Estonian aspect, verbs, and case.

Referenes

Nordlinger, Rachel and Louisa Sadler. 2004. 'Tense Beyond the Verb: Encoding Clausal Tense/Aspect/Mood on Nominal Dependents.' *Natural Language and Linguistic Theory* 22. 597–641.

Tamm, Anne. 2004. *Relations between Estonian verbs, aspect, and case.* PhD thesis, Budapest.

Verkuyl, H. 1993. *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure.* Cambridge: Cambridge University Press.

Computational aspects of an automatic recognizer of Italian clitics

Marco Tomatis
Università degli Studi di Torino
Via Sant'Ottavio 20 10124 Torino
m-tomatis@tiscali.it

Introduction

The aim of this paper is to present the main features of a computational system for the electronic recognition of Italian clitics.

One of the many problems which may be encountered when preparing a corpus[1] for further (automatic or manual) analysis lies undoubtedly in the so-called text tokenization; the splitting of different words (or lexical units) from all those non-alphabetic graphic signs they may be tied to.[2]

When handling morphologically rich languages like Italian, the problem can not simply be solved by using a text pre-processor: the solution may require considering a more specific level, that is, the morphological structure of the very word. Indeed, for a proper interpretation of the linguistic data of a corpus, the researcher is often obliged to extend the tokenization process even within different words in order to isolate the very word from the clitic it is tied to.

Such a process is, in most cases, very difficult to manage; moreover, when handling large corpora, it has to be done automatically as far as possible. So, to make these tasks easier, an automatic clitic recognising program has been developed.

The system, wholly implemented using a procedural scripting language named "GAWK"[3], acts on an already tokenized text. For it to work correctly, it requires a very complete list of flexed Italian words; obviously without clitics. Since the system is basically founded on linguistic rules, the existence of such a list of words provides a rapid way to check the linguistic hypotheses that are inferred by the rules themselves.

[1] Barnbrook (1996); Kennedy (1998).

[2] Grefenstette (1999)

[3] For the complete guide, please read the official Gawk manual:
GAWK: Effective AWK Programming: A User's Guide for GNU Awk. 3rd edition. Free Software Foundation, Inc. 2001.
The whole manual can be freely downloaded from the address:
http://www.gnu.org/software/gawk/manual/gawk.html
It is also available online at the address: http://it.tldp.org/man/man1/awk.1.html

Methodological approach

The methodological approach, which has been adopted to develop the system, is in some ways innovative. Studies conducted to date on the clitics phenomenon focused attention largely on the leading element, the verb; consequently the clitics issue has been examined mainly from a syntactic or semantic point of view (Borer, 1986) or, in some cases, by adopting a lexical framework which tried to reconduct the behaviour of clitics to that of suffixes (Monachesi, 1999). The approach discussed in this paper, instead, has moved the focus towards the enclitic piece of word only, trying as far as possible to track down the main features of Italian clitics according to their ability to select the verb tense they are attached to. The study of Italian clitics form and behaviour led to drawing up a resumptive table of their main features. For greater terminological clarity, in this paper the general term "clitic" is referred both to simple particles and to multiple clitics chains structured starting from simpler bits, unless stated otherwise.

Clitics	Conjugation	Tense	Number of letters	Final Letter
ccela	1 - 3	imperative	2	A - I
ccele	=	=	=	=
cceli	=	=	=	=
ccelo	=	=	=	=
ccene	=	=	=	=
mmela	=	=	=	=
mmele	=	=	=	=
mmeli	=	=	=	=
mmelo	=	=	=	=
mmene	=	=	=	=
mmici	=	=	=	=
mmiti	=	=	=	=
ttela	=	=	=	=
ttele	=	=	=	=
tteli	=	=	=	=
ttelo	=	=	=	=
ttene	=	=	=	=
cci	=	=	=	=
lla	=	=	=	=

lle	=	=	=	=
lli	=	=	=	=
llo	=	=	=	=
mmi	=	=	=	=
mme	=	=	=	=
tti	=	=	=	=
gli	1 - 2 - 3	infinitive gerund imperative past participle	> = 2	A - E - I - O - R
gliela	=	=	=	=
gliele	=	=	=	=
glieli	=	=	=	=
glielo	=	=	=	=
gliene	=	=	=	=
ci	=	=	> 2	=
cela	=	=	=	=
cele	=	=	=	=
celi	=	=	=	=
celo	=	=	=	=
cene	=	=	=	=
mi	=	=	=	=
mela	=	=	=	=
mele	=	=	=	=
meli	=	=	=	=
melo	=	=	=	=
mene	=	=	=	=
ti	=	=	=	=
tela	=	=	=	=
tele	=	=	=	=
teli	=	=	=	=
telo	=	=	=	=
tene	=	=	=	=
vi	=	=	=	=
vela	=	=	=	=
vele	=	=	=	=
veli	=	=	=	=
velo	=	=	=	=
vene	=	=	=	=
mici	=	=	=	=

tici	=	=	=	=
glici	=	=	=	=
leci	=	=	=	=
vici	=	=	=	=
la	=	=	=	=
le	=	=	=	=
li	=	=	=	=
lo	=	=	=	=
ne	=	=	=	=
si	=	indicative infinitive gerund present part. past participle subjunctive	=	A - E - I - O - N - R
sela	=	infinitive gerund past participle	=	A - E - I - O - R
sele	=	=	=	=
seli	=	=	=	=
selo	=	=	=	=
sene	=	=	=	=
misi	=	=	=	=
tisi	=	=	=	=
glisi	=	=	=	=
lesi	=	=	=	=
cisi	=	=	=	=
visi	=	=	=	=

Table 1

The table above is divided into five fields. The first column on the left includes the list of all the standard modern Italian clitics. The second and the third fields include respectively the conjugation and the tense of the verb which the clitic can be attached to. The fourth field, instead, lists the syllabic structure of the leading verb by specifying the exact number or the minimum threshold of letters which it has to be made of in order to be selected by a particular set of clitics. Finally, the fifth field lists the different verbal flexions which the clitic can be tied to.

Analysis of data

The table shows that Italian clitics can be divided into two main groups; those starting with a geminate consonant and all the rest, which may be further classified following their own intrinsic features.

Although it is quite easy to spot the rules useful to handle the set of clitics that start with a geminate consonant, for the bigger group made of simple and articulate clitics a more thorough examination of their verb selection behaviour is required.

After a detailed analysis of the data displayed in table 1, it is possible to group the clitics into homogeneous macro areas. A first area should include the clitic "gli" and its articulate forms "gliela", "gliele", "glielo", "glieli" and "gliene". Such clitics show a very interesting behaviour; like a hybrid entity they share the features of both the clitics having a geminate consonant (which take monosyllabic imperatives) and a great part of the remaining. Yet two forms which do not belong to such set exist; they are "glici" and "glisi". In fact the latter behave differently from the other compound clitics starting with "gli"; in particular they cannot be tied to monosyllabic verbs. For this reason they should be included into different areas.

Another large group includes the simple clitics "ci". "mi" "ti" "vi" and their articulate forms which take the pronouns "lo" "la" "li" "le" and the adverb "ne". Such set of clitics is different from the previous one because it can not take monosyllabic imperatives but only plurisyllabic ones, plus verbs conjugated in the infinitive, gerund and past participle forms. This group may also include the subset filled with those compound clitics starting with a personal pronoun "mi" "ti" "gli" "le" "vi" and ending with the second adverbial element "ci", though such particular combinations appear more and more rarely in contemporary Italian (i.e. "porta***mici***").

Another area includes the simple form "si" only. This clitic proves to have a different selectivity level compared to the other ones because it takes a wider range of verb conjugations. As a matter of fact it can be tied to the infinitive, gerund, present and past participle and the third person singular and plural of the simple present tense (e.g. "compra***si***", "vendon***si***", etc.), plus the third person singular of the present subjunctive tense (e.g. "legga***si***", etc.).

Finally, the last clitics that may be grouped in a homogeneous set are the compound forms of the personal pronoun "si"; they can only take verbs conjugated in the infinitive, gerund and past participle forms.

The kind of approach described so far has helped to formulate a set of handy rules to distinguish, within the specific lexical entry, what is to be considered a real clitic from what is simply a bare segment of the whole word (e.g. vedi***ne*** - "see of it" vs. pedi***ne*** - "pawns"). A brief description of the general setting of the system and its recognition rules follow.

Software architecture

The automatic Italian enclitic tracking system “ClitRec” examines the longest clitic blocks first. When the whole set of rules returns a positive result, the clitic particle is divided from the word itself and marked by a univocal sign, then the result is printed out and the control routine moves to the next word in the text line. Otherwise what may happen is that after ending the entire analysis procedure, the system cannot match the hypothetical clitic with the list of possible Italian clitics. So the whole word is printed with no changes and the routine starts again from the next lexical unit. Inversely, once the software successfully matches the probable clitic with one in the list, the linguistic validity of the clitic is further evaluated against the set of linguistic rules. After that, the software may print out the positive result or it may carry on its task until it finishes all the linguistic material to be examined.

Pre-processing activities

In order to enhance the computational efficiency of the system discussed in this paper, as well as optimizing the algorithm used in the clitic recognition activity, a number of pre-processing rules have been defined. Since their main task is to filter out irrelevant or noise producing lexical material, the said rules accomplish the following actions:

- Filtering of all those words whose final part does not match with a possible clitic.
- Filtering of those words containing clitic elements which are ambiguous due to the lack of graphic stressing marks (e.g. “mangia***telo***” vs. “mangiate***lo***”, “guarda***tene***” vs. “guardate***ne***”)
- Filtering of those forms identical to compound prepositions (e.g. “da***llo***”, “da***gli***”, etc.)
- Checking of the existence in the Italian lexicon of the piece of word obtained by depriving the whole word of the hypothetical clitic element. The word would be excluded from further analysis if this check returned a negative result. This check is not run both on the verbal forms conjugated in the infinitive form and in the third apocopate plural person of all the tenses involved.
- Exclusion of those words containing probable, but not real, clitics.
    - In the case of a word ending with the vowel “a” (e.g. “peda***la***”, “affi***la***”, etc.) the checking system provides for the addition to the end of the word of the part corresponding to the infinitive verbal

flexion, followed by a check of the existence of such new form into the Italian lexicon ("peda**la*re***", "affi**la*re***", etc.)

- Cases different from the one above require further tests to ascertain the existence of the whole word in the lexicon. The system will exclude the existence of a clitic in the word if, after substituting its final vowel with those bearing a particular value of gender and number, the new form proves to be part of the Italian lexicon used by the software. (e.g. "fifo***ne***" - "fifon**a**" "fifon**i**"; "feri***ti***" - "ferit**a**" "ferit**e**" "ferit**o**")

Analysis rules

After having described the general structure of the software and the first stage filter rules, a sample description of some linguistic rules used by the program follow.

