Coling 2008

22nd International Conference on Computational Linguistics

Proceedings of the workshop on Cognitive Aspects of the Lexicon

Workshop chairs:
Michael ZOCK and Chu-Ren HUANG

24 August 2008
Manchester, UK

Order copies of this and other Coling proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-905593-56-9

Design by Chimney Design, Brighton, UK
Production and manufacture by One Digital, Brighton, UK

Preface

Information access and exchange play a major role in our globalized world. Hence, building resources (lexica, thesauri, ontologies or annotated corpora) and providing access to words become an important goal. The lexicon is a vital resource for building applications. It is also a crucial element in the study of human language processing.

The spirit of this workshop multidisciplinary, the goal being to gather experts with various backgrounds and to allow them to exchange ideas, to compare their methodologies and theoretical perspectives, to create synergy, and to encourage future collaborations. In sum, the participants will be discussing questions concerning the **cognitive aspects** of the lexicon, and their answers should guide the design of on-line dictionaries.

While completeness is a virtue, the quality of a dictionary depends not only on coverage (number of entries) and granularity, but also on accessibility of information. Access strategies vary with the task (text understanding vs. text production) and the knowledge available at the moment of consultation (word, concept, sound). Unlike *readers*, who look for meanings, *writers* start from them, searching for the 'right' words. While paper dictionaries are static, permitting only limited strategies for accessing information, their electronic counterparts promise dynamic, proactive search via multiple criteria (meaning, sound, related word) and via diverse access routes. Navigation takes place in a huge conceptual-lexical space, and the results are displayable in a multitude of forms (as trees, as lists, as graphs, or sorted alphabetically, by topic, by frequency).

Many lexicographers work nowadays with huge digital corpora, using language technology to build and to maintain the resource. But access to the potential wealth in dictionaries remains limited for the common user. Yet, the new possibilities of electronic media in terms of comfort, speed and flexibility (multiple inputs, polymorph outputs) are enormous and probably beyond our imagination. More than just allowing electronic versions of paper-bound dictionaries, computers provide a freedom for rethinking dictionaries, thesauri, encyclopedia, etc., a distinction necessary in the past for economical reasons, but not justified anymore.

The goal of this workshop is to perform the groundwork for the next generation of electronic dictionaries, that is, to study the possibility of integrating the different resources, as well as to explore the feasibility of taking the users' needs, knowledge and access strategies into account.

To reach this goal we have asked authors to address one or more of the following:

1. **Conceptual input of a dictionary user:** what is present in speaker's/writer's minds when they are generating a message and looking for a (target) word? Does the user have in mind conceptual primitives, semantically related words, some type of partial definition, something like synsets, or something completely different?

2. **Access, navigation and search strategies:** how can search be supported by taking into account prior, i.e. available knowledge? Entries should be accessible in many ways: by word forms, by meaning, by sounds (syllables), or in a combined form, and this even if input is given in an incomplete, imprecise or degraded form. The more precise the conceptual input, the less

navigation should be needed and vice versa. How can we create manageable search spaces, and provide a user with the tools for navigating within them?

3. **Indexing words and organizing the lexicon:** Words and concepts can be organized in many ways, varying according to typology and conceptual systems. For example, words are traditionally organized alphabetically in Western languages, but by semantic radicals and stroke counts in Chinese. The way words and concepts are organized affects indexing and access. Indexing must robustly allow for multiple ways of navigation and access. What efficient organizational principles allow the greatest flexibility for access? What about lexical entry standardization? Are universal definitions possible? What about efforts such as the Lexical Markup Framework (LMF) and other global structures for the lexicon? Can ontologies be combined with standards for the lexicon?

4. **NLP Applications:** Contributors can also address the issue of how such enhanced dictionaries, once embedded in existing NLP applications, can boost performance and help solve lexical and textual-entailment problems such as those evaluated in SEMEVAL 2007, or, more generally, generation problems encountered in the context of summarization, question-answering, interactive paraphrasing or translation.

We've received 18 papers, of which 6 were accepted as full papers, while 8 were chosen as poster presentations. While we did not get papers on all the issues mentioned in our call, we did get a quite rich panel on ideas as divers as use of ontologies; sense extraction; computation of associative responses to multi-word stimuli; saliency relations; lexical relationships within collocations and word association norms; cognitive organization of dictionaries; user-adapted views on a lexicographic database; access based on conceptual input; search in onomasiological dictionaries, access based on underspecified input; dictionary use for authoring aids or MT, use of feature vectors, corpora and machine learning, etc..

It was also interesting to see the variety of languages in which these issues are addressed. The proposals range from Japanese, English, German, Russian, Dutch, Bulgarian, Romanian, Spanish, to French and Chinese. In sum, the community working on dictionaries is dynamic, and there seems to be a growing awareness of the importance of some of the problems presented in our call for papers.

We would like to express here our sincerest thanks to all the specialists who have assisted us to assure a good selection of papers, despite the very tight schedule. Their reviews were helpful not only for us as decision makers, but also for the authors, helping them to improve their work. In the hope that the results will inspire you, provoke fruitful discussions and result in future collaborations.

Michael Zock and Chu-Ren Huang

Organizers:

Michael Zock,LIF, CNRS, Marseille,(France)
Chu-Ren Huang, Sinica, (Taiwan)

Programme Committee:

Slaven Bilac, Google-Tokyo, (Japan)
Pierrette Bouillon, ISSCO, Geneva, (Switzerland)
Dan Cristea, University of Iasi, (Romania)
Christiane Fellbaum, Princeton, (USA)
Olivier Ferret, CEA LIST, (France)
Thierry Fontenelle, Microsoft, Redmont, (USA)
Gregory Grefenstette, CEA LIST, (France)
Graeme Hirst, University of Toronto, (Canada)
Ed Hovy, ISI, Los Angeles, (USA)
Chu-Ren Huang, Sinica, (Taiwan)
Terry Joyce, Tama University, Kanagawa-ken, (Japan)
Adam Kilgarriff, Brighton, Lexical Computing Ltd, (UK)
Philippe Langlais, University of Montreal, (Canada)
Dekang Lin, Google, Mountain View, California, (USA)
Rada Mihalcea, University of North Texas, (USA)
Alain Polguère, University of Montreal, (Canada)
Reinhard Rapp, University of Tarragona, (Spain)
Sabine Schulte im Walde, University of Stuttgart, (Germany)
Gilles Serasset, Imag, Grenoble, (France)
Anna Sinopalnikova, FIT, BUT, Brno, (Czech Republic)
Takenobu Tokunaga, Titech, Tokyo, (Japan)
Dan Tufis, RACAI, Bucharest, (Romania)
Jean Véronis, Université d'Aix-Marseille, (France)
Yorick Wilks, Oxford Internet Institute, (UK)
Michael Zock, LIF, CNRS, Marseille, (France)
Pierre Zweigenbaum, Limsi, Orsay, (France)

Table of Contents

Conference Programme

Sunday, August 24, 2008

9:00–9:10 Opening Remarks

Session 1: Regular Talks

9:10–9:50 *Comparing Lexical Relationships Observed within Japanese Collocation Data and Japanese Word Association Norms*
Terry Joyce and Irena Srdanović

10:50–10:30 *Lexical access based on underpecified input*
Michael Zock and Schwab Didier

10:30–11:00 Cofee Break + Poster Installation

11:00–11:40 *Accessing the ANW Dictionary*
Fons Moerdijk, Carole Tiberius and Jan Niestadt

Session 2: Poster Presentations (8 minutes each)

11:40–11:48 *ProPOSEL: a human-oriented prosody and PoS English lexicon for machine-learning and NLP*
Claire Brierley and Eric Atwell

11:48–11:56 *Natural Language Searching in Onomasiological Dictionaries*
Gerardo Sierra

11:56–12:04 *First ideas of user-adapted views of lexicographic data exemplified on OWID and elexiko*
Carolin Möller-Spitzer and Christine Möhrs