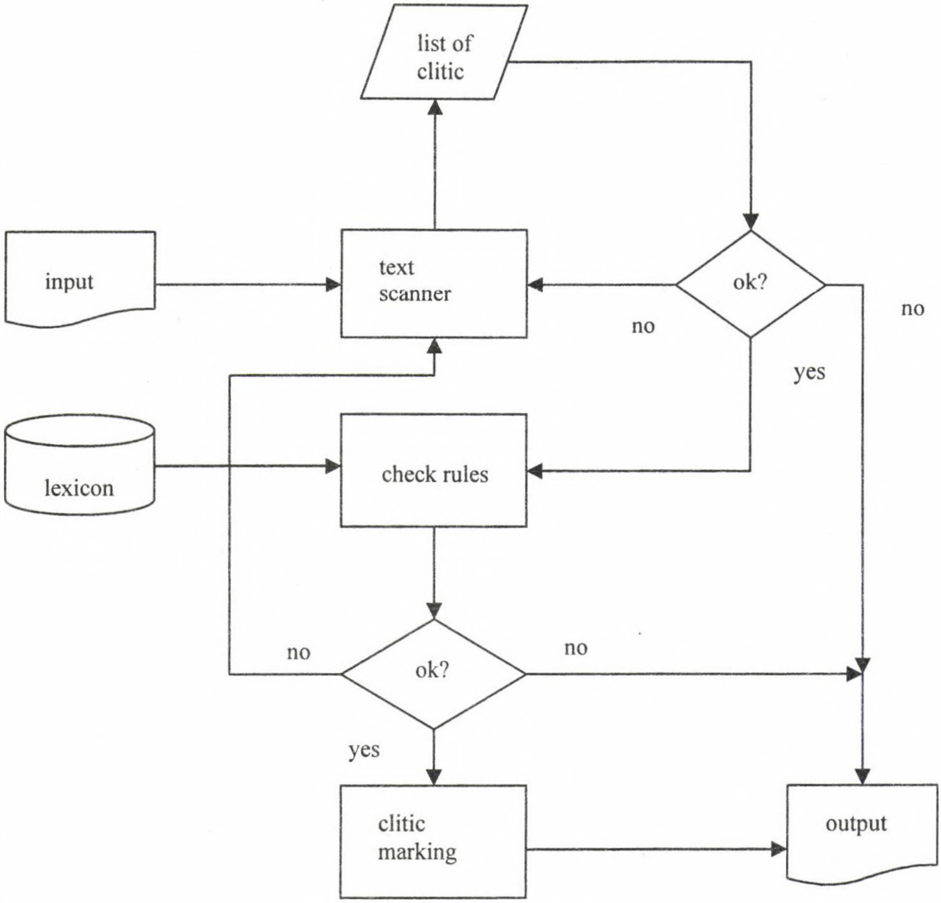
- Clitics starting with geminate consonant: they can only take monosyllabic imperative verbs or their iterative forms (e.g. "di***mmi***", "*ri*di***mmi***", etc.). This rule avoids automatically analysing those lexical forms which prove ambiguous due to their transcategorization (i.e. "fa***lla***", "fa***llo***", etc.)[4]
- Clitic "**gli**", simple form. It is always recognised as a clitic when found tied to the monosyllabic imperatives "di", "da", "fa" and their own iterative forms. The imperative conjugation of the verb "andare" ("va") has not been taken into account by this rule in order to avoid ambiguity at morphologic level between that word and the present subjunctive singular of the verb "vagliare" ("vagli")
- Clitic "**gli**", compound forms. They are always recognised as clitics when found tied to the monosyllabic imperatives "di", "da", "fa", "va" and their own iterative forms.
- Clitic "**ne**": rule to clear the ambiguity with augmentative and diminutive forms. This check acts substituting the last vowel of the piece of word in exam, previously deprived of the potential clitic, with the 2nd and 3rd conjugation of the gerund flexion ("endo"). For clarity's sake let's examine the following words as examples: "prendi***ne***" and "costi***ne***". Adopting the criterion explained before, the particle "ne" belonging to "costine" will be never analysed by the system as a clitic; instead it will the "ne" part of "prendine" simply because within the Italian reference lexicon this particular rule will find the word "prend**endo**", while something like "cost**endo**" will never be found. Moreover, it is important

[4] The term "transcategorization" means that a particular word can belong to different part of speech (or grammar categories). In the example given, the term "falla" belongs to both categories of nouns and verbs.

to notice that, in order to avoid misleading analysis due to non-existing words that may be generated by spelling errors (i.e. "disti***ne***" instead of "distinte"), the program does not limit itself to the tests on the lexicon discussed before, but it runs a specific one which substitutes the gerund flexion with the 2nd and 3rd conjugation of the simple imperfect tense flexion. Even though this further check may seem redundant, it is in fact extremely significant and useful. Although the last example takes into account a badly formed Italian word, the part of word that is deprived from the potential clitic particle ("disti") is exactly the same as the 1st, 2nd and 3rd persons singular of the present subjunctive of the verb "distare". Since the lexicon is not supposed to be a POS (part of speech) tagged text, the system could not but recognize the particle "ne" as a real clitic, even though such kind of clitics can not tie themselves to verbs conjugated in the subjunctive. So, given these premises, if the rule simply substituted the final vowel "i" with the whole gerund flexion "endo", the newly formed word "dist**endo**" would correspond to the 1st person singular simple present conjugation of the verb "distendere", which would again lead to the wrong result of validating the hypothesis that the particle "ne" is a real clitic and not a mere part of the word. As a matter of fact, only the test which substitutes the gerund flexion with the 2nd and 3rd persons singular of imperfect (*"dist**iva**"; *"dist**eva**") and then use the Italian lexicon to test their existence allows the system to avoid such a mischievous error. Finally, it is important to remark that the test uses both the flexion of 2nd and 3rd conjugation of the simple imperfect tense because it is not possible to infer the correct conjugation a verb belongs to simply by examining its imperative form (e.g. "bevi***ne***" "bev**endo**" "bev**eva**" - "apri***ne***" "apr**endo**" "apr**iva**").

- Rules for clitics which are tied to morphologically irregular verbs in the imperative and subjunctive tenses[5] (i.e. "sappi" - "sappia", "siedi" - "segga", etc.). Since in the Italian lexicon such verbs are very few in number, if only the piece of word deprived of the potential clitic should be found in the reference lexicon, but not the complete word, it is possible to infer with a good error margin that the specific particle under exam is a real clitic, not merely the final part of a word.

[5] In Italian, when moving from the subjunctive to imperative tense, regular verbs simply change their final vowel only. The 1st person conjugation provides for the verb to take a flexion "a" in the imperative form (i.e. "guard**a**") and a flexion "i" in the subjunctive tense (i.e. "guard**i**"). On the contrary, the 2nd and 3rd conjugations behave in the opposite way; verbs take a "i" in the imperative tense (i.e. "prend**i**") and an "a" in the subjunctive tense (i.e. "prend**a**"). All those verbs which do not behave in the standard way must be managed using specific rules.



Flow chart of the clitics recognizing system

Future developments

The system described in this paper is still in working progress; its quality could be improved by using a preliminary stochastic part of speech tagger in order to disambiguate words which could not be treated in other ways (i.e. "mangia***tene***" - "eat some of it" (2nd person singular) vs. "mangiate***ne***" - "eat some of it" (2nd person plural). The said tagger should work together with a morphological analysis system, which is currently being developed, to help recognize and divide possible prefixes from the lexical stem (i.e. "**stra**dilungarsi" - "to linger unduly") in order to free the clitic recognizing system from the need to use a control lexicon of all the words existing in the Italian language.

Conclusion

This paper has described a system which helps to track the pronominal and adverbial enclitic part of word in a non POS tagged corpus. Even though this system is not based on stochastic inference functions, a complex set of rules enables us to reach a rather high analytical level. Finally, the adoption of a comprehensive Italian lexical database helps the different rules to optimise their inference.

The program has been written using a scripting language named "GAWK" which allows the developer to create a fast, portable, easy to maintain program which does not require the installation or running of complex procedures by the final user.

Bibliographic References

Aarts J. and Meijs W. "Theory and practice in corpus linguistics." Amsterdam & Atlanta: Rodopi 1990.

Barnbrook G. "Language and Computers. A Practical Introduction to the Computer Analysis of Language." Edinburgh: Edinburgh University Press 1996.

Borer H. (ed .), The Syntax of Pronominal Clitics, "Syntax and Semantics" 19, New York : Academic Press, 1986

Calabrese, A. "I pronomi clitici." In: L. Renzi (ed.) Grande Grammatica Italiana di Consultazione. Vol.1. Il Mulino, Bologna, 1988

Grefenstette, Gregory. "Tokenization." In van Halteren, chap. 9, pp. 117–133 1999.

Kennedy G. "An Introduction to Corpus Linguistics." Longman: London & New York 1998.

Monachesi, P. "A Lexical Approach to Italian Cliticization." Stanford: CSLI Publ. 1999.

Simpson J. & M. Withgott, "Pronominal clitic clusters and templates" in Borer (ed.) 1986: 149-174

Reverse Lemmatizing of the Dictionary of Middle Dutch (1885-1929) Using Pattern Matching

John van der Voort van der Kleij

voortkleij@inl.nl

Institute for Dutch Lexicology

Leiden, The Netherlands

Abstract

The Integrated Language Database (ILD), a project of the Institute for Dutch Lexicology (See: http://www.inl.nl), will contain various kinds of Dutch language data from the earliest to the most recent periods, such as electronic dictionaries, text files and files with linguistic data (lexica). There will be a well-balanced selection of sources, linguistic annotation of the texts and the linking of sources. This projected database will be a research tool for various aspects of the Dutch language and culture throughout the centuries.

The Dictionary of Middle Dutch (MNW), now available in electronic form, is a classic lexicographic source of nine large volumes that will be incorporated in this database. To link the ca. 74000 entries of this dictionary with the corresponding entries of other lexicographic sources in the database (for example the dictionary of the Dutch Language on historical principles, WNT) modern Dutch entry forms are being added.

Our paper concerns the links between Middle Dutch wordforms (tokens) in their context with the entries in the dictionary. For lemmatizing Middle Dutch texts we need a lexicon covering the paradigms of the entries. To build such a lexicon we developed a sophisticated pattern matching program that links the entry forms with their paradigmatic types in the quotations. Basis for the matching are the dictionary entries and their listed variants.

A wider perspective is, that this lexicon of paradigmatic forms may be extended into a morpholexical computer lexicon. Of course part of speech information needs to be added. Other Middle Dutch dictionaries will also be exploited, like the Dictionary of Early Middle Dutch (VMNW) and the concise one volume dictionary of Middle Dutch, an excerpt of the MNW and for many articles a revision of its source.

1. The Integrated Language Database of 6th-21st Century Dutch

The Integrated Language Database (ILD; Kruyt 2004), a long-term project of the Institute for Dutch Lexicology (INL, see http://www.inl.nl), will contain various kinds of Dutch language data from the earliest to the most recent periods, such as electronic dictionaries, text files (diachronic text corpus) and files with linguistic data (lexica). There will be a well-balanced selection of sources, linguistic annotation of the texts (in particular with part of speech and headword) and the linking of sources. This projected database will be a web-accessible research tool for various aspects of the Dutch language and culture throughout the centuries. Links with digital data collections of other institutes are foreseen to create a supra-institutional research instrument.

1.1. The ILD dictionary component

One of the three components which will be mutually linked, is the dictionary component dealing with the vocabulary from 1200 up to 1976. The most comprehensive dictionaries of the Dutch language (and some smaller, supplementary dictionaries) will be incorporated: the Dictionary of Old Dutch (ONW, covering ca. 500-1200, on-going INL project); the Dictionary of Early Middle Dutch (VMNW, covering 1200-1300, 4 volumes, finished INL project); the Dictionary of Middle Dutch MNW (covering 1250-1550, 11 volumes); the Dictionary of the Dutch Language (WNT, covering 1500-1976, 43 volumes, finished INL project) and the Dictionary of Standard Dutch (ANW, covering 1970-2020, ongoing INL project). These linguistic dictionaries have been compiled by many generations of mainly Dutch scholars since the late nineteenth century.

1.2. MNW introduced

The MNW, started in 1885 and completed in 1929, is a classic lexicographic source for the study of Middle Dutch in nine large volumes. Vol. X describes the sources used and vol. XI is a specialised addendum mainly about water management. This privately financed work will be an important part of the ILD dictionary component as it covers three centuries. It was published on a (read-only) CD-ROM together with many Middle Dutch texts (CD-ROM Middle Dutch 1998). The dictionary files (including markup ca. 130 MB) became computer-processable recently.

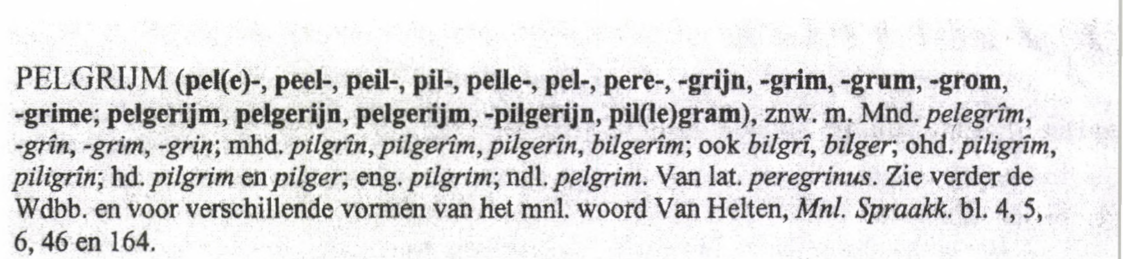
To link the 74758 entries with quotations of this dictionary with the corresponding entries of other lexicographic sources in the database, for example the WNT, modern Dutch entry forms are being added to the entries (in archaic spelling). When modern equivalents do not exist, etymologically justified entry forms are created.

2. Reverse lemmatizing of the MNW by pattern matching

Our paper concerns the linking of Middle Dutch wordforms (tokens) in their context (i.e. quotation) with the entries in the dictionary. Thus, we aim to lemmatize in reverse order, reconstructing, one might say, the normal lemmatization of the MNW editors.
For lemmatizing Middle Dutch texts we need a lexicon covering the paradigms of the headwords and their variants. To build such a lexicon we developed a cumulative pattern matching Perl program (cf. **2.4.**) that links the entry forms with their paradigmatic types in the dictionary quotations. Basis for the matching are the dictionary entries and their listed variants. There are many different variants, as Middle Dutch is not a homogeneous language but a collection of dialects that have grown in the course of four centuries into a more standard national language with a less varying orthography. Variants (and headwords) are often listed in a shortened form and need manual expansion before they can be used.

2.1. Headword and variants in shortened notation

Shortened notations are constructed with parentheses and/or hyphens. We give some examples to illustrate this. The entry EERSA(E)MLIJC 'honest' has the printed variants: *(eersam(e)lijc), -like*. This expands into the headword forms and variants: *eersamlijc, eersaemlijc, eersamlijc, eersamelijc, eersamlike, eersamelike*.
The entry ERFVOREWAERDE 'hereditary contract' has the printed variants: *(erfvorwarde, -voor-, voirwaerde)*. This expands into: *erfvorwarde, erfvoorwarde, erfvoirwaerde*.
Sometimes the hyphen is absent in the digitized version, as for the variant *–doem* of entry HEILICHDOEM 'shrine' with the electronic variants : *helich-, hillich-, heilic-, doem, -echdoem, -doom, -dom* . The expanded variants are: *helichdoem, hillichdoem, heilicdoem, heilechdoem, heilechdoom, heilechdom*. Variants of less then five characters therefore needed a check on missing hyphens.
An exceptional case is the entry PELGRIJM 'pilgrim'. We give a scan of the head of the article:



PELGRIJM **(pel(e)-, peel-, peil-, pil-, pelle-, pel-, pere-, -grijn, -grim, -grum, -grom, -grime; pelgerijm, pelgerijn, pelgerijm, -pilgerijn, pil(le)gram)**, znw. m. Mnd. *pelegrîm, -grîn, -grim, -grin*; mhd. *pilgrîn, pilgerîm, pilgerîn, bilgerîm*; ook *bilgrî, bilger*; ohd. *piligrîm, piligrîn*; hd. *pilgrim* en *pilger*; eng. *pilgrim*; ndl. *pelgrim*. Van lat. *peregrinus*. Zie verder de Wdbb. en voor verschillende vormen van het mnl. woord Van Helten, *Mnl. Spraakk.* bl. 4, 5, 6, 46 en 164.

Even for humans it is difficult to expand the numerous variants correctly: *pelgrime, pelegrime, peelgrime, peilgrime, pilgrime, pellegrime, peregrime, pelegrom, pelgrom, peelgrom, peilgrom, pilgrom, pellegrom, peregrom, pelegrum, pelgrum, peelgrum, peilgrum, pilgrum, pellegrum, peregrum, pelegrim, pelgrim, peelgrim, peilgrim, pilgrim, pellegrim, peregrim, pelegrijn, pelgrijn, peelgrijn, peilgrijn, pilgrijn, pellegrijn, peregrijn, pelegrijm, pelgrijm, peelgrijm, peilgrijm, pilgrijm, pellegrijm, peregrijm, pilgerijm, pelgerijn, pelgerijm* (listed twice!), *pilgerijn* (the preceding hyphen is erroneous), *pilgram, pillegram.*

2.2. Fuzzy pattern matching with the tool Nr-grep?

The programming language Perl was the choice for pattern matching (cf. **2.**). An alternative could have been Nr-grep, abbreviation for: nondeterministic reverse grep (recent version 1.1.1; to be downloaded for free from: http://www.dcc.uchile.cl/~gnavarro/pubcode/). This tool was developed specially for fuzzy pattern matching (Navarro 2000). It permits four transpositions (one more than similar tools): insertion, deletion, substitution and transposition. The allowed number of each type of transposition is customizable. Another attractive quality is that this tool has a very good overall performance.We carried out some experiments to test the usability: unfortunately it proved to be unsuitable for the form variation of Middle Dutch dialects.

2.3. Use of headwords and listed variants (basis for matching)

Articles are processed in dictionary order. Plain headwords are extracted by the program from each of the 58066 MNW articles with quotations. The listed variants (ca. 11000, including expanded headwords and variants) are read for each entry from a database file in which they were stored during a preparatory pass. First these forms are used to find an exact string match in the quotations of an article. Next, when there is no succesful match, a number of substitutions is used to transform these forms into regular expressions with character alternation(s) at special positions. Then the program tries to match the resulting regular expressions against the words of a quotation. This will be explained in more detail in the following section

2.4. Regular expression matching with Perl

The pattern matching program (or script) is written in the versatile computer language Perl (version 5.8, see Wall e.o. 2000, in particular the rewritten chapter 5) and uses its powerful regular expression capabilities to manipulate strings of characters (for a detailed exposition see the standard textbook Friedl 2002, translated in French, German, Russian, Polish, Japanese and Korean).

Several types of expressions are used (examples will be given hereafter):

- expressions anchored at the beginning of a string, useful for prefixes or left word elements;
- unanchored expressions, useful for the substitution of a complete string;
- expressions anchored at the end of a string, useful for suffixes or right word elements;
- expressions with global substitutions, useful for as many substitutions as possible within a string;
- expressions with lookbehind and lookahead, special options only available in Perl (since vs. 5.8);
- expressions to match discontinuous forms, special for a.o. Dutch and German.

The first program action is to try an exact string match of the headword or its (expanded) variant(s), the searchword, on the words in the quotation(s) of each entry. This is fast and simple. If no match is found, a range of cumulative regular substitutions is applied to the searchword to transform it into a regular expression with character alternation(s) at special positions. Regular expression matching on the output with the UNIX tool egrep (we used the GNU version) helped to develop these while inspecting unmatched quotations (an iterative process). Per dictionary volume about 350 substitutions have been formulated.They are sometimes hierarchically ordered to prevent unwanted expansions.

The second program action is to find a match for the resulting extended regular expressions in quotations without a match. Special rules are formulated to recognize frequent discontinuous forms and to mark them as such (cf. **2.4.2.**).

Finally, when no match is found, the program checks for some frequent patterns to detect incorrect encoding of a quotation or an abbreviated form of the headword. Tags are used to mark these 'faulty' quotations (cf. **2.5.**).

2.4.1. Examples of regular expressions for substitution

The part between the first and second slash in the examples below will be replaced by the part between the second and third slash. The pattern modifier "i" after the third slash indicates case-insensitive pattern matching, the pattern modifier "g" indicates global matching.
- hierarchical complexity; this substitution of "aenvaerden" 'to accept' (anchored at the beginning and end of the string with optional prefix "g(h)e" for past participles) has to be executed before the next one for "aen" 'to', otherwise it will be missed:
s/^aenvaerden$/((aen|ane)vaerden?|anferden|(aen|ann?e)(gh?e)?vae+i?r[td]|(aen|ann?e)(
gh?e)?ve+rt)/i;

- anchored at the beginning of a string; substitute "aen" 'to' into an expression with occurring variations, including an optional prefix "g(h)e" for past participles:
s/^aen/(a[e]?n[e]?|anne)(gh?e)?/i;

- not anchored; useful for "tekenen" 'to sign' and e.g. for "aentekenen" 'to note'
s/teken/te[iy]ken/i;

- anchored at the end of a string; substitute "gaen" at the end of a word into an expression with occurring variations; useful e.g. for all compounds with "gaen" 'to go':
s/gaen$/((gh?e)?gaen|ga(ne)?|g[ia]n[gc+(e|t|ic|s)?|ga?e+ts?|gae?[ns]?|g[ei]+n[cg](ic)?|gi
ngen(ic|si)?)/i;

- global substitutions; typically used for variant spellings of the long vowel a:
s/ae/(ae|ai|aij|ay|aa|a)/gi;

- lookbehind and lookahead; substitute all occurrences in a headword/variant of "i" not preceded by a, e, i, o, u or y and not followed by a, e, i, j, o, u, y or [(the bracket indicates a present character class notation) by the alternation i, y or ij:
s/(?<![aeiouy])i(?![aeijouy\|])/(i|y|ij)/gi;

- matching discontinuous forms; if the headword/variant e.g. is "aenleggen" 'to construct', then substitute "aen" into "aen " (space at the right!) in order to match (see **2.4.2**) the paradigm of "legghen" 'to put'; that paradigm has a substituting paradigm later in the program:
$lem[2]=~ s/^aen/aen /i.

2.4.2. The resulting expanded regular expression

These regular expressions (cumulated by regular substitutions) are *not for human reading*. The computer processes them fast. Sometimes, unfortunately, we have to inspect them to check correct execution.

Example "aenleggen" 'construct' transformed into "aen" + "leggen" (discontinuous elements!).

Expansion of "aen" and optional 'g(h)e' (pipes indicate alternatives):

```
((ae|ai|a(ij|y|i+)|ay|aa|(a|ae|ai|aa))n|(a|ae|ai|aa)n|(a|ae|ai|aa)n(e|ei|ee)|(a|ae|ai|aa)nn(e|ei|ee)|
(ae|ai|a(ij|y|i+)|ay|aa|(a|ae|ai|aa))n(e|ei|ee))(gh?(e|ei|ee))?
```

Expansion of "leggen":

```
(l(e|ei|ee)[(c|ch|ck|c?k)gh?]gh?(e|ei|ee)n?|l(e|ei|ee)gh?(e|ei|ee)?t|l(e|ei|ee)gh?s|l(e|ei|ee)[yi
]+[dt](s(e|ei|ee)|d(i|y|(ij|y|i+)))?|l(e|ei|ee)[yi]+d(e|ei|ee)n?|l(e|ei|ee)cht?(e|ei|ee)?|gh?(e|ei|
ee)l(e|ei|ee)[yi]+[td]?|gh?(e|ei|ee)l((a|ae|ai|aa)|(e|ei|ee)+)cht|gh?(e|ei|ee)l(e|ei|ee)ght|gh?(e
|ei|ee)l(e|ei|ee)gh?(e|ei|ee)[nt])
```

Example: mark discontinuous forms in the quotation with 1 or 2 followed by an @-sign:
Milen 20 dusentech ende vier hondert, daer toe *2@lech* 29 milen der *1@an*, dien ommeganc hevestu dan,

2.5. Faulty (encoded) quotations

When no match has been found in a quotation, a pattern matching check is executed to detect incorrect encodings in the dictionary file or an abbreviated headword/variant. A tag is inserted in the output if the check was positive. Eight pattern variations for faulty quotations have been formulated in the program and one for an abbreviation problem. Afterwards the tags will be used to correct the quotation encodings of the MNW.

AENSIEN <i>Chron</i> <FT5>: this is a part of a source title;
AENSTAL In het <FT5>: this the start of a source indication;
AENTALEN <i>van</i> 1329 <i>aang. bij</i> <FT2>: this is a part of a source title;
AENTASTEN Dl.2, <FT7>: this indicates a volume of a source;
BERADEN E. pine b., <FTABR>;
BERADEN E. scade ende verdriet b., <FTABR>.

2.6. Results for vols. I/IX

The bracketed numbers represent the results of exact string matching, the others the results with cumulative pattern matching:

```
I       COUNTED 34956 [35454] MATCHED 32522 [18637]: 93.0 % [52.5 %]
II      COUNTED 49927 [50632] MATCHED 46541 [23523]: 93.2 % [46.4 %]
III     COUNTED 39334 [39859] MATCHED 36473 [24022]: 92.7 % [60.2 %]
IV      COUNTED 44707 [45169] MATCHED 41962 [28033]: 93.8 % [62.0 %]
V       COUNTED 35263 [35700] MATCHED 32545 [18372]: 92.2 % [51.4 %]
VI      COUNTED 24766 [25437] MATCHED 23149 [16657]: 93.4 % [66.2 %]
VII     COUNTED 43391 [44412] MATCHED 40530 [27168]: 93.4 % [65.4 %]
VIII    COUNTED 44429 [45427] MATCHED 40646 [22083]: 91.4 % [48.6 %]
IX      COUNTED 56531 [58241] MATCHED 51631 [33390]: 91.3 % [57.3 %]
```

A comparison of the results for exact string matching and for cumulative pattern matching shows that the latter is superior: exact matching produces a score between 46.4 % and 65.4 %, while cumulative pattern matching scores between 91.3 % and 93.8 %. When a headword has less than five characters, errors can be expected and therefore a manual check has to be performed on the output in that case. Apart from that, an inspection of a random sample taken from the first three processed volumes showed that there was an error rate between 1 and 1.5 % in the matches. This makes correction less imperative.