12:04–12:12 *Multilingual Conceptual Access to Lexicon based on Shared Orthography: An ontology-driven study of Chinese and Japanese*
Chu-Ren Huang, Ya-Min Chou, Chiyo Hotani, Sheng-Yi Chen and Wan-Ying Lin

12:12–12:20 *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*
Neculai Curteanu, Alex Moruz and Diana Trandabăţ

12:20–12:28 *Lexical-Functional Correspondences and Their Use in the System of Machine Translation ETAP-3*
Andreyeva Sasha

Sunday, August 24, 2008 (continued)

12:28–12:36 *The "Close-Distant" Relation of Adjectival Concepts Based on Self-Organizing Map*
Kyoko Kanzaki, Noriko Tomuro and Hitoshi Isahara

12:36–12:45 *Looking up phrase rephrasings via a pivot language*
Aurelien Max and Michael Zock

12:45–14:00 Lunch

Session 3: Regular Talks

14:00–14:40 *Toward a cognitive organization for electronic dictionaries, the case for semantic proxemy*
Bruno Gaume, Karine Duvignau, Laurent Prévot and Yann Desalle

14:40–15:20 *Cognitively Salient Relations for Multilingual Lexicography*
Gerhard Kremer, Andrea Abel and Marco Baroni

15:20–15:50 Coffee Break + Poster Session

15:50–16:30 *The Computation of Associative Responses to Multiword Stimuli*
Reinhard Rapp

Session 4: Poster Session + Wrap Up Discussion

16:30–17:00 Poster Session

17:00–17:30 Wrap Up Dission

17:30–17:30 End of the Workshop

Comparing Lexical Relationships Observed within Japanese Collocation Data and Japanese Word Association Norms

Terry Joyce
School of Global Studies, Tama University,
802 Engyo, Fujisawa, Kanagawa,
252-0805, JAPAN
terry@tama.ac.jp

Irena Srdanović
Tokyo Institute of Technology,
2-12-1 Ookayama, Meguro-ku,
Tokyo 152-8552, JAPAN
srdanovic.i.ab@m.titech.ac.jp

Abstract

While large-scale corpora and various corpus query tools have long been recognized as essential language resources, the value of word association norms as language resources has been largely overlooked. This paper conducts some initial comparisons of the lexical relationships observed within Japanese collocation data extracted from a large corpus using the Japanese language version of the Sketch Engine (SkE) tool (Srdanović et al., 2008) and the relationships found within Japanese word association sets taken from the large-scale Japanese Word Association Database (JWAD) under ongoing construction by Joyce (2005, 2007). The comparison results indicate that while some relationships are common to both linguistic resources, many lexical relationships are only observed in one resource. These findings suggest that both resources are necessary in order to more adequately cover the diverse range of lexical relationships. Finally, the paper reflects briefly on the implementation of association-based word-search strategies into electronic dictionaries proposed by Zock and Bilac (2004) and Zock (2006).

1 Introduction

Large-scale corpora and various corpus query tools have long been recognized as extremely important language resources. The impact of corpora and corpus query tools has been particularly significant in the area of compiling and developing lexicographic materials (Kilgarriff and Rundell, 2002) and in the area of creating various kinds of lexical resources, such as WordNet (Fellbaum, 1998) and FrameNet (Atkins et al., 2003; Fillmore et al., 2003).

In contrast, although the significance of databases of free word association norms have long been recognized within psychology in providing insights into higher cognitive processes (Cramer, 1968; Deese, 1965; Nelson et al., 1998; Steyvers and Tenenbaum, 2005), their value as a language resource has been largely overlooked. However, as Sinopalnikova and Pavel (2004) point out, databases of word association norms represent an extremely useful supplement to the range of traditional language resources, such as large-scale corpora, thesauri, and dictionaries, and can potentially contribute greatly to the development of more sophisticated linguistic resources.

This paper seeks to demonstrate the potential value of word association databases as language resources. Specifically, we conduct some initial comparisons of the lexical relationships observed within Japanese collocation data, as extracted from a large corpus with the Japanese language version of the Sketch Engine (SkE) tool (Srdanović et al., 2008), with those found within Japanese word association sets, which were created through the ongoing construction of the large-scale Japanese Word Association Database (JWAD) (Joyce, 2005, 2007). Interesting similarities and differences between the two language resources in terms of captured lexical relationships affirm the value of word association databases as rich linguistic resources. In concluding, we speculate briefly on how the wider range of lexical relationships identifiable through the combination of collocation data and word associ-

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 1–8
Manchester, August 2008

ation databases could be utilized in organizing lexical entries within electronic dictionaries in ways that are cognitively salient. While we fully acknowledge that the challenges involved are formidable ones (Zock, 2006), the principled incorporation of word association knowledge within electronic dictionaries could greatly facilitate the development of more flexible and user-friendly navigation and search strategies (Zock and Bilac, 2004).

2 Basic Concepts: Word Sketches and Word Association Norms

This section briefly provides some background information about SkE, which is the corpus query tool used in this study to extract and display word collocation data, and about word association norms as gathered through psychological experimentation.

2.1 Sketch Engine (SkE): Word Sketches and Thesaurus Tools

Sketch Engine (SkE) (Kilgarriff et al. 2004) is a web-based corpus query tool that supports a number of functions. These include fast concordancing, grammatical processing, 'word sketching' (one-page summaries of a word's grammatical and collocation behavior), a distributional thesaurus, and robot use. SkE has been applied to a number of languages. In this study, we utilize the Word Sketches and Thesaurus functions for the Japanese language. As both tools process raw collocation data by organizing words according to grammatical and lexical relationships, they are particularly suited to the conducted comparisons with the word association data.

Word Sketches (Kilgarriff and Tugwell, 2001) present the most frequent and statistically-salient collocations and grammatical relations for a given word. These relations are derived as the results of grammatical analysis (a gramrel file) that employs regular expressions over PoS-tags.

The distributional thesaurus groups together words that occur in similar contexts and have common collocation words. Estimations of semantic similarity are based on 'shared triples'. For example, <read a book> and <read a magazine> share the same triple pattern of <read a ?>, and because 'book' and 'magazine' exhibit high salience for the triple, they are both assumed to belong to the same thesaurus category. This approach is similar to conventional techniques for automatic thesaurus construction (Lin, 1998).

2.2 Word Association Norms

In contrast to the Word Sketch collocation and thesaurus tools that take the corpus as the basic input language resource, databases of word association norms are the results of psychological experiments. The free word association task typically asks the respondent to respond with the first semantically-related word that comes to mind on presentation of a stimulus word.

The collection of word association normative data can be traced back to the seminal study by Kent and Rosanoff (1910) which gathered word association responses for a list of 100 stimulus words. However, despite the insightful remarks of Deese (1965) and Cramer (1968) that word associations closely mirror the structured patterns of relations that exist among concepts—claims that undoubtedly warrant further investigation—there are, unfortunately, still relatively few large-scale databases of word association norms. The notable exceptions for the English language include the Edinburgh Association Thesaurus (EAT) (Kiss et al., 1973), which consists of approximately 56,000 responses to a stimulus list of 8,400 words, and the University of South Florida Word Association, Rhyme, and Word Fragment Norms compiled by Nelson et al. (1998), consisting of nearly three-quarters of a million responses to 5,019 stimulus words. Another database deserving mention is the Russian Association Thesaurus compiled by Karaulov et al. (1994, 1996, 1998) which has approximately 23,000 responses for 8,000 stimulus words (cited in Sinopalnikova and Pavel, 2004).

3 Japanese Language Resources

This section introduces the Japanese language resources utilized in this study: namely, the Japanese Word Sketches and Thesaurus (Srdanović et al., 2008) and the Japanese Word Association Database (Joyce, 2005, 2007).

3.1 Japanese Word Sketches and Thesaurus

The Japanese version of SkE is based on JpWaC (Erjavec et al., 2007; Srdanović et al., 2008), which is a 400-million word Japanese web corpus that has been morphologically analyzed and POS-tagged with the ChaSen tool (http://chasen.naist.jp/). The Word Sketches are based on Japanese grammatical analysis results (gramrel file), where 22 grammatical relations are defined based on ChaSen PoS tags and tokens (Srdanović et al 2008). Figure 1 presents

parts of word sketches for the noun *fuyu* (冬 winter), showing adjective modifications and two verb relations involving the particles of *wa* (は topic marker) and *ni* (に time marker), respectively.

冬 JpWaC freq = 18546

modifier_Ai 844 9.5		はverb 909 3.7		にverb 1586 3.1	
寒い	304 9.31	越せる	5 7.13	積もる	8 5.83
涼しい	17 7.31	冷え込む	5 6.23	凍る	7 5.78
厳しい	125 7.02	枯れる	6 5.51	備える	33 5.74
暖かい	20 6.46	埋もれる	5 5.37	枯れる	6 5.3
冷たい	19 6.27	冷える	5 5.17	咲く	16 5.28
暗い	25 6.13	降る	25 5.02	降る	23 4.86
長い	108 5.93	着る	7 2.96	履く	5 4.32
暑い	15 5.45	過ごす	8 2.85	向かう	36 4.02
温かい	6 5.24	引く	7 2.49	向ける	32 3.98
白い	8 4.1	続く	13 2.27	欠く	8 3.9

Figure 1. Parts of the Word Sketch results for the noun *fuyu* (冬 winter).

3.2 Japanese Word Association Database

To an even greater extent than for the English language, there has been a serious lack of word association norms for the Japanese language. While Umemoto's (1969) survey collected associations from 1,000 university students, the limited set of just 210 words merely underscores the deficient. More recently, Okamoto and Ishizaki (2001) compiled an Associative Concept Dictionary (ACD) consisting of 33,018 word association responses provided by 10 respondents for 1,656 nouns. However, it should be noted that the ACD is not strictly free association data because response category was specified as part of the task.

Under ongoing construction by Joyce (2005, 2007), the Japanese Word Association Database (JWAD) aims to eventually develop into a very large-scale database of free word association norms for the Japanese language in terms of both the number of stimulus items and the numbers of association responses collected. The present JWAD stimulus list consists of 5,000 basic Japanese kanji and words. The currently available JWAD Version 1 (JWAD-V1) consists of 104,800 free word association responses collected through a paper questionnaire survey with a sample of 2,099 items presented to up to 50 respondents. The association sets compared with work sketch profiles in the subsequent sections are from JWAD-V1.

4 Conducted Comparisons

This section presents the results of our initial comparison for the lexical relationships observed within the Japanese collocation data with those in the Japanese word association sets. The comparisons focused on approximately 350 word association responses constituting the association sets for the two verbs of *kizuku* (気付く to notice) and *sagasu* (探す to search for), the adjective of *omoshiroi* (面白い interesting), and the three nouns of *jitensha* (自転車 bicycle), *natsu* (夏 summer), and *yama* (山 mountain), as examples of basic Japanese vocabulary. Taking into account the considerable degree of orthographic variation present with the Japanese writing system, all possible orthographic variations were searched for in the SkE, such as *kizuku* (気付く/気づく) and *omoshiroi* (面白い/おもしろい).

4.1 Word Sketches and Thesaurus Versus Word Association Norms

The Japanese SkE employs a large-scale Japanese corpus and detailed grammatical analysis based on ChaSen POS tags. Accordingly, numerous lexical relationships are identified in the word sketches and thesaurus results. For example, *kizuku* appears 12,134 times in the corpus in approximately 200 collocation examples in total, which are grouped under 12 different collocation and grammatical relations and sorted according to the statistical salience of the relation's frequency within the corpus (note that searches were conducted with the default setting of only including collocations with frequencies of five or more). The thesaurus function also yields numerous results, typically displaying around 60 salient relations that are clustered into five semantic groups. In contrast, while JWAD-V1 is quite large-scale for a word association databases, it is naturally far smaller than the Japanese SkE corpus. As already noted, it consists of word association collected from about 50 respondents (although there are 100 respondents in the case of *kizuku*), and where some responses would obviously be provided by multiple respondents.

Comparisons of the SkE results with the sets of word association responses revealed that there is considerable overlap in the range of lexical relationships observed in the two linguistics resources. However, the comparisons also identified many lexical relationships that are only present in one of the language resources.

Because of the large differences in the overall sizes of the association responses in JWAD-V1 and the collocations in SkE, it is not surprising that the word association data does not cover the numerous collocation words present in the SkE results. (In future studies, we plan to examine the kinds of relationships that are extracted from the corpora but which are not observed in the word association database). However, it is very interesting to note that a considerable number of the JWAD word associations were not present in the SkE results, even though the tool is drawing on a much larger resource. In this study, we concentrate on describing these lexical relationships.

Table 1. The numbers of word association norms present (+) and absent (-) in the Word Sketches (WS) and the Thesaurus (T) results

Norms	Ass. Freq ≧ 2			Ass. Freq = 1		
	WS+	WS-	T+	WS+	WS-	T+
omoshiroi	6	5	2	1	16	2
kizuku	6	8	3	9	44	2
sagasu	4	8	1	2	13	1
jitensha	7	13	0	2	10	0
natsu	3	4	1	5	13	1
yama	6	3	2	8	7	2

Table 1 shows that considerable numbers of word association responses with frequencies of two or more, as well as many with frequencies of one, are not observed in the word sketches and thesaurus results. While these results could be indicating a need to consider new methods or approaches to corpus-extraction in addition to those currently employed, these findings also strongly suggest that some of the lexical relationships might be unique to the normative word association data. Both resources unquestionably tap into fundamental aspects of lexical relationships, but the resources would seem to be quite different in nature. Accordingly, the present results suggest that investigations into lexical relationships would do well to employ both corpus-based results and databases of word association norms in complementary ways, in order to provide more comprehensive coverage of the diverse range of lexical relationships.

The thesaurus function only outputs lexical relationships between words of the same word class. This function also yields synonym relationships that are also found in the word association norms, and are rated as being highly salient for the thesaurus results. For example, *tanoshii* and *kyomibukai* (興味深い interesting) are word association responses for *omoshiroi*.

4.2 Lexical Relationships that are Common to Both the Corpus-Based Results and the Word Association Norms

This section discusses some of the lexical relationships common to the two resources. The most frequent of these are presented in Table 2.

The first 'coord' group includes *kawa* (川 river) with the noun of *yama*, *tanoshii* (楽しい pleasant) with the adjective of *omoshiroi*, and *odoroku* (驚く to be surprised) with the verb of *kizuku*. Other frequent relationships are verbal phrases involving appropriate particles (such as nounNI (e.g., *jitensha ni noru* (自転車に乗る to ride a bicycle), noPronom, nounWO (e.g., *michi wo sagasu* (道を探す to look for a road), deVerb, niVerb). Table 2 also includes a number of modification relationships (modifier_Adv, modifier_Ai (e.g., *atsui natsu* (暑い夏 hot summer)). Note that these terms are those employed in the Word Sketch results.

Table 2. Lexical relationships common to both the Word Sketch (WS) results and the word association norms

Relationship	WS	Example
Coord	15	山・川 (*yama/kawa*), 面白い・楽しい (*omoshiroi/tanoshii*), 気付く・驚く (*kizuku/ odoroku*)
nounNI	8	間違いに気付く (*machigai ni kizuku*)
noPronom	7	自転車のかぎ (*jitensha no kagi*) 山の緑 (*yama no midori*)
gaAdj	5	山がきれい (*yama ga kirei*)
nounWO	4	道を探す (*michi wo sagasu*)
waAdj	4	夏は好き (*natsu wa suki*)
waVerb	4	自転車は走る (*jitensha wa hashiru*)
deVerb	3	自転車で転ぶ (*jitensha de korobu*)
modifier_Adv	3	ふと気付く (*futo kizuku*)
modifier_Ai	3	暑い夏 (*atsui natsu*)
niVerb	3	自転車に乗る (*jitensha ni noru*)
nounWA	3	話は面白い (*hanashi wa omoshiroi*)
woVerb	3	自転車をこぐ (*jitensha wo kogu*)

4.3 Relations Specific to Association Norms

While acknowledging that it could be beneficial to examine the types of lexical relationships observed in the corpus-based results but not in the word association data, given the relative differences in the sizes of the two resources, the present study focuses on the relationships that were only present in the database of word association norms. Briefly, these relationships can be classified under six categories.