2.7. Coverage of types (A / R)

Below we present the coverage of the resulting type list in rhyming and prose text.
Note that a rough filtering has been applied for mark up tags, Roman numerals and French and Modern Dutch text.
Rhyming texts (total: 84231 types) of *Cd-rom Middelnederlands*:
Freq. 1 COUNTED 42979 MATCHED 11462: 26.66 %
Freq. > 1 COUNTED 41252 MATCHED 21083: 51.10 %
Prose texts (total: 88458 types) of *Cd-rom Middelnederlands*:
(many of them not used for the MNW):
Freq. 1 COUNTED 45077 MATCHED 8065: 17.89 %
Freq. > 1 COUNTED 37965 MATCHED 18697: 43.09 %
To explore the coverage of the proposed lexicon, a type list was extracted from quotations with a match and compared with a type frequency list of real Middle Dutch text types (beginning with the letters A-R). These text types have been collected from rhyming and prose texts of the CD-ROM Middle Dutch. After roughly filtering these texts for mark up tags, Roman numerals, French and Modern Dutch text, we got the following results (looking at types with a frequency of 2 or higher): the types of the rhyming texts were covered for 51.10 %, the types of the prose texts for 43.09 %. Notice that a substantial part of the prose texts was not excerpted for the MNW - this reduces the number of possible matches essentially - and until now six processed volumes have been used.

3. Future work

From a wider perspective, this lexicon of paradigmatic forms is meant to be extended into a morpholexical computer lexicon. Of course, part of speech information needs to be added. Later on Middle Dutch dictionaries should deliver additional materials (a. o. new entries and/or variants): the concise one-volume dictionary of Middle Dutch (MNHW, an excerpt of the MNW and a revision of its source in the case of many articles; it delivers in fact much of the supplement already promised in vol. I), the supplement to this dictionary (MNHWS), especially for eastern Middle Dutch (underrepresented in the MNW), and the Dictionary of Early Middle Dutch (VMNW). Then more recent lexicographic research on Middle Dutch can also be incorporated in the ILD.

References

ANW. *Algemeen Nederlands Woordenboek*. Instituut voor Nederlandse Lexicologie. In preparation.

CD-ROM Middle Dutch (1998), *Cd-rom Middelnederlands: Woordenboek en teksten*. Instituut voor Nederlandse Lexicologie, Sdu Uitgevers/Standaard Uitgeverij, Den Haag/Antwerpen.

Friedl, Jeffrey E. F. (2002), *Mastering Regular Expressions*. O'Reilly [2nd ed.; with excellent index]

ILD: Kruyt, J.G. (2004), 'The Integrated Language Database of 8th - 21st-Century Dutch'. In: M.T. Lino, M. F. Xavier, F. Ferreira, R. a.o. (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation*; ELRA, Paris , pp. 1751-1754.

MNHW (1932), *Middelnederlandsch handwoordenboek*. Bewerkt door J. Verdam. [...] van het woord *Sterne* af opnieuw bewerkt door C. H. Ebbinge Wubben. M. Nijhoff, 's-Gravenhage 1932. [3rd ed.]

MNHWS (1983), Voort van der Kleij, J.J. van der, *Verdam Middelnederlandsch handwoordenboek: Supplement*. M. Nijhoff, Leiden/Antwerpen 1983.

MNW (1929), Verwijs E., J. Verdam, *Middelnederlandsch woordenboek* I-IX. Voltooid door F.A. Stoet. M. Nijhoff, 's-Gravenhage 1885-1929. Vol. X *Bouwstoffen* (1927-'52), vol. XI *Aanvullingen* [...] (1941).

Navarro, G. (2000), 'Nr-grep: A fast and flexible pattern matching tool'. *Technical Report TR/DCC-2000-3*, Dept. of Computer Science, Univ. of Chile, 2000.

ONW. *Oudnederlands woordenboek*. Instituut voor Nederlandse Lexicologie. To be completed in 2007.

VMNW (2001), *Vroegmiddelnederlands woordenboek*. Woordenboek van het Nederlands van de dertiende eeuw in hoofdzaak op basis van het Corpus-Gysseling. Bewerkt door W.J.J. Pijnenburg, K.H. van Dalen-Oskam, K.A.C. Depuydt, T.H. Schoonheim. Instituut voor Nederlandse Lexicologie, Leiden.

Wall, Larry, Tom Christiansen, Jon Orwant 2000. *Programming Perl*.O'Reilly [3rd ed.].

WNT (1998), *Woordenboek der Nederlandsche taal* I-XXIX. Bewerkt door M. de Vries, L.A. te Winkel e.a. M. Nijhoff enz., 's-Gravenhage enz. 1882-1998.

From paper slips to the electronic archive Cross-linking potential in 90 years of lexicographic work at the *Wörterbuch der bairischen Mundarten in Österreich (WBÖ)*

EVELINE WANDL-VOGT
Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX)
der Österreichischen Akademie der Wissenschaften (ÖAW)
1010 Wien, Postgasse 7/1
eveline.wandl-vogt@oeaw.ac.at

Abstract

In 1911 the *Kommission zur Schaffung des Österreichisch-Bayerischen Wörterbuches und zur Erforschung unserer Mundarten* was founded by the nowaday's *Österreichischen Akademie der Wissenschaften (ÖAW)* in conjunction with the nowaday's *Bayerischen Akademie der Wissenschaften (BAdW)*. After a period of joint work, in 1961 they decided to publish two separate parts of the dictionary: part I dealing with Austria and nearby regions (*WBÖ*) and part II dealing with Bavaria (↓1). The material of the *WBÖ*, for example the about 4 million paper slips of the main archive, the so-called *Hauptkatalog (HK)*, shows cross-linking potential to a very high degree (↓2.1). The lexicographer has to know very much or to use a lot of source material to understand the whole information on a single paper slip.

The entries of the *WBÖ* show cross-linking potential to a very high degree, too (↓2.2). The user of the *WBÖ* can follow the internal links and find the correct information, but to follow external links means a lot of effort.

The *DBÖ*, a new project started in 1993, is a data base system with several data bases (↓3). It combines example data bases with source-material data bases. The main data base is the *Hauptkatalogdatenbank (HkDb)*. The cross-linking potential of the *HK* was considered by the *DBÖ*-development.

The *DBÖ* has not been finished, yet it can be used for writing entries for the *WBÖ* (↓4.1). The work the lexicographers do remains the same, but it is done in another *way* using the new media. The digital texts e.g. are a valuable source for the lexicographer freeing him from the monotony of checking over and over again, thus leaving him more time for his proper work, namely the writing of entries for the *WBÖ*.

Finally, we have to think about the cross-linking potential of the *WBÖ*, because links stay static in a print-publication, a link to a digital source is problematic (e.g. *s.DBÖ*) and every link in a print-publication means effort for the reader (↓4.2). Therefore the *WBÖ* team should develop an editing system implementing possibilities of cross-linking to a very high degree and take advantage of single-sourcing. A first step to this ambitious aim was set in 2004 when the *WBÖ* team developed a MS Word based editor.

1 The *Wörterbuch der bairschen Mundarten in Österreich (WBÖ)* Historical survey

It was in 1911, nearly one hundred years after the first papers on the Bavarian language by Schmeller appeared, when the *Kommission zur Schaffung des Österreichisch-Bayerischen Wörterbuches und zur Erforschung unserer Mundarten* was founded by the nowaday's *Österreichischen Akademie der Wissenschaften (ÖAW)* in conjunction with the nowaday's *Bayerischen Akademie der Wissenschaften (BAdW)*. They decided to start a new project introducing new standards in lexicography, which were to give a complete and detailed overview of the bavarian dialect variants in the Austrian-Hungarian Monarchy and the Kingdom of Bavaria, to present the complete wordiness of these areas with detailed definitions and contextualized examples, to record the authentic pronunciation, to define the grammatical coding to each single word entry, to trace the etymology to each lexicon-item and to register expert knowledge in the fields of rural techniques, traditional folk medicine and customs.

After a time of conception, 1913 the nowaday's *Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX)* was founded and the collection of the material for the dictionary started. From 1913-1932 109 questionnaires and 9 auxiliary questionnaires with approx. 24,000 detailled questions were sent to selected Austrian municipalities.

In 1961 the Austrian and Bavarian Academy decided to publish two separate parts of the lexicon: Part I dealing with Austria (with the exception of the province of Vorarlberg) and the (former) German-speaking parts across the nowaday's borders of Italy, Slovenia, Slovakia, Hungary and the Czech Republic[1], and part II dealing with the Bavarian dialects in Germany. Part I, the *Wörterbuch der bairischen Mundarten in Österreich (WBÖ)*, is published since 1963 (4 volumes, 5 parts: *A – [auf]ge-dunsen*) and is supposed to consist of 12 volumes by the time of its completion in 2020.[2]

2 The dictionary itself: an ideal hypertext? Material and entries

Hypertexts are non-linear, non-sequential texts, consisting of a number of modules which are connected to each other by so-called links.[3] But even traditional linear,

[1] See Straffungskonzept (1998) § 1 in: WBÖ-Beiheft 2, 11 and the enclosed map *Das Bearbeitungsgebiet des Wörterbuchs der bairischen Mundarten in Österreich (WBÖ) und seine Lage im gesamtbairischen Sprachraum.*

[2] Details to the historical survey see WBÖ 1,V- XVI (*Vorwort*) and Bergmann (2003). Part II, the *Bayerisches Wörterbuch (BWB)*, is published since 1995 (see ↓References).

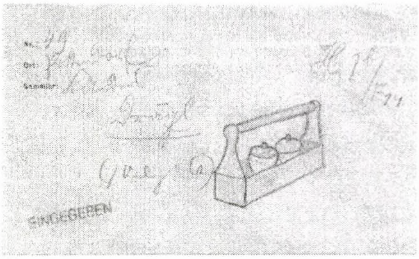
[3] See Nentwich (2003) 263.

academic texts have what one may call hyper-elements.[4] A dictionary like the *WBÖ* uses such elements to a very high degree.[5]

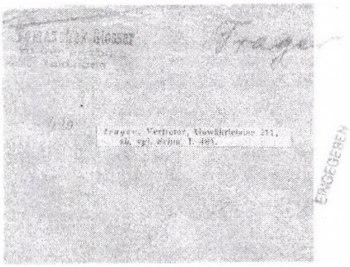
2.1 Material and its cross-linking potential

The main archive, the so-called *Hauptkatalog (HK)*, is a collection of about 4 million paper slips, representing the variety of the regional, social and historical bavarian dialects in Austria and nearby regions.
About 55% of the material are answers to systematic questionnaires (↓map1) and free collections (indirect inquiry methods). About 35% of the material was systematically investigated by experts, who recorded and transcribed the pronunciation (direct inquiry methods). Finally, about 10% of the material are excerpts of historical (↓map2) and dialect texts.

map1: paper slip with drawing


EINGEGEBEN

map2: paper slip of a historical excerpt


EINGEGEBEN

There is lot of cross-linking potential left on every paper slip.
For example ↑map1: The number *49H1b/F14* is a link to the questionnaire (↓3, *Fragebogendatenbank*). The placename *Pottendorf* is a link to a gazetteer and a hierarchically superordinated local unit (↓3, *Gebietsdatenbank*). The surname *Anders* is a link to the collection of coworkers and collectors of the *WBÖ* (↓3, *Mitarbeiterdatenbank*).
The picture itself is an indirect link to the *Bilddatenbank* (↓3).
The shorthand note is an indirect link to the list of *Gabelsberger Steno*.
The *ā* in the word *drāgl* is an indirect link to the collection of coworkers and collectors of the *WBÖ* (↓3, *Mitarbeiterdatenbank*) and to the list of rules for diacritical marks collectors of the *WBÖ* should use.[6]
For example ↑map2: The red word *Trager* is a link to the lemma in the *HK* and the *DBÖ* (↓3). The stamp *Tomaschek Glossar /../* is a literature-link (↓3, *Literaturdatenbank*) and the number *420* is a link to a certain page of a specific

[4] See Nentwich (2003) 258.

[5] Another example is discussed in Kammerer (1998) (*Frühneuhochdeutsches Wörterbuch*).

[6] Belehrung (1913).

literature source. Finally, the excerpt contains a link to two pages of another text (*211, 45;* ↓3, *Textkorpus*) and to another dictionary (*vgl. Schm. 1,481*; ↓3, *Literaturdatenbank*).

All of these links mentioned are external links.
There are internal links on the paper slips, too, e.g. to link synonyms.
Concluding it could be said, that the lexicographer and everyone who uses the material of the *HK* for research needs a lot of specific information or has to follow many links to apprehend the whole information presented on a single paper slip.

2.2 WBÖ-entries and their cross-linking potential

The lemma of the *WBÖ*-entry is elaborated by the lexicographer due to etymological rules and rules of pronunciation.
The catchword entries of the *WBÖ* are sorted according to the "Grundwortprinzip". Compounds are registered in the same entry as the basic word, e.g. the word *Apfelbaum* is alphabetically sorted under the entry *Pāum* as *(Apfel)pāum* with the variant *(Epfel)pāum*. There are links to derivation entries. Furthermore, some derivations are discussed in the entry of the etymologic base word (see ↓map 3: There is a link to the derivation *Trënschel*; but the derivation *Trënschlach* is dealt with in the position of word family of the *WBÖ*-entry *trënscheln*).
The *WBÖ*-entries are of a modular structure. These are the main positions of a standard *WBÖ*-entry[7]: superordinated positions: lemma of the entry, simplex, compound(s), word family, author; subordinated positions: certain lemma, grammar, pronunciation, etymology, sense, links to other specific dictionaries.[8]

One has to differentiate between internal and external, direct an indirect links. There are four types of direct links in *WBÖ*-entries:

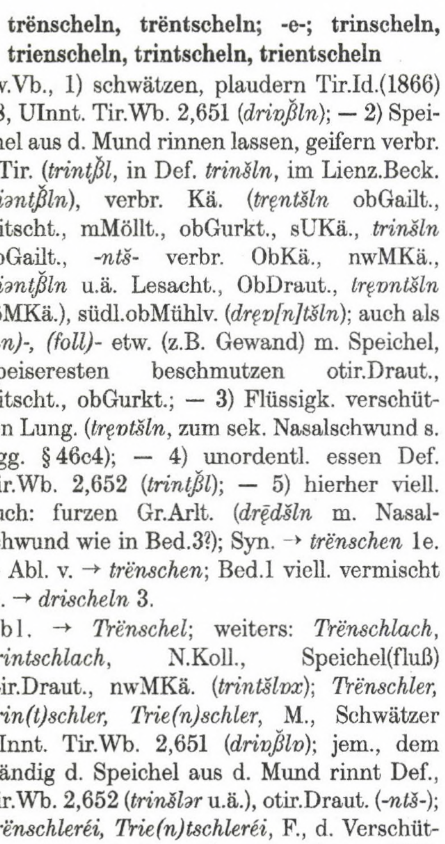
1. internal links to a certain lemma, marked with → *lemma*
   e.g. links to synonyms , see ↓map 3: Syn. → *trënschen* 1e;
   e.g. links to derivations, see ↓map 3: Abl. → *Trënschel*;
2. internal links to a certain sense, unmarked
   see ↓map 3: 5) hierher viell. auch: furzen Gr.Arlt. (/../ wie in Bed.3?)
3. internal links from the compounds to the simplex, marked with → *Simpl.*
   e.g. Komp. (→ Simpl.1a): /../
4. external links to the *DBÖ* (↓3), marked with *(Näh.) s.DBÖ* (↓map5).

[7] There are some types of entries with special linking functions (e.g. *DBÖ*-entries; see Wandl-Vogt [2004]) that have not been taken in account. For the structure of an entry see WBÖ-Beiheft 2, 14-17 (*Artikelaufbau*) and, more detailed, Wandl-Vogt (2005).
[8] These subordinated positions could be split in further, more detailed subordinated positions.

Most of the links in the *WBÖ*-entries are indirect or *potential* links.[9]

It means a lot of effort for the user to follow external links, e.g. to find information about specific abbreviations (general abbreviations, abbreviations of literature sources and local units). It is still more difficult to get information about unpublished sources, e.g. the *DBÖ*, rules of lemmatization and macrostructure.

map3: *WBÖ*-entry[10]



trënscheln, trëntscheln; -e-; trinscheln, trienscheln, trintscheln, trientscheln sw.Vb., 1) schwätzen, plaudern Tir.Id.(1866) 88, UInnt. Tir.Wb. 2,651 (*driɒßln*); — 2) Speichel aus d. Mund rinnen lassen, geifern verbr. OTir. (*trintß̌l*, in Def. *trinšln*, im Lienz.Beck. *triəntß̌ln*), verbr. Kä. (*tręntšln* obGailt., Gitscht., mMöllt., obGurkt., sUKä., *trinšln* obGailt., *-ntš-* verbr. ObKä., nwMKä., *triəntß̌ln* u.ä. Lesacht., ObDraut., *tręɒntšln* söMKä.), südl.obMühlv. (*dręɒ[n]tšln*); auch als *(an)-*, *(foll)-* etw. (z.B. Gewand) m. Speichel, Speiseresten beschmutzen otir.Draut., Gitscht., obGurkt.; — 3) Flüssigk. verschütten Lung. (*tręɒtšln*, zum sek. Nasalschwund s. Lgg. § 46c4); — 4) unordentl. essen Def. Tir.Wb. 2,652 (*trintß̌l*); — 5) hierher viell. auch: furzen Gr.Arlt. (*drę̄dšln* m. Nasalschwund wie in Bed.3?); Syn. → *trënschen* 1e. — Abl. v. → *trënschen*; Bed.1 viell. vermischt m. → *drischeln* 3.

Abl. → *Trënschel*; weiters: *Trënschlach*, *Trintschlach*, N.Koll., Speichel(fluß) otir.Draut., nwMKä. (*trintšlɒx*); *Trënschler*, *Trin(t)schler*, *Trie(n)schler*, M., Schwätzer UInnt. Tir.Wb. 2,651 (*driɒß̌lɒ*); jem., dem ständig d. Speichel aus d. Mund rinnt Def., Tir.Wb. 2,652 (*trinšlər* u.ä.), otir.Draut. (*-ntš-*); *Trënschleréi*, *Trie(n)tschleréi*, F., d. Verschütten v. Flüssigk. Lung. (*tręɒtšlɒráe*). *W.B.*

links

ë is an external link to the rules of lemmatization and *trënscheln* is an external link to the *HK* and the *DBÖ* (↓3).
Tir.Wb. is an external literature-link (↓3, *Literaturdatenbank*), and the number *2,651* is an external link to a specific volume and page.
š is an external link to the list of diacritical marks.
söMKä. is an external link to a certain local unit (↓3, *Gebietsatenbank*).
The - in *(an)-* is an internal link to the lemma of the entry and the whole word is an external link to the lemma in the *HK* and the *DBÖ* (↓3).
→ *trënschen* is an internal *WBÖ*-link (lemma of another *WBÖ*-entry).
Bed.1 is an internal link ot the entry.

N.Koll. is an external link to the list of abbreviations.

W.B. is an external link to the abbreviations and an external link to the collection of co-workers and collectors of the *WBÖ* (↓3, *Mitarbeiterdatenbank*).

[9] To potential links see Blumenthal/Lemmnitzer/Storrer (1988) and Kammerer (1998) 164f.

[10] Another example for a *WBÖ*-entry (text-cut-out) see ↓map5.

3 The *Datenbank der bairischen Mundarten in Österreich (DBÖ)* Structure and cross-linking functionalities

In 1993, in order to support the lexicographer's work decisively, the I DINAMLEX started a new project: the so-called *Datenbank der bairischen Mundarten in Österreich (DBÖ)*, evaluated and financed by the ÖAW.

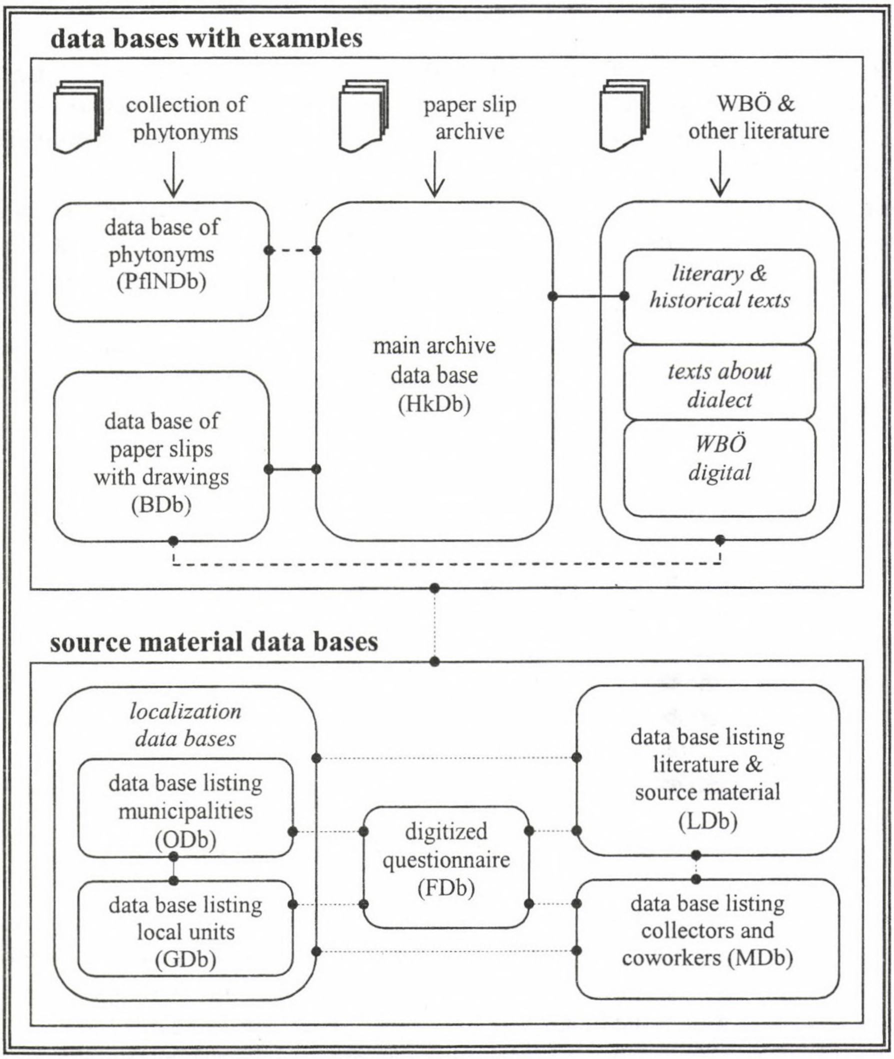
The *DBÖ* is intended to be the electronic archive, storing not only the digitized paper slips but also the whole source materials of the *WBÖ*.

The *DBÖ* format is TUSTEP (*Tübinger System von Textverarbeitungsprogrammen*), except the *Bilddatenbank* (TIFF) and the digitized *WBÖ* (TUSTEP; planned for 2005/2006: TIFF, PDF, XML; see ↓4.2).

The main archive is the *Hauptkatalogdatenbank (HkDb)*. Every paper slip of the *HK* is fed the computer mask by one of the 8 datatypists. Usually there are no picture files, except paper slips with drawings (↑map 1).
A collection of dialect phytonyms is digitized in the *Pflanzennamendatenbank (PflNDb)*.
The *Textkorpus (TK)* is a data base of about 100 Austrian dialectal and historical electronic texts (text-cut-out see ↓map6). One part of it is the digitized *WBÖ*.
Furthermore, there are several data bases storing source material:
In the *Ortsdatenbank (ODb)* every municipality taken into account in the *WBÖ* is listed and assigned to specific superordinated local units (3,212 entries). In the *Gebietsdatenbank (GDb)* all local units and their subordinated local units are noted down (1,492 entries). The *Fragebogendatenbank (FDb)* stores the digitized questionnaire (24,000 entries). The *Mitarbeiterdatenbank (MDb)* lists names and information about collectors and coworkers of the *WBÖ* (3,900 entries).
The informations the collectors got (e.g. Belehrung [1913]) and informations about the rules of lemmatization are to be digitized.[11]
Finally, the literature with abbreviations used for writing the *WBÖ* is registered in the *Literaturdatenbank* (*LDb*; 2,105 entries).
The list of the general abbreviations and the list of diacritical marks are digitized, yet not in a data base format (both DOC, PDF).
Just few shorthand notes of *Gabelsberger Steno* are in the source material of the *WBÖ*; these are not digitized at the time being.

[11] The rules of lemmatization published in WBÖ 1,8-18 should be adapted for the actual situation and published with user-friendly information about the macrostructure of the *WBÖ* (e.g. how the lemmata are sorted in the *WBÖ*).

map 4: Survey: The *DBÖ*


data bases with examples
collection of phytonyms
paper slip archive
WBÖ & other literature
data base of phytonyms (PflNDb)
main archive data base (HkDb)
literary & historical texts
texts about dialect
WBÖ digital
data base of paper slips with drawings (BDb)
source material data bases
localization data bases
data base listing municipalities (ODb)
data base listing local units (GDb)
digitized questionnaire (FDb)
data base listing literature & source material (LDb)
data base listing collectors and coworkers (MDb)

4 Changes in writing – changes in thinking Cross-linking : challenge and chance

4.1 Cross-linking as the new (?) principle of writing WBÖ-entries

As we have seen the *WBÖ*-material (↑2.1) is a collection of paper slips with one specific information and a lot of links. Writing a certain *WBÖ*-entry the lexicographer follows the links; many of them even without realizing the linking he does.