(1) Relationships involving a specific concept related to the stimulus word and its contextual meaning. In Table 3 below, many of these are classified as 'typically associated' words. Examples include *omoshiroi* and *warai* (笑・お笑い・わらう laughter), *kizuku* and *chūi* (注意 attention), and *natsu* and *taiyō* (太陽 sun). These relationships are neither collocational nor grammatical in nature, and so the grammatical analysis currently employed in the word sketches cannot identify them. On the other hand, while they are semantically related, because they often belong to different word classes, the thesaurus function also fails to identify them.

(2) Relationships that are semantically similar (could be regarded as close synonyms) but do not belong to the same word class. Examples include *sagasu* and *tankyū* (探検 search) and *kizuku* and-*kikubari* (気配り care, attention). While these are not grammatical or collocational relations, again, the thesaurus function is also unable to find them because they belong to different word classes.

(3) Association responses consisting of more than one word. Examples include explanatory phrases such as *kibun ga ii* (気分がいい lit. 'feeling is good', comfortable) as response to *omoshiroi*, as well as concepts denoted by phrases, such as *hito no kao* (人の顔 human faces), also a response to *omoshiroi*.

(4) Relationships that could be recognized by the SkE, but which the present version fails to detect. These would seem to reflect limitations with the present ChaSen dictionary (e.g., it does not list *chari* / *charinko* (チャリ・チャリンコ casual words for bicycle) or morphological/POS-tagging errors with ChaSen, or relationships that are not regarded as being sufficiently salient within the complete corpus, because they may appear frequently as both independent words and as constituents of many poly-morpheme words (e.g., *omoshiroi hito* (面白い人 interesting person)).

(5) Relationships that can be identified when search is executed for orthographic variants of the word, such as *tsumaranai* (つまらない boring) being found when *omoshiori* is written in hiragana (as おもしろい).

(6) Word association responses that are rather idiosyncratic in nature, often reflecting private experiences of a single respondent. The importance of such responses in word association databases should be judged on the size of the database, although one also should be cautious about sampling issues with lower respondent numbers.

While it would certainly be interesting to conduct further comparisons between the association norms and other kinds of corpora, such as literary works, newspapers, or more balanced corpora, processed by the SkE, the main purpose of the present paper is to draw attention to the value of word association databases as linguistic resources. Although the lexical relationships in categories 1 and 2 were not observed in the present corpus-based results, they are unquestionably of great relevance to efforts to develop more principled organizations of the lexicon for navigational purposes, and would enhance existing lexical resources, such as WordNet. With trends to increasingly include multiple word idioms and phrases within various dictionaries and linguistic resources, the multiple-word association responses of category 3 may provide further insights into how such items are stored and processed. Moreover, categories 4 and 5 clearly suggest that free word association norms can be a very useful resource for evaluating and further improving morphological analyzers, as well as corpus query tools.

5 Lexicographical Implications: Organizing Lexicons According to Association Relationships

As the merits of SkE and its significant contributions to the compilation of a number of major dictionaries are discussed in detail elsewhere (e.g., Kilgarriff and Rundell, 2002), and because Srdanović and Nishina (2008) outline some possible lexicographical applications of the Japanese language version of the SkE, in this section, we focus on the lexical relationships observed within the JWAD and their lexicographical implications for realizing a principled association-based organization of the lexicon.

Table 3. Tentative classification of the word association responses elicited for *fuyu* (冬 winter)

Relationship	Description	Examples
Modification	Attribute: Temperate	寒い・さむい (*samui* cold)
Modification	Attribute: Color	白・白い (*shiroi* white)
Modification	Attribute: Emotion	切ない (*setsunai* bitter, severe)
Lexical siblings	Hyponyms of 'seasons'	夏 (*natsu* summer), 春 (*haru* spring)
Typically associated	Meteorological phenomena	雪 (*yuki* snow), 氷 (*koori* ice)
Typically associated	Activity	冬眠 (*tōmin* hibernation), 越冬 (*ettō* passing of winter), 休憩 (*kyūkei* rest), 休み (*yasumi* rest, holiday)
Typically associated	Cultural artifacts	こたつ (*kotatsu* quilt for lower body when sitting around low table), かまくら (*kamakura* snow hut)
Typically associated	Time	冬至 (*tōji* winter solstice)
Typically associated	Location	北 (*kita* north)
Typically associated	Animal	くま (*kuma* bear)
Typically associated	Cultural symbolization	冬将軍 (*fuyu-shōgun* General Winter; hard winter; Jack Frost)

5.1 Linguistic Approaches to Association Data and Its Potential

As previously commented, Deese (1965) and Cramer (1968) have both argued that word associations closely mirror the structured patterns of relations that exist among concepts. Indeed, as Sinopalnikova and Pavel (2004) note, Deese (1965) was the first to conduct linguistic analyses of word association norms, such as measurements of semantic similarity based on his convictions that similar words evoke similar word association responses—an approach that is somewhat reminiscent of Church and Hanks' (1990) notion of mutual information.

However, as we have also remarked already, the linguistic value of word association data has, regrettably, been largely overlooked. In a similar spirit to Hirst's (2004) claim that, notwithstanding certain caveats on the complex relationships between them, a lexicon can often serve as a useful basis for developing a practical ontology, we believe that a very promising approach to organizing the lexicon would be to more fully appreciate and utilize the rich variety of associative relationships that exist within word association norms. While the required, more thoroughgoing investigation into how to appropriately classify the complex nature of associative relationships is beyond the scope of this present study, in the next sub-section, we attempt to highlight the potential contributions that word association norms could provide to efforts seeking to explore lexical knowledge.

5.2 Tentative Classification of Association Relationships

To illustrate some of the issues for developing a comprehensive, yet a parsimonious, classification of associative relationships, it is useful to briefly consider the notion proposed by Zock and Bilac (2004) and Zock (2006) of word search strategies in electronic dictionaries based on associations. Their outline of how such a look-up system might function employs three kinds of basic association relationships; namely, 'a kind of' (AKO), 'subtype' (ISA), and 'typically involved object, relation or actor' (TIORA). While we accept that the limited set of just three types was probably motivated primarily in the interests of simplicity, given Zock's (2006) suggestion to enhance the navigability of the system by categorizing relationships, clearly the classification of association relationships is a fundamental issue.

Table 3 presents a tentative classification of the word association responses for the noun *winter*. As the comparisons introduced in Section 4 clearly demonstrate, it is usually possible to extract the modification and lexical sibling relationships included in Table 3 from corpora with corpus query tools such as SkE. However, the comparisons also highlighted the fact that it is far more difficult to identify the kinds of relationships classified in Table 3 as typically associated with such linguistic resources alone. While highly provisional in nature, we believe that the attempt to classify the association relationships within the association responses for *fuyu* can

serve to highlight some important issues for Zock and Bilac's (2004) approach.

While the lexical siblings relationships between *fuyu* and the two response words of *natsu* (夏 summer) and *haru* (春 *spring*) could feasibly be represented by AKO or ISA relationship links to *shiki* (四季 the four seasons) outside of the association set itself, having to rely on external references would not be a very satisfactory approach to classifying the direct association relationships. Incidentally, although the 'hyponyms of 'seasons'' description would seem fairly natural from the perspective of a thesaurus, the absence of *aki* (秋 autumn) from the set would indicate that the strengths of associations can vary even among lexical siblings (although the absence of *aki* from the present data could simply be due to sampling issues).