An example is to illustrate the use of the *DBÖ* for writing *WBÖ*-entries (see paper slip *Trager* ↑map.2). The lexicographer knows that the *Tomaschek Glossar* is a certain literature, in which he will get exactly the same information as presented on the paper slip. He knows that the numbers behind the sense are linking to the certain literature, where he finds the whole context. He knows, what literature it is (because it is not mentioned on the paper slip itself) and where he can find it. He furthermore knows, that the catchword as registered on the copy of the paper slip must not be exactly the same to be found in the literature.
According to the guidelines of the Straffungskonzept 1998, § 1.3.1 he knows that he has just to cite *one* example per sense and century and – according to § 2.5. ebd. – he knows, that the has to choose the *best* and *shortest* section.
Finally, he has to

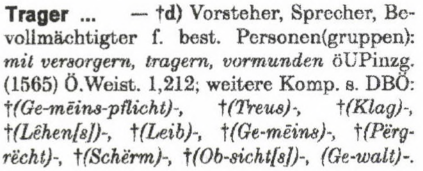
1. locate the text-cut-out
   first: municipality (↑3 *ODb*)
   second: unit, the municipality belogs to (↑3 *GDb*)
   Where does the author come from? What kind of dialect does/did he use?
2. bring the text-cut-out into a chronological order
   When was the text written? Sometimes – e.g. in the case of historical documents such as the example *Trager* (↑map.2) demonstrates – this is a highly sophisticated question.
3. decide, whether to use the text-cut-out or not, and insert it into the *WBÖ*-entry.
4. find the correct abbreviation for the cited literature (↑3 *LDb*).

If the text is part of it, the lexicographer uses the *DBÖ* (↑3) to speed up the working process. Every single file in the *Textkorpus (TK)* has an alphanumeric key, which allows one to retrieve each quotation from the data base into a chronological order and to sort it according to its localization (↓map6): eg. *{1565:U\5.2b} öUPinzg.:UPinzg.:Pinzg.:Sa.:Öst.* means: the example is a legal instrument (*U*) of the year *1565*. It was written in the eastern part oft the *Unterpinzgau* (*öUPinzg.*), which is part of the superordinated local units *UPinzg., Pinzg.,* the federal county *Salzburg* (*Sa.*) and the nation *Österreich* (*Öst.*).
Due to this standardization and user-friendlyness increase. The sidesteps to other sources need no longer be done by the lexicographer himself, because the information of all data bases is linked[12].

[12] Unfortunately, at the time being the linking of the data bases is static, meaning, one change in one data base does not automatically produce changes in all linked data bases.

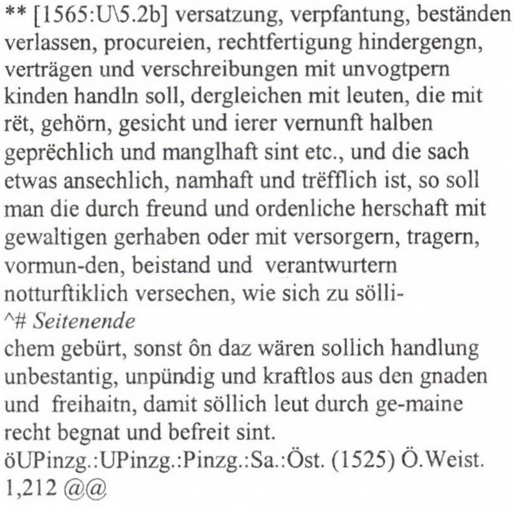
The data base frees the lexicographer from the monotony of checking over and over again, thus leaving him more time for his proper work, namely the writing of the entries for the *WBÖ*.[13]

map5: *WBÖ*-entry (text-cut-out)



Trager ... — **†d)** Vorsteher, Sprecher, Bevollmächtigter f. best. Personen(gruppen): *mit versorgern, tragern, vormunden* öUPinzg. (1565) Ö.Weist. 1,212; weitere Komp. s. DBÖ: †*(Ge-mēins-pflicht)-*, †*(Treus)-*, †*(Klag)-*, †*(Lêhen[s])-*, †*(Leib)-*, †*(Ge-mēins)-*, †*(Përg-rëcht)-*, †*(Schërm)-*, †*(Ob-sicht[s])-*, *(Ge-walt)-*.

map 6: TK (text-cut-out)
corresponding with ↑map 2 and ← map5



** [1565:U\5.2b] versatzung, verpfantung, beständen verlassen, procureien, rechtfertigung hindergengn, verträgen und verschreibungen mit unvogtpern kinden handln soll, dergleichen mit leuten, die mit rët, gehörn, gesicht und ierer vernunft halben geprëchlich und manglhaft sint etc., und die sach etwas ansechlich, namhaft und trëfflich ist, so soll man die durch freund und ordenliche herschaft mit gewaltigen gerhaben oder mit versorgern, tragern, vormun-den, beistand und verantwurtern notturftiklich versechen, wie sich zu sölli-
^# *Seitenende*
chem gebürt, sonst ôn daz wären sollich handlung unbestantig, unpündig und kraftlos aus den gnaden und freihaitn, damit söllich leut durch ge-maine recht begnat und befreit sint.
öUPinzg.:UPinzg.:Pinzg.:Sa.:Öst. (1525) Ö.Weist. 1,212 @@

4.2 From cross-linking potential to cross-media publishing

The cross-linking potential of the material (↑2.1) has been taken into account by the *DBÖ*-development (↑3).

But what about the cross-linking potential of the *WBÖ* itself (↑2.2)? How should the static print publication of the *WBÖ* allow the linking to the *DBÖ* as required in the Straffungskonzept 1998, § 1.2?

This is just one reason why the publication for the *WBÖ* should change. One first step was taken in 2004, when a new MS Word-based editing system was developed. The new principle is to mark what could not be parsed afterwards with data base support, and to tag it according to its function, e.g. mark differently the lemma of an *WBÖ*-entry and the lemma of a compound (both of them are printed bold).

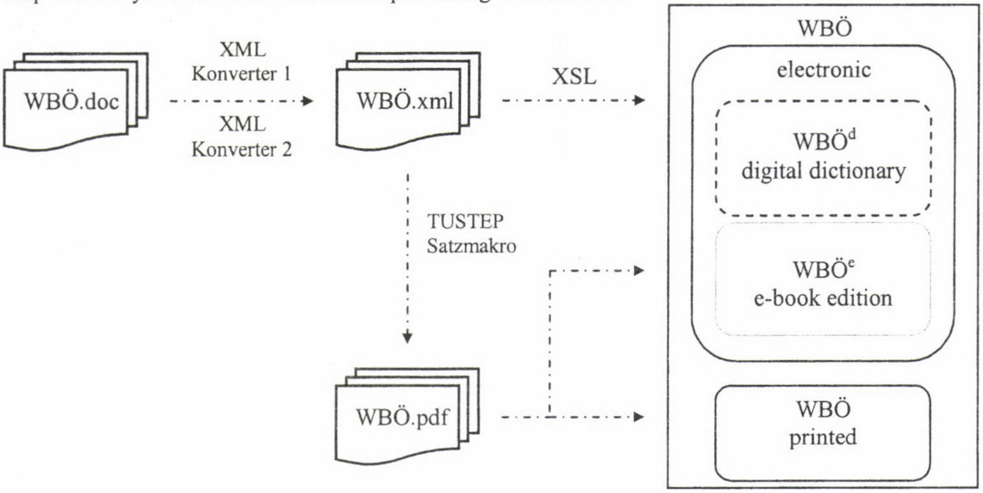
In the end the Word-file is converted into a xml-file, which should be the base of several publishing formats, such as

1. the print publication

[13] More about the *TK* and the lexicographers work with it see Wandl-Vogt (2003).

2. the e-book edition (planned for 2005/2006)
3. a digital dictionary, in which the *WBÖ* on the one hand, and the *DBÖ* on the other are *one* modular complex, where the user could use the sources as well as the results for his research.

map 7: Survey: The idea of cross-media publishing with the *WBÖ*



5 Conclusion Old wine in new hoses?

Actually, we do nothing new when linking. It is just the *way* we do it that is new. We should enable the user to take a look behind the “high walls of wisdom and knowledge”. We should encourage him to research himself with material that is interpreted in the *WBÖ*-entry and not just to use the results. This means transparency (and somehow “control”) and democracy. We should open the doors to let new ideas get in.

We are searching for the best ways to write, publish and read the *WBÖ* and to use the *DBÖ*. Yet, we have to realize that the ways are not always easily to recognize and that they are to be gone with much more effort and difficulty.

6 References

Belehrung (1913)
Belehrung für die Sammler des bayrisch-österreichischen Wortschatzes. Wien 1913.

Bergmann (2003)
Hubert Bergmann: Streiflichter aus der Geschichte des Instituts für Österreichische Dialekt- und Namenlexika. PPt. – online: http://www.oeaw.ac.at/dinamlex/power_point_tagg_copy.ppt (18.04.2005).

Blumenthal/Lemmnitzer/Storrer (1988)
Anreas Blumenthal, Lothar Lemmnitzer, Angelika Storrer: Was ist eigentlich ein Verweis? Konzeptionelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung. In: Gisela Harras (Hg.): Das Wörterbuch. Artikel und Verweisstrukturen. Jahrbuch 1987 des Instituts für deutsche Sprache. Bielefeld 1988, 351-373 (Sprache der Gegenwart LXXIV).
BWB
Bayerisches Wörterbuch (BWB). Hg. von der Kommission für Mundartforschung der Bayerischen Akademie der Wissenschaften. München 1995-lfd. (Bayerisch-österreichisches Wörterbuch: II. Bayern).

Kammerer (1998)
Matthias Kammerer: Hypertextualisierung gedruckter Wörterbuchtexte: Verweisstrukturen und Hyperlinks. Eine Analyse anhand des Frühneuhochdeutschen Wörterbuches. In: Angelika Storrer, Bettina Harriehausen (Hg.): Hypermedia für Lexikon und Grammatik. Tübingen 1998, 145-171 (Studien zur deutschen Sprache 12).

Nentwich (2003)
Michael Nentwich: cyberscience. Research in the Age of the internet. Wien 2003.

Straffungskonzept (1998)
Neues Straffungskonzept für das „Wörterbuch der bairischen Mundarten in Österreich (WBÖ)“. Masch.schriftl. Wien. – published in: WBÖ-Beiheft 2,11-13; – online: http://www.oeaw.ac.at/dinamlex/Straffungskonzept_1998.pdf (18.04.2005).

TUSTEP
Tübinger System von Textverarbeitungsprogrammen (TUSTEP). – information: online: http://www.uni-tuebingen.de/zdv/tustep/index.html (18.04.2005).

WBÖ
Wörterbuch der bairischen Mundarten in Österreich. Wien 1970-lfd. (Bayerisch-österreichisches Wörterbuch: I. Österreich).

WBÖ-Beiheft 2
Institut für Österreichische Dialekt- und Namenlexika (Hg.): Wörterbuch der bairischen Mundarten in Österreich. Beiheft Nr. 2. Erläuterungen zum Wörterbuch. Wien 2005 (Bayerisch-österreichisches Wörterbuch: I. Österreich).

Wandl-Vogt (2003)
Eveline Wandl-Vogt: Digitale Volltexte als Arbeitsbehelf für die Dialektlexikographie am Beispiel des Textkorpus zum „Wörterbuch der bairischen Mundarten in Österreich (WBÖ)“. In: Thomas Burch, Johannes Fournier, Kurt Gärtner, Andrea Rapp (Hg.): Standards und Methoden der Volltextdigitalisierung. Stuttgart 2003, 177-185, 340f. (Akademie der Wissenschaften und der Literatur. Abhandlungen der Geistes- und sozialwissenschaftlichen Klasse. Einzelveröffentlichung 9).

Wandl-Vogt (2004)
–: Verweisstrukturen in einem datenbankunterstützten Dialektwörterbuch am Beispiel des Wörterbuchs der bairischen Mundarten in Österreich (WBÖ). In: Stephan Gaisbauer, Hermann Scheuringer (Hg.): Linzerschnitten. Bayerisch-österreichische Dialektologentagung 2001, zugleich 3. Arbeitstagung zu Sprache und Dialekt in Oberösterreich, in Linz, September 2001. Linz 2004, 423-435 (Schriften zur Literatur und Sprache in Oberösterreich 8).

Wandl-Vogt (2005)
–: Überlegungen zur Artikelstruktur im Wörterbuch der bairischen Mundarten in Österreich (WBÖ) unter Berücksichtigung des Neuen Straffungskonzeptes von 1998. Dargestellt an Hand ausgewählter Wörterbuchartikel von Günter Lipold. In: Christiane Pabst (Hg.): Sprache als System und Prozess. Festschrift für Günter Lipold zum 60. Geburtstag. Wien. (to be printed).