Given that *fuyu* is a noun, the presence of several modification relationships is not very surprising, at least not for the prime associate of *samui* (寒い cold), but the idea of *fuyu* having a color attribute is perhaps initially more startling (while one many not expect 'winter' to have a default color slot within its range of attributes, the association of *shiroi* (白い white) with *fuyu* is initiatively appealing).

For the *fuyu* association set, the most relevant of the association relationships specified by Zock and Bilac (2004) is the TIORA relationship. However, even for this relatively small association set containing just 11 main relationship types, because seven of them can be initially classified as 'typically associated', clearly this designation alone is too encompassing to be a useful classification category. The inclusion of the description field in Table 3 is an attempt to further define meaningful sub-categories. In the case of the sub-category 'meteorological phenomena', it would seem to be well motivated to explain the associations between *fuyu* as the stimulus word and *yuki* (雪 snow) and *kōri* (氷 *ice*) as two response words. However, while the sub-category of 'cultural artifacts' clearly goes some way to pinpointing the underlying association between *fuyu* and *kotatsu* (こたつ), it does rely on a certain cultural familiarity with the kind of *quilted kind of blanket that are used for keeping one's legs warm when sitting around a low family table during winter.* A natural association for anyone who has ever lived in Japan during the winter months, but 'typically associated' + 'cultural artifact' seems to miss something of the naturalness.

6 Conclusions

This paper has compared the lexical relationships observed within Japanese collocation data extracted from a large corpus using the Japanese language version of the Sketch Engine (SkE) tool and the relationships found within Japanese word association sets taken from the large-scale Japanese Word Association Database (JWAD).

The comparison results indicate that while many lexical relationships are common to both linguistic resources, a number of lexical relationships were only observed in one of the resources. The fact that some lexical relationships might be unique to word association norms demonstrates the value of word association databases as linguistic resources. The present findings suggest that both resources can be effectively used in combination in order to provide more comprehensive coverage of the wide range of lexical relationships.

Finally, we presented a tentative classification of the association relationships in the association set for *fuyu*. Our brief discussion of the classification sought to reflect on some of the challenges to realizing a principled association-based organization of the lexicon as a fundamental step toward implementing cognitively-salient word-search strategies based on associations in electronic dictionaries.

References

Atkins, Sue, Charles J. Fillmore, and Christopher R. Johnson. 2003. Lexicographic Relevance: Selecting Information from Corpus Evidence. *International Journal of Lexicography*, 16(3):251-280.

Church, Kenneth W., and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and. Lexicography. Computational Linguistics, 16(1): 22-29.

Cramer, Phebe. 1968. *Word Association*. New York and London: Academic Press.

Deese, John. 1965. *The Structure of Associations in Language and Thought*. Baltimore: The John Hopkins Press.

Erjavec, Tomaž, Adam Kilgarriff, Irena Srdanović Erjavec. 2007. A Large Public-access Japanese Corpus and its Query Tool. *The Inaugural Workshop on Computational Japanese Studies*. Ikaho, Japan.

Fellbaum, Christiane (Ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA, MIT Press.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography,* 16(3):235-250.

Hirst, Graeme. 2004. Ontology and the Lexicon. Steffen Staab, and Rudi Studer (Eds.) *Handbook of Ontologies*. pp. 209-229. Berlin: Springer-Verlag.

Joyce, Terry. 2005. Constructing a Large-scale Database of Japanese Word Associations. Katsuo Tamaoka. *Corpus Studies on Japanese Kanji*. (Glottometrics 10), 82-98. Tokyo, Japan; Hituzi Syobo and Lüdenschied, Germany: RAM-Verlag.

Joyce, Terry. 2007. Mapping Word Knowledge in Japanese: Coding Japanese Word Associations. *Proceedings of the Symposium on Large-scale Knowledge Resources (LKR2007)*, 233-238. Tokyo, Japan: Tokyo Institute of Technology.

Karaulov, Ju. N., G. A. Cherkasova, N. V. Ufimtseva, Ju. A. Sorokin, and E. F. Tarasov. 1994, 1996, 1998. *Russian Associative Thesaurus*. Moscow.

Kent, Grace H., and A. J. Rosanoff. 1910. A Study of Association in Insanity. *American Journal of Insanity*, 67:317-390.

Kilgarriff, Adam and Michael Rundell, 2002. Lexical Profiling Software and its Lexicographic Applications: A Case Study. Anna Braasch and Claus Povlsen (Eds). *Proceedings of the Tenth EURALEX International Congress*. pp. 807-818. Copenhagen, Denmark.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress.* pp 105-116. Lorient, France.

Kilgarriff Adam, and David Tugwell. 2001. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proceedings of the 39th ACL Workshop on Collocations: Computational Extraction, Analysis and Exploitation*, 32-38. Toulouse: France.

Kiss, George, Christine Armstrong, Robert Milroy, and J. Piper. 1973. An associative thesaurus of English and its computer analysis. A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (Eds.). *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Lin, Dekang. 1998. Automatic retrieval; and clustering of similar words. *COLING-ACL* Montreal: 768-774.

Nelson, Douglas L., Cathy L. McEvoy, and Thomas A. Schreiber. 1998. *The University of South Florida Word Association, Rhyme, and Word Fragment Norms*. http://www.usf.edu/FreeAssociation.

Okamoto, Jun, and Shun Ishizaki. 2001. Associative Concept Dictionary and its Comparison with Electronic Concept Dictionaries. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 94–103.

Sinopalnikova, Anna, and Pavel Smrž. 2004. Word Association Norms as a Unique Supplement of Traditional Language Resources. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1557-1561. Lisbon, Portugal: Centro Cultural de Belem.

Srdanović Erjavec, Irena, Tomaž Erjavec, and Adam Kilgarriff. 2008. A web corpus and word-sketches for Japanese. *Journal of Natural Language Processing*, 15/2.

Srdanović, I. E. and Nishina, K. (2008). "Ko-pasu kensaku tsu-ru Sketch Engine no nihongoban to sono riyou houhou (The Sketch Engine corpus query tool for Japanese and its possible applications)." Nihongo kagaku (Japanese Linguistics), 24, pp. 59–80.

Steyvers, Mark, and Joshua B. Tenenbaum. 2005. The Large-scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29:41-78.

Umemoto, Tadao. 1969. *Table of Association Norms: Based on the Free Associations of 1,000 University Students*. (in Japanese). Tokyo: Tokyo Daigaku Shuppankai.

Zock, Michael. 2006. Navigational Aids, a Critical Factor for the Success of Electronic Dictionaries. Reinhard Rapp, Peter Sedlmeier and Gisela Zunker-Rapp (Eds.) *Perspectives on Cognition: A Festschrift for Manfred Wettler*. Pabst Science Publishers, Lengerich.

Zock, Michael and Slaven Bilac. 2004. Word Lookup on the Basis of Associations: From an Idea to a Roadmap. *Workshop on Enhancing and Using Electronic Dictionaries at the 20th International Conference on Computational Linguistics*. Geneva, Switzland.

Lexical Access Based on Underspecified Input

Michael ZOCK
LIF-CNRS
Équipe TALEP
163, Avenue de Luminy
F-13288 Marseille Cedex 9
michael.zock@lif.univ-mrs.fr

Didier SCHWAB
Groupe GETALP
Laboratoire d'Informatique de Grenoble
385 avenue de la Bibliothque - BP 53
F-38041 Grenoble Cedex 9
didier.schwab@imag.fr

Abstract

Words play a major role in language production, hence finding them is of vital importance, be it for speaking or writing. Words are stored in a dictionary, and the general belief holds, the bigger the better. Yet, to be truly useful the resource should contain not only many entries and a lot of information concerning each one of them, but also adequate means to reveal the stored information. Information access depends crucially on the organization of the data (words) and on the navigational tools. It also depends on the grouping, ranking and indexing of the data, a factor too often overlooked.

We will present here some preliminary results, showing how an existing electronic dictionary could be enhanced to support language producers to find the word they are looking for. To this end we have started to build a corpus-based *association matrix*, composed of *target words* and *access keys* (meaning elements, related concepts/words), the two being connected at their intersection in terms of *weight* and *type of link*, information used subsequently for grouping, ranking and navigation.

1 Context and problem

When speaking or writing we encounter basically either of the following two situations: one where everything works automatically, somehow like magic, words popping up one after another like spring water, and another where we look deliberately and often painstakingly for a specific, possibly known word. We will be concerned here with this latter situation: a speaker/ writer using an electronic dictionary to look for such a word. Unfortunately, alphabetically organized dictionaries are not well suited for this kind of *reverse lookup* where the inputs are meanings (elements of the word's definition) or conceptually related elements (collocations, associations), and the outputs the target words.

Without any doubt, lexicographers have made considerable efforts to assist language users, building huge resources, composed of many words and lots of information associated with each one of them. Still, it is not unfair to say most dictionaries have been conceived from the reader's point of view. The lexicographers have hardly taken into account the language producer's perspective,[1] considering conceptual input, incomplete as it may be, as starting point. While *readers* start with words, looking generally for their corresponding meanings, *speakers* or *writers* usually start with the opposite, meanings or concepts,[2] which should be the entry points of a dictionary, which ideally is neutral in terms of access direction.[3]

The problem is that we still don't know very well what *concepts* are, whether they are compositional and if so, how many *primitives* there are (Wilks, 1977; Wierzbicka, 1996; Goddard, 1998).

[1] Roget's *thesaurus* (Roget, 1852), Miller and Fellbaum's *WordNet* (Fellbaum, 1998) and Longman's *Language Activator* (Summers, 1993), being notable exceptions (For more details, see next section).

[2] Of course, this does not preclude, that we may have to use *words* to refer to them in a concept-based query.

[3] While we agree with Polguère theoretically when he pleads for dictionary neutrality with regard to lexical access (Polguère, 2006), from a practical point of view the situation is obviously quite different for the speaker and listener, even if both of them draw on the same resource.

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 9–17
Manchester, August 2008

Neither do we know how to represent them. Yet, there are ways around this problem as we will show. Whether *concepts* and *words* are organized and accessed differently is a question we cannot answer here. We can agree though on the fact that getting information concerning words is fairly unproblematic when reading, at least in the case of most western languages. Words can generally be found easily in a dictionary, provided the user knows the spelling, the alphabet and how to build lemma starting from an inflected form. Unlike *words*, which are organized alphabetically (in western languages) or by form (stroke counts in Chinese), *concepts* are organized topically: they are clustered into functional groups according to their role in real world, or our perception of it.

Psychologist have studied the difficulties people have when trying to produce or access words (Aitchinson, 2003). In particular, they have studied the *tip-of-the-tongue phenomenon* (Brown and McNeill, 1996) and the effects an input can have on the *quality* of an output (error analysis (Cutler, 1982)) and on the *ease* of its production: positive or negative priming effect (activation/inhibition). Obviously, these findings allow certain conclusions, and they might guide us when developing tools to help people find the needed word. In particular, they reveal two facts highly relevant for our goal:

1. even if people fail to access a given word, they might know a lot about it: *origin*, *meaning* (word definition, role played in a given situation), *part of speech*, *number of syllables*, *similar sounding* words, etc. Yet, despite all this knowledge, they seem to lack some crucial information to be able to produce the phonetic form. The word gets blocked at the very last moment, even though it has reached the *tip-of-the-tongue*. This kind of nuisance is all the more likely as the target word is rare and primed by a similar sounding word.

2. unlike words in printed or electronic dictionaries, words in our mind may be inexistent as tokens. What we seem to have in our minds are decomposed, abstract entities which need to be synthesized over time.[4] According to Levelt (Levelt, 1996) the generation of words (synthesis) involves the following stages: conceptual preparation, lexical selection, phonological- and phonetic encoding, articulation. Bear in mind that having performed 'lexical selection' does not imply access to the phonetic form (see the experiments on the *tip-of-the-tongue phenomenon*).

What can be concluded from these observations? It seems that underspecified input is sufficiently frequent to be considered as normal. Hence we should accept it, and make the best out of it by using whatever information is available (accessible), no matter how incomplete, since it may still contribute to find the wanted information, be it by reducing the search space. Obviously, the more information we have the better, as this reduces the number of words among which to choose.

2 Related work and goal

While more dictionaries have been built for the reader than for the writer, there have been some onomasiological attempts as early as in the middle of the 19th century. For example, Roget's *Thesaurus* (Roget, 1852), T'ong's *Chinese and English instructor* (T'ong, 1862), or Boissiere's *analogical dictionary* (Boissière, 1862).[5] Newer work includes Mel'čuk's *ECD* (Mel'čuk et al., 1999), Miller and Fellbaum's *WordNet* (Fellbaum, 1998), Richardson and Dolan's *MindNet* (Richardson et al., 1998), Dong's *HowNet* (Dong and Dong, 2006) and Longman's *Language Activator* (Summers, 1993). There is also the work of

[4] This may be very surprising, yet, this need not be the case if we consider the fact that speech errors are nearly always due to competing elements from the same level or an adjacent one, unless they are the result of a surrounding concept which has been activated, or which is about to be translated into words. Put differently, we do not store words at all in our mind, at least not in the layman's or lexicographer's sense who consider word-forms and their meanings as one. If we are right, than rather continue to consider the human mind as a *word store* we could consider it as a *word factory*. Indeed, by looking at some of the work done by psychologists who try to emulate the mental lexicon (for a good survey see (Harley, 2004), pages 359-374) one gets the impression that words are synthesized rather than located and read out. Taking a look at all this work, generally connectionist models, one may conclude that, rather than having words in our mind we have a set of more or less abstract features (concepts, syntactic information, phonemes), distributed across various layers, which need to be synthesized over time. To do so we proceed from abstract meanings to concrete sounds, which at some point were also just abstract features. By propagating energy rather than data (as there is no message passing, transformation or cumulation of information, there is only activation spreading, that is, changes of energy levels, call it weights, electronic impulses, or whatever), that we propagate signals, activating ultimately certain peripheral organs (larynx, tongue, mouth, lips, hands) in such a way as to produce movements or sounds, that, not knowing better, we call words.

[5] For a more recent proposal see (Robert et al., 1993).

(Fontenelle, 1997; Sierra, 2000; Moerdijk, 2008), various *collocation dictionaries* (BBI, OECD) and Bernstein's *Reverse Dictionary*.[6] Finally, there is M. Rundell's MEDAL, a thesaurus produced with the help of Kilgariff's Sketch Engine (Kilgarriff et al., 2004).

As one can see, a lot of progress has been accomplished over the last few years, yet more can be done, especially with regard to unifying *linguistic* and *encyclopedic* knowledge. Let's take an example to illustrate our point.

Suppose, you were looking for a word expressing the following ideas: 'superior dark coffee made from beans from Arabia', and that you knew that the target word was neither *espresso* nor *cappuccino*. While none of this would lead you directly to the intended word, *mocha*, the information at hand, i.e. the word's definition or some of its elements, could certainly be used. In addition, people draw on knowledge concerning the *role* a concept (or word) plays in language and in real world, i.e. the associations it evokes. For example, they may know that they are looking for a *noun* standing for a *beverage* that *people* take under certain circumstances, that the *liquid* has certain properties, etc. In sum, people have in their mind an encyclopedia: all words, concepts or ideas being highly connected. Hence, any one of them has the potential to evoke the others. The likelihood for this to happen depends, of course, on factors such as *frequency* (associative strength), *distance* (direct vs. indirect access), *prominence* (saliency), etc.