The *True, Deep Happiness*: Towards the Automatic Semantic Classification of Adjective-Noun Collocations

LEO WANNER
ICREA and Pompeu Fabra University
Passeig de Circumval.lació, 8
08003 Barcelona
Spain
leo.wanner@icrea.es

BERND BOHNET
University of Stuttgart
Universitätsstr. 38
70569 Stuttgart
Germany
bernd.bohnet@iis.uni-stuttgart.de

MARGARITA ALONSO, NANCY VÁZQUEZ
Faculty of Philology
University of La Coruña
Campus de Zapateira
15071 La Coruña
lxalonso@udc.es, lxnveiga@udc.es

Most of the current techniques for the extraction of collocations from corpora either provide plain lists of collocations or focus on the retrieval of a few specific types of verb-noun collocations – support verb constructions being the most prominent of them. Only little work has been done so far on the extraction and semantic classification of adjective-noun collocations. In this paper, we concentrate on the classification of adjective-noun collocations according to an adjective-noun fragment of the fine-grained semantically oriented collocation typology – the *lexical function* typology as introduced in the *Explanatory Combinatorial Lexicology*. For the classification, we use different Machine Learning (ML) techniques that draw upon the semantic description of the adjective-noun bigrams. The descriptions are obtained from an external lexico-semantic resource. So far, experiments have been carried out mainly on Spanish. In experiments discussed in the paper, two different ML-techniques have been applied to bigrams from the semantic field of emotions: the Nearest Neighbor Classification and a variant of the Bayesian Network Classification – the Tree Augmented Network Classification. As the external lexico-semantic resource, the Spanish part of the EuroWordNet has been used.

1. Introduction

Most of the current techniques developed for the extraction of collocations from the corpus return plain lists of word combinations estimated to be collocations. However, while being useful as a resource for manual dictionary construction, plain lists of collocations are only of restricted use in Natural Language Processing (NLP) and for second language learning. In order to be also useful in these two areas, collocations must be assigned a semantic description. If not done during the extraction stage, this is done, as a rule manually, in a subsequent stage (cf., e.g., Smadja & McKeown, 1991). True, recently, several techniques have been proposed to deal with collocation semantics. However, these techniques focus, as a rule, on one specific type of collocations (such as *support verb constructions*; cf., among others, Grefenstette & Teufel, 1995; Tapanainen *et.al.*, 1998; Stevenson *et al.* 2004).

The goal of our research in Computational Lexicography is twofold: (a) to develop algorithms that substitute the traditionally manual stage of assigning semantics to collocations, i.e., algorithms that classify collocations obtained from elsewhere (either by a collocation extraction program or from a traditional collocation dictionary) according to a fine-grained rich semantically-motivated collocation typology; (b) to develop algorithms that identify collocations in the corpus and classify them according to the given collocation typology. Obviously, the algorithms in (b) are designed as extensions of the algorithms in (a). Our language of investigation has been so far mainly Spanish. In our previous work, we focused on noun-verb collocations; cf. (Wanner, 2004, Wannner et al. submitted) for results on (a) and (Wanner et al., 2005) for results on (b). In this paper, we address the problem of the semantic classification of adjective-noun collocations. The collocation typology that we draw upon for classification is the typology of the *syntagmatic lexical functions* (LFs) as known from the *Explanatory Combinatorial Lexicology*, ECL (Mel'cuk *et al.* 1995; Mel'cuk, 1996). The syntagmatic LF typology is the most detailed collocation typology available to date.

2. Syntagmatic Adjectival Lexical Functions as Collocation Typology

A syntagmatic LF encodes a standard abstract lexico-semantic relation between two lexical units among which one of the units (the *base*) controls the lexical choice of the other unit (the *collocate*). "Standard" means that this relation is sufficiently common; "abstract" means that this relation is sufficiently generic to group all relations that possess the same semantic nucleus. Typical adjective-noun relations that fulfil these criteria are 'intense', 'not intense' 'appropriate', 'inappropriate', 'positive', 'negative', etc. For instance, the LF with the meaning 'intense' captures the relation between *smoker* (the base) and *chain* (the collocate): *chain smoker*; the LF with the meaning 'positive' captures the relation between *performance* (the base) and *brilliant* (the collocate): *brilliant performance*; the LF with the meaning 'negative' captures the relation between *performance* and *poor*: *poor performance*; and so on.

As LF names, Latin abbreviations are used: Magn, Bon, AntiBon, etc. In the ECL-terminology, *brilliant performance* is said "to be an instance of the LF Bon, *poor* *performance* "an instance of the LF AntiBon", etc. In total, about 15 adjective noun LFs are available.[1] Table 1 summarizes the six most common adjective-noun LFs.

Table 1: Some adjective-noun lexical functions (with English examples)

'intense', 'big'(**Magn**)		'not intense', 'little' (**AntiMagn**)	
sharp	FLUCTUATION	*Minor*	FLUCTUATION
high	HOPE	*Faint*	HOPE
big	CROWD	*Slight*	MISUNDERSTANDING
'appropriate' (**Ver**)		'inappropriate' (**AntiVer**)	
appropriate	USE	*suppressed*	EXCITEMENT
absolute	HONESTY	*muffled*	EXCLAMATION
lasting	FRIENDSHIP	*cursory*	INSPECTION
'positive' (**Bon**)		'negative' (**AntiBon**)	
good	USE	*Bad*	USE
bold	HYPTHESIS	*Cheap*	OPTION
perfect	OPPORTUNITY	*Poor*	PERFORMANCE

3. The Approach

Our objective is to classify adjective-noun combinations from a given list according to an LF-typology that consists of the adjective-noun LFs – including the six LFs from Table 1. For the classification task, we use machine learning techniques that exploit the semantic descriptions of the lexical items which co-occur in a combination. Semantic descriptions may be of a varying detail. Thus, some may contain different semantic features, others may be rather superficial – consisting, e.g., of the synonyms of the item in question. In our experiments, we use the Spanish part of the lexical database *EuroWordNet* (Vossen, 1998), henceforth, SpEWN, as the external source of the semantic descriptions of lexical items.

3.1 Basic Assumptions

To be able to classify a candidate bigram with respect to the LF-typology, the characteristic features shared by the instances of each LF **L** in this typology must be known. In corpus-based NLP, characteristic features of a word pattern are most often captured in terms of word frequency counts (i.e., how often the words in the

[1] Contrary to the tradition in ECL, we count AntiMagn, AntiVer, etc. as "simple" LFs (rather than as being composed of Anti and Magn, Anti and Ver, and so on). This is to account for their prominence as well as for the rare appearance of the LF Anti as a "stand-alone" LF.

pattern in question occur together). In contrast, we suggest to learn specific characteristics of **L** from the *semantics* of the instances of **L**, i.e., using *semantic component* (or *concept*) counts. More precisely, we assume that:

1. The meaning of any lexeme (be it an element of a collocation or an element of a free word combination) is decomposable. This means that for a given collocation, the meaning of the base can be viewed as consisting of a set of semantic components $\{b_1,b_2,\ldots,b_n\}$ and the meaning of the collocate as consisting of a set of components $\{c_1,c_2,\ldots,c_m\}$. See, e.g., (Wierzbicka, 1982; Dixon, 1991) for a cognitively-motivated argumentation that supports this view. The componential description of lexical meanings is expected to be available from an external lexico-semantic resource. Any sufficiently comprehensive and sufficiently formalized lexico-semantic resource can be used.
2. Despite the partial idiosyncrasy of collocations, in a given semantic field, a correlation holds between the semantics of a base and the semantics of collocates this base co-occurs with; see (Mel'cuk & Wanner, 1996) for an empirical study on this topic. This means that (a) if several lexemes occur as bases in collocations expressed by the same LF, these lexemes share one or several meaning components, and (b) if a given collocation is expressed by a specific LF, meaning component pairs of the form (b_i,c_j) can be identified that are characteristic of this LF.
3. Starting from a representative set of manually compiled, semantically decomposed disambiguated instances for the LF **L** (a training set for **L**), we can learn what it means for a word combination to be an instance of **L**. "Disambiguated" means in this context that if the collocate or the base of a given instance in the training set are polysemous, only the decomposition of the sense which comes to bear in this instance is considered.

An approach that is based on the above three assumptions has two major advantages. Firstly, it is not bound to the occurrence frequency of a candidate bigram in the corpus. This is crucial because the frequency criterion is a serious obstacle for the identification of less common collocations. Secondly, it naturally generalizes over collocates with the same meaning. The concept count allows us, for instance, to detect the close semantic similarity between Sp. *profundo* 'deep' and Sp. *extraordinario* 'extraordinary' in co-occurrence with Sp. *admiración* 'admiration' and between Sp. *franco* 'frank' and Sp. *puro* 'pure' in co-occurrence with Sp. *hostilidad* 'hostility'.

3.2 Learning the LFs

According to assumption 2 from above, the profile of an LF **L** can be "learned" in three different ways: (a) by simply assuming that the decomposition $\{b_1,b_2,\ldots,$

$b_n\}\oplus\{c_1,c_2,\ldots,c_m\}$ of each instance within the training set of **L** reveals a representative semantic pattern of **L**; (b) by estimating the relevance of each meaning component pair that occurs within the decomposed description of any instance of the training set of **L** for the characteristics of **L** – drawing then upon the most relevant pairs; (c) by deriving a "centroid" from the descriptions of the instances of the training set, i.e., compiling an artificial "ideal" semantic representative of **L** whose description contains the most prominent meaning components encountered in the instances of the training set of **L**.

In this paper, we focus on (a) and (b), using for each a different machine learning technique. For (a), we use the *Nearest Neighbor* (NN) *Classification*, and for (b) we use a variant of the *Bayesian Network Classification* – the *Tree Augmented Bayesian Network* (TAN) *Classification*. For work on (c), see (Wanner, 2004).

3.2.1 Nearest Neighbor Classification

Unlike the other ML-techniques, NN-Classification does not include, strictly speaking, a learning stage. It can be thought of as consisting of a representation stage and a classification stage. In abstract terms, the representation can be described in terms of a pair of vector space models (Salton, 1980). Assume a training set of instances for each LF $\mathbf{L_1}$, $\mathbf{L_2}$, ..., $\mathbf{L_n}$ in the LF-typology. As mentioned above, the meaning of each instance is considered being composed of a set of base meaning components $\{b_1,b_2,\ldots,b_n\}$ and a set of collocate meaning components $\{c_1,c_2,\ldots,c_m\}$. Accordingly, the set of all distinct base meaning components that occur in the meaning description of any of the instances in any of the training sets is a union over the individual instance base sets: $\{b_1,b_2,\ldots,b_n,\ldots b_N\}$, and the set of all distinct collocate meaning components is a union over the individual instance collocate sets: $\{c_1,c_2,\ldots,c_m,\ldots,c_M\}$. The two sets can be ordered to component *vectors*: $V_B = (b_1,b_2,\ldots,b_n,\ldots b_N)$, $V_C = (c_1,c_2,\ldots,c_m,\ldots,c_M)$. The representation of a specific training instance I is then given by two sequences, I_B and I_C, of '1's and '0's, with I_B being of the same length as V_B, and I_C of the same length as V_C. At the position of a component b_i (c_j) in V_B /V_C, which is available in the meaning description of this instance, I_B /I_C contains a '1'; at the position of a component b_k (c_l), which is not available in the instance's description, I_B /I_C contains a '0'.

In the classification stage, when a candidate adjective-noun bigram $K = (A,N)$ is to be assigned an LF-label, i.e., classified according to the LF-typology, the classification stage consists of (1) the decomposition of the meaning of A and N by looking up the meaning component descriptions for A and N in an external semantic resource; (2) representation of the meaning descriptions of A and N in terms of '1/0'-sequences K_A and K_N as outlined above for the training instances (K_N is constructed with V_B, and K_A with V_C); (3) comparison of K_N and K_A with

I_B /I_C of all instances; the candidate bigram is assigned the LF-label of the instance whose I_B /I_C are most similar to K_N/K_A.

To determine the similarity between (K_N , K_A) and the different (I_B,I_C)s, we use a metric that calculates to what extent the '1/0'-sequences of K_N and I_B and of K_A and I_C coincide:

$$sim(K,I) = \beta \frac{fb}{fb_{max}} + \gamma \frac{fc}{fc_{max}}$$

where *fb* is the number of dimensions (i.e. '1's) shared by the noun in *K* and the base vector of *I*, fb_{max} is the maximal number of dimensions shared by the noun of *K* and a base vector of any of the instances in the training set for the LF **L** of which *I* is an instance, *fc* is the number of dimensions shared by the adjective in *K* and *I*'s collocate vector, and fc_{max} is the maximal number of dimensions shared by the adjective of *K* and a collocate vector of any of the instances in the training set of **L**.