How is this supposed to work for a dictionary user? Suppose you were looking for the word *mocha* (target word: t_w), yet the only token coming to your mind were *computer* (source word: s_w). Taking this latter as starting point, the system would show all the connected words, for example, *Java, Perl, Prolog* (programing languages), *mouse, printer* (hardware), *Mac, PC* (type of machines), etc. querying the user to decide on the direction of search by choosing one of these words. After all, s/he knows best which of them comes closest to the t_w. Having started from the s_w 'computer', and knowing that the t_w is neither some *kind of software* nor a *type of computer*, s/he would probably choose *Java*, which is not only a *programming language* but also an *island*. Taking this latter as the new starting point s/he might choose *coffee* (since s/he is looking for some kind of *beverage*, possibly made from an ingredient produced in Java, coffee), and finally *mocha*, a type of *beverage* made from these beans. Of course, the word *Java* might just as well trigger *Kawa* which not only rhymes with the s_w, but also evokes *Kawa Igen*, a javanese volcano, or familiar word of *coffee* in French.

As one can see, this approach allows word access via multiple routes: there are many ways leading to Rome. Also, while the distance covered in our example is quite unusual, it is possible to reach the goal quickly. It took us actually very few moves, four, to find an indirect link, between two, fairly remotely related terms: *computer* and *mocha*. Of course, *cyber-coffee* fans might be even quicker in reaching their goal.

3 The lexical matrix revisited

The main question that we are interested in here is how, or in what terms, to index the dictionary in order to allow for quick and intuitive access to words. Access should be possible on the basis of meaning (or meaning elements), various kinds of associations (most prominently 'syntagmatic' ones) and, more generally speaking, underspecified input. To this end we have started to build an *association matrix* (henceforth AM), akin to, yet different from G. Miller's initial proposal of WN (Miller et al., 1990). He suggested to build a lexical matrix by putting on one axis all the *forms*, i.e. words of the language, and on the other, their corresponding *meanings*. The latter being defined in terms of synsets. The corresponding meaning-form relations are signaled via a boolean (presence/absence). Hence, looking at the intersection of meanings and forms, one can see which meanings are expressed by, or converge toward what forms, or conversely, what form expresses which meanings. Whether this is the way WN is actually implemented is not clear to us, though we believe that it is not. Anyhow, our approach is different, and we hope the reader will understand in a moment the reasons why.

We will also put on one axis all the form elements, i.e. the *lemmata* or expressions of a given language (we refer to them as *target words*, henceforth t_w). On the other axis we will place the *triggers* or *access-words* (henceforth a_w), that is, the words or concepts capable and likely to evoke the t_w. These are typically the kind of words psy-

[6] There is also at least one electronic incarnation of a dictionary with reverse access, combining a dictionary (WordNet) and an encyclopedia (Wikipedia) (http://www.onelook.com/reverse-dictionary.shtml).

chologists have gathered in their association experiments (Jung and Riklin, 1906; Deese, 1965; Schvaneveldt, 1989). Note, that instead of putting a boolean value at the intersection of the t_w and the a_w, we will put *weights* and the *type of link* holding between the co-occurring terms. This gives us quadruplets. For example, an utterance like "this is the key of the door" might yield the a_w (key), the t_w(door), the link type l_t(part of), and a weight (let's say 15).

The fact that we have these two kinds of information is very important later on, as it allows the search engine to cluster by type the possible answers to be given in response to a user query (word(s) provided as input) and to rank them. Since the number of hits, i.e. words from which the user must choose, may be substantial (depending on the degree of specification of the input), it is important to group and rank them to ease navigation, allowing the user to find directly and quickly the desired word, or at least the word with which to continue search.

Obviously, different word senses (homographs), require different entries (bank-money vs bank-river), but so will synonyms, as every word-form, synonym or not, is likely to be evoked by a different key- or access-word (similarity of sound).[7]

Also, we will need a new line for every different relation between a a_w and a t_w. Whether more than one line is needed in the case of identical links being expressed by different linguistic resources (the lock *of* the door vs. the door*'s* lock vs. the door *has* a lock) remains an open empirical question.

Let us see quickly how our AM is supposed to work. Imagine you wanted to find the word for the following concept: *hat of a bishop*. In such a case, any of the following concepts or words might come to your mind: church, Vatican, abbot, monk, monastery, ceremony, ribbon, and of course rhyming words like: brighter, fighter, lighter, righter, tighter, writer,[8] as, indeed, any of them could remind us of the t_w: *mitre*. Hence, all of them are possible a_w.

Once this resource is built, access is quite straightforward. The user gives as input all the words coming to his mind when thinking of a given idea or concept,[9] and the system will display all connected words. If the user can find the item he is looking for in this list, search stops, otherwise it will continue, the user giving other words of the list, or words evoked by them.

Of course, remains the question of how to build this resource, in particular, how to populate the axis devoted to the trigger words, i.e. *access-keys*. At present we consider three approaches: one, where we use the words occurring in word definitions (see also, (Dutoit and Nugues, 2002; Bilac et al., 2004)), the other is to mine a well-balanced corpus, to find co-occurrences within a given window (Ferret and Zock, 2006), the size depending a bit on the text type (encyclopedia) or type of corpus. Still another solution would be to draw on the association lists produced by psychologists, see for example http://www.usf.edu/, or http://www.eat.rl.ac.uk.

Of course, the idea of using matrices in linguistics is not new. There are at least two authors who have proposed its use: M. Gross (Gross, 1984) used it for coding the syntactic behavior of lexical items, hence the term *lexicon-grammar*, and G. Miller, the father of WN (Miller et al., 1990) suggested it to support lexical access. While the former work is not relevant for us here, Miller's proposal is. What are the differences between his proposal and ours? There are basically four main differences:

1. we use, collocations or *access-words*, i.e a_{ws} rather than *synsets*; Hence, any of the following a_{ws} (cat, grey, computer device, cheese, Speedy Gonzales) could point toward the t_w 'mouse', none of them are part of the meaning, leave alone synonyms.

2. we mark explicitly the *weight* and the *type of link* between the t_w and the a_w (isa, part_of, etc.),[10] whereas WN uses only a binary value. Both the *weight* and *link* are necessary information for ranking and grouping, i.e. navigation.

3. our AM is corpus-sensitive (see below), hence, we can, at least in principle, accommo-

[7] Take, for example, the nouns *rubbish* and *garbage* which can be considered as synonyms. Yet, while the former may remind you of a *rabbit* or (horse)-*radish*, the latter may evoke the word *cabbage*.

[8] The question, whether rhyming words should be computed is not crucial at this stage.

[9] The quantifier *all* shouldn't be taken too literally. What we have in mind are "salient" words available in the speaker's mind at a given moment

[10] Hence, if several links are possible between the t_w and the a_w, several cells will be used. Think of the many possible relations between a city and a country, example: *Paris* and *France* (part of, biggest city of, located in, etc.)

date the fact that a speaker is changing topics, adapting the weight of a given word or find a more adequate a_w in this new context. Think of 'piano' in the contexts of a concert or moving your household. Only the latter would evoke the notion of weight.

4. relying on a corpus, we can take advantage of *syntagmatic associations* (often encyclopedic knowledge), something which is difficult to obtain for WN.

4 Keep the set of lexical candidates small

Here and in the next section we describe how the idea of the AM has been computationally dealt with. The goal is to reduce the number of hits, i.e. possible t_{ws} (output), as a function of the input, i.e, the number of relevant a_{ws} given by the speaker/writer. To achieve this goal we apply lexical functions to the a_{ws}, considering the intersection of the obtained sets to be the relevant t_{ws}.

4.1 Lexical Functions

The usefulness of *lexical functions* for linguistics in general and for language production in particular has been shown by Mel'čuk (Mel'čuk, 1996). We will use them here, as they seem to fit also our needs of information extraction or lexical access.

Mel'čuk has coined the term *lexical functions* to refer to the fact that two terms are systematically related. For example, the lexical function *Gener* refers to the fact that some term (let's say ‹*cat*›) can be replaced by a more general term (let's say ‹*animal*›).