Since the meanings of *N* and *A* are not disambiguated, we have to consider for each *(A,N)* a variety of (K_A, K_N)s, namely the cross-product of all possible senses of *N* and all possible senses of *A*. Therefore, the distance of all (K_A, K_N) sense bigrams of a given candidate bigram to LF-instance vectors is examined; the LF-label that is most often encountered among the closest neighbors of the sense bigrams provides the LF-label of the candidate bigram.

3.2.2 Tree Augmented Bayesian Network Classification

Bayesian networks are a very popular representation for machine learning techniques in corpus-based Computational Linguistics.[2] In our application, a *Bayesian* network is a network (formally: a *labeled directed acyclic graph*) in which one node is assigned an LF-label and all other nodes are assigned semantic component names. For each LF **L** in the typology, an own network is built up. An LF-network contains as many semantic component nodes as are available in the meaning descriptions of the training instances for the LF in question. An edge between two nodes symbolizes the dependency between these two nodes. That is, an edge between a node with the LF-label and a meaning component node means that the corresponding meaning component occurs in the meaning description of some training instances for this LF. The edge is labeled with the probability that this happens given the set of current training instances for this LF. Analogously, an edge between two meaning component nodes means that these two meaning components occur together in the description of some instances. Again, the edge is assigned the probability of this co-occurrence.

[2] See, e.g., (Heckerman, 1996) for an introduction to Bayesian networks and to learning with Bayesian networks.

The different realizations of the Bayesian networks vary with respect to the number and type of edges they introduce. The most wide-spread classification algorithm using Bayesian networks is the so-called *naïve Bayesian classifier*. The naïve Bayesian classifier assumes that the attribute variables are conditionally independent given the LF-label and introduces thus edges only between the LF-label and meaning component nodes, not between meaning component nodes. The independence assumption let it perform poorly in applications where attribute variables depend on each other – as in the case of LF-based collocation classification. Experiments described in (Wanner et al., submitted) buttress this assumption.

Friedman et al. (1997) proposed a classifier network (the *TAN-classifier network*) whose structure is based on the structure of naïve Bayes, i.e., that requires that the LF-label node be parent of every meaning component node, but which captures correlations between meaning components by additional edges between component nodes. Once TAN-networks are built up for each LF in the typology, the classification of lexeme bigrams can take place. The classification is rather straightforward: Given a candidate lexeme bigram (*A*,*N*) whose elements are decomposed (as for NN-classification, see the representation stage in Subsection 3.2.1), the joint probability over the meaning components of *A* and *N* is calculated with each LF-network. The joint probability is calculated by multiplying the probabilities assigned in the network in question to the edges between any of two meaning component nodes that are present in the description of *A* or *N* (see Wanner et al., submitted for formal details). The LF whose network leads to the highest joint probability is chosen as the LF-label for (*A*,*N*).

4. EuroWordNet

In order to be able to apply the above techniques in practice, we need, in the first place, an external lexico-semantic resource that provides us with the componential descriptions of lexemes. As mentioned above, we use the Spanish part of the EuroWordNet, SpEWN, for both the componential description of LF-instances in the training sets and the description of the candidate bigrams. More precisely, we use the *hyperonymy hierarchies* provided for nouns and *synonym sets* provided for adjectives. This is because for adjectives, SpEWN does not contain yet the full scale of information foreseen for the description of lexical items.

EuroWordNet (EWN), of which SpEWN is a part, is a multilingual lexical database which comprises lexico-semantic information organized following the relational paradigm. In contrast to the original Princeton WordNet (Miller, 1990; Fellbaum, 1998), where the hyperonymy hierarchy of a lexical item is purely lexical (i.e. contains only hyperonyms), in SpanWN (as in most WNs in the EuroWN), the hyperonym hierarchy of each lexical item consists of:

- its hyperonyms and synonyms (i.e., words that combine with the lexical item in question to form a (*synset*)
- its own *Base Concepts* (BCs) and the BCs of its hyperonyms
- the *Top Concepts* (TCs) of its BCs and the TCs of its hyperonyms

BCs are general semantic labels that subsume a sufficiently large number of synsets. Examples of such labels are: change, feeling, motion, and possession. Thus, DECLARACIÓN**3** 'declaration' is specified as communication, MIEDO**1** 'fear' as feeling, PRESTAR**3** 'lend' as possession, and so on.[3] Unlike *unique beginners* in the original WN, BCs are mostly not "primitive semantic components" (Miller, 1998); rather, they can be considered labels of semantic fields. The set of BCs used across different WNs in the EuroWN consists of 1310 different tokens. The language-specific synsets of these tokens constitute the cores of the individual WNs in EuroWN.

Each BC is described in terms of TCs – language-independent features such as Agentive, Dynamic, Existence, Mental, Location, Social, etc. (in total, 63 different TCs are distinguished). For instance, the BC change is described by the TCs Dynamic, Location, and Existence.

Cf., Figure 1, which shows the hyperonym hierarchy (including synonyms, BCs and TCs) of ADMIRACIÓN**3** 'admiration' and PROFUNDO**6** 'deep from the collocation *admiración profunda* lit. 'deep admiration).

((4. feeling ADMIRACIÓN**3**

3. feeling AFICIÓN**2** GUSTO**5**

2. Tops Dynamic | Experience | Mental SENTIMIENTO**1**

1. Tops Mental | Property RASGO-PSICOLÓGICO**1**

(1. PROFUNDO**6** EXTREMO**6** SEVERO**13**))

Figure 1: Hyperonym hierarchies for ADMIRACIÓN**3** and PROFUNDO**6** in the collocation *admiración profunda* (lexical items are written in small capitals, BCs and TCs are in sans serif, and the TCs start with a capital; individual TCs are separated by the '|' sign)

5. Experiments

Intuitively, collocations that belong to a single semantic field are semantically more homogeneous than collocations that belong to different fields. Therefore, their classification can also be expected to be of higher quality. To verify this

[3] The numbers indicate the corresponding senses in SpanWN.

hypothesis, it is useful to carry out experiments on both individual semantic fields and across semantic fields.

So far, we carried out experiments on emotion adjective-noun collocations with the following four LFs: AntiMagn, AntiVer, Magn, and Ver. Experiments that cover a broader field, namely that of jurisdiction, are underway.

The lists of collocations used in the experiments on emotion adjective-noun collocations stem from the *Diccionario de Colocaciones del Español*, DiCE (Alonso Ramos, 2003). In total, we used 39 AntiMagn-instances, 56 AntiVer-instances, 116 Magn-instances and 86 Ver-instances. 95% of the instances of each LF have been used as training material, the remaining 5% have been used as test material.[4] In order to avoid the distortion of the resulting quality figures by a biased training set, we performed 500 runs, each time selecting randomly 95% of the instances of an LF as training set. The final figures are the mean of the precision and recall over the 500 runs.

Table 2 summarizes the results of the performance of the NN-classifier and the TAN-classifier in our experiments, contrasting them with a baseline. As baseline, we used the co-occurrence of a noun with the most common collocate of the respective LF in the emotion field: *ligero* 'light' for AntiMagn, *fingido* 'faked' for AntiVer, *grande* 'big' for Magn, and *sincero* 'sincere' for Ver. As usual, precision p is given by $p = |A_{LF}|/|A_{CL}|$ (with A_{LF} being the instances of the LF **L** correctly classified as instances of **L** and A_{CL} being all instances classified as instances of **L**), and recall r is given by $r = |A_{LF}|/|A^{T}_{LF}$ (with A^{T}_{LF} being all instances of **L** available in the test material).

Table 2: Results of the NN- and TAN-classification experiments

		AntiMagn	AntiVer	Magn	Ver
NN	*p*	0.61	0.68	0.70	0.70
	r	0.40	0.76	0.72	0.76
TAN	*p*	0.44	0.40	0.91	0.90
	r	0.57	0.84	0.82	0.57
baseline	*p*	1.00	1.00	1.00	1.00
	r	0.10	0.16	0.047	0.17

The precision figures obtained with TAN show that in the case of Magn and Ver, the presence of specific pairs of semantic components in the semantic decomposition of a given adjective-noun bigram is a very good indicator that this

[4] To be more precise, we experimented with 5%, 10%, 25%, 50%, 75%, and 95% of the material as training set (and the remaining percentage as test set). Due to the lack of space, we do not discuss the performance of the ML-techniques with smaller training sets.

bigram is an instance of Magn / Ver. Over 80% of all Magn-instances possess these pairs – unlike Ver-instances, of which more than 40% do not possess them.

The instances of AntiMagn and AntiVer are semantically more heterogeneous and do not reveal typical semantic component pairs with such a prominence as the instances of Magn and Ver do. Therefore, in the TAN-experiments especially *p* decreases considerably. For AntiVer, NN-classification appears more appropriate than TAN: *p* and *r* obtained with NN for this LF suggest that AntiVer-instances cluster reasonably well into several semantically similar groups.

The baseline shows a major discrepancy between *p* and *r*. The precision is 1.0 since in our experiments, the test list of bigrams contains only collocations. It would without any doubt dramatically decrease if the list would contain free bigrams (such as [*un*] *edificio grande* lit. '[a] big building', *peso ligero* 'light weight', [*un*] *hombre sincero* '[a] sincere man', etc.). Furthermore, in the field of emotions, the most common collocate of an instance of a given LF is unlikely to appear as collocate of a different LF. The very low recall for all LFs is due to the fact that, obviously, more than one collocate forms instances of a given LF with the emotion nouns.

6. Conclusions and Future Work

Most of the work on identification and classification of collocations described in the literature has been carried out for verb-noun collocations. Our experiments show that an automatic classification according to a fine-grained typology is equally possible for adjective-noun collocations.

So far, we focused on Spanish material, using SpEWN as an external lexico-semantic resource. Three strands of future work will be pursued: (1) exploring strategies for easing the need for external resources; (2) applying NN, TAN and other machine learning techniques to the identification and classification of collocations in language corpora different from Spanish; (3) extending the experiments to cross-field adjective-noun combinations. Experiments on German that address (1) and (2) and on Spanish that address (3) are currently being defined and some of them are already being carried out.

References

Alonso Ramos, M. 2003. "Hacia un diccionario de colocaciones del español y su codificación." In Fernández, A. et al. (eds.) *Lexicografía computacional y semántica*. Barcelona: University of Barcelona Press.

Dixon, R. 1991. *A New Approach to English Grammar, On Semantic Principles.* Oxford: Clarendon Paperbacks.

Fellbaum, Ch. (ed.). 1998. *WordNet. An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Friedman, N., D. Geiger & M. Goldszmidt. 1997. "Bayesian network classifiers." *Machine Learning*. Vol. 29.2–3:131–163.

Grefenstette, G. & S. Teufel. 1995. "Corpus-based method for automatic identification of support verbs for nominalizations." In *Proceedings of the Biannual Meeting of the European Chapter of the Association for Computational Linguistics*. 27–31.

Heckerman, D. 1996. *A Tutorial on Learning with Bayesian Networks*. Report MSR-TR-95-06. Redmond, WA: Microsoft Advanced Technology Division.

Me'cuk, I. A., A. Clas & A. Polguére.1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Mel'cuk, I.A.1996. "Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon." In Wanner, L. (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. 37–102. Amsterdam: Benjamins Academic Publishers.

Mel'cuk, I.A. & L. Wanner. 1996. Lexical Functions and Lexical Inheritance for Emotion Lexemes in German. In Wanner, L. (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. 209– 278. Amsterdam: Benjamins Academic Publishers.

Miller, G.A. (ed.).1990. "WordNet: An on-line lexical database." *International Journal of Lexicography*. Vol. 3:4.

Smadja, F. and K. McKeown. 1991. Using Collocations for Language Generation. Computational Intelligence 7(4), 229-239.

Salton, G. 1980. "Automatic term class construction using relevance: A summary of work in automatic pseudo-classification". *Information Processing and Management*. Vol. 16.1:1–15.

Stevenson, S. et al. 2004. Statistical Measures of the Semi-Productivity of Light Verb Constructions. In *ACL 2004 Workshop on Multi Word Expressions: Integrating Processing*.

Tapanainen, P., J. Piitulainen & T. Järvinen. 1998. "Idiomatic object usage and support verbs." In *Proceedings of the COLING/ACL*. 1289–1293. Montréal.

Vossen, P. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Wanner, L. 2004. "Towards automatic fine-grained semantic classification of verb-noun collocations." *Natural Language Engineering Journal* 10(2), 95-143.

Wanner, L., B. Bohnet, M. Giereth and V. Vidal. 2005. "The first steps towards the automatic compilation of specialized collocation dictionaries." *Terminology* 11(1), 137-174.

Wanner, L., B. Bohnet, M. Giereth, M. Alonso & A. Martí. submitted. "Making Sense of Collocations."

Wierzbicka, A. 1982. "Why Can You Have a Drink when You Can't Have an Eat?". *Language* 58, 753-789.