Lexical functions encode the combinability of words. While 'big' and 'strong' express the same idea (intensity, magnitude), they cannot be combined freely with any noun: *strong* can be associated with *fever*, whereas *big* cannot. Of course, this kind of combinability between lexical terms is language specific, because unlike in English, in French one can say *grosse fièvre* or *forte fièvre*, both being correct (Schwab and Lafourcade, 2007). Our AM handles, of course these kind of functions. Here is a list of some of them:

- *paradigmatic associations*: hypernymy (‹*cat*› - ‹*animal*›), hyponymy, synonymy, or antonymy,...;

- *syntagmatic associations*: collocations (‹*fear*› being associated with ‹*strong*› or ‹*little*›);

- *morphological relations* ie. terms being derived from another part of speech: applying the *change-part-of-speech* lexical function f_{cpos} to ‹*garden*› will yield: $f_{cpos}(\text{‹}garden\text{›}) = \{\text{‹}to\ garden\text{›}, \text{‹}gardener\text{›}, \dots\}$

- *sound-related items*: homophones, rhymes.

4.2 Assumptions concerning search

The purpose of using lexical functions is to reduce the number of possible outcomes from which the user must choose. The list contains either the t_w or another promising a_w the user may want use to continue search. Hence, lexical functions are useful for search provided that:

1. the speaker/writer is able to specify the kind of relations s/he wants to use. The problem here lies in the nature and number of the functions, some of them being very well specified, while others are not.

2. the larger the number of trigger words the smaller the list of words from which to choose: the speaker/writer can add or delete words to broaden or narrow the scope of his/her query.

These hypotheses are being modeled by using set properties of lexical functions. The idea is to apply all functions, or a selection of them, to the a_{ws} and to give the speaker/writer the intersection as result (see section 5.3.5 for an example)

5 Experiment

We have started with a simple, preliminary experiment. Only one lexical function was used: neighborhood (henceforth f_{neig}). Let f_{neig} be the function producing the set of co-occurring terms within a given window (sentence or a paragraph).[11] The result produced by the system and returned to the user is the intersection of the application of f_{neig} to the a_{ws}. In the next section we explain how this function is applied to two corpora (Wordnet and Wikipedia), to show their respective qualities and shortcomings for this specific task.

5.1 WordNet

5.1.1 Description

WordNet (henceforth WN) is a lexical database for English developed under the guidance of G.

[11] The scope or window size will vary with the text type (normal text vs. encyclopedia). The optimal size is at this point still an empirical question.

Miller (Miller et al., 1990). One of his goals was to support lexical access akin to the human mind, association-based. *Knowledge* is stored in a network composed of nodes and links (nodes being words or concepts and the links are the means of connecting them) and *access to knowledge*, i.e. search, takes place by entering the network at some point and follow the links until one has reached the goal (unless one has given up before). This kind of *navigation* in a huge conceptual/lexical network can be considered equivalent to *spreading activation* taking place in our brain.

Of course, such a network has to be built, and navigational support must be provided to find the location where knowledge or words are stored. This is what Miller and his coworkers did by building WN. The resource has been built manually, and it contains at present about 150.000 entries.

The structure of the dictionary is different from conventional, alphabetical resources. Words are organized in WN in two ways. Semantically similar words, i.e. synonyms, are grouped as clusters. These sets of synonyms, called *synsets*, are then linked in various ways, depending on the kind of relationship they entertain with the adjacent synset. For example, their neighbors can be more *general* or *specific* (hyperonymy vs. hyponymy), they can be *part of* some reference object (meronymy: car-motor), they can be the *opposite* (antonymy: hot-cold), etc. While WN is a resource it can also be seen as a corpus.

5.1.2 Using WN as a corpus

There are many good reasons to use WN for learning f_n. For one, there are many extensions, and second, the one we are using, eXtended WN (Mihalcea and Moldovan, 2001) spares us the trouble of having to address issues like: (a) segmentation: we do not need to identify sentence boundaries ; (b) semantic ambiguity: words being tagged, we get good precision; (c) lemmatization: since only verbs, nouns, adjectives and adverbs are tagged, we need neither a stoplist nor a lemmatizer.

Despite all these qualities, two important problems remain nevertheless for this kind of corpus: (a) size: though, all words are tagged, the corpus remains small as it contains only 63.941 different words; (b) in consequence, the corpus lacks many *syntagmatic associations* encoding encyclopedic knowledge.

5.2 Using Wikipedia as corpus

Wikipedia is a free, multilingual encyclopedia, accessible on the Web.[12] For our experiment we have chosen the English version which of this day (12th of may 2008) contains 2,369,180 entries.

Wikipedia has exactly the opposite properties of WN. While it covers well encyclopedic relations, it is only raw text. Hence problems like text segmentation, lemmatisation and stoplist definition need to be addressed.

Our experiments with Wikipedia were very rudimentary, given that we considered only 1000 documents. These latter were obtained in response to the term ‹*wine*›, by following the links obtained for about 72.000 words.

5.3 Prototype

5.3.1 Building the resource and using it.

Building the resource requires processing a corpus and building the database. Given a corpus we apply our neighborhood function to a predetermined window (a paragraph in the case of encyclopedias).[13] The result, i.e. the co-occurrences, will be stored in the database, together with their *weight*, i.e. number of times two terms appear together, and the *type of link*. As mentionned above, both kinds of information are needed later on for *ranking* and *navigation*.[14]

At present, cooccurences are stored as triplets (t_w, a_w, *times*), where *times* represents the number of times the two terms cooccur in the corpus, the scope of coccurence being here the paragraph.

5.3.2 Processing of the Wikipedia page

For each Wikipedia page, a preprocessor converts HTML pages into plain text. Next, a part-of-speech tagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/) is used to annotate all the words of the paragraph under consideration. This allows the filtering of all irrelevant words, to keep but a bag of words, that is, the nouns, adjectives, verbs and adverbs occuring in the paragraph. These words will be used to fill the triplets of our database.

[12] `http://www.wikipedia.org`

[13] The optimal window-size depends probably on the text type (encyclopedia vs. unformatted text). Yet, in the absence of clear criteria, we consider the optimal window-size as an open, empirical question.

[14] This latter aspect is not implemented yet, but will be added in the future, as it is a necessary component for easy navigation (Zock and Bilac, 2004; Zock, 2006; Zock, 2007).

5.3.3 Corpus Building

We start arbitrarily from some page (for our experiment, we have chosen "wine" as input), apply the algorithm outlined here above and pick then randomly a noun within this page to fetch with this input a new page on Wikipedia. This process is repeated until a given sample size is obtained (in our case 1000 pages). Of course, instead of picking randomly a noun, we could have decided to process all the nouns of a given page, and to add then incrementally the nouns of the next pages. Yet, doing this would have led us to privilege a specific topic (in our case 'wine') instead of a more general one.

5.3.4 Usage

We have developed a website in Java as a servlet. Interactions with humans are simple: people can add or delete a word from the current list (see *Input* in the figure on top of the next page). The example presented shows that with very few words, hence very quickly, we can obtain the desired word.

Given some input, the system provides the user with a list of words cooccuring with the a_{ws}. The output is an ordered list of words, the order depending on the overall score, i.e. number of cooccurrences between the a_w and the t_w. For example, if the a_{ws} 'wine' and 'harvest' co-occur with the t_w 'bunch' respectively 5 and 8 times, then the overall score of cooccurence of 'bunch' is 13: ((wine, harvest), bunch, 13). Hence, all words with a higher score will precede it, while those with a lower score will follow it.

5.3.5 Examples and Comparison of the results of the two corpora

Here below are the examples extracted from the WN corpus (see figure-1). Our goal was to find the word ‹*vintage*›. Trigger words are ‹*wine*› and ‹*harvest*›, yielding respectively 488 and 30 hits, i.e. words. As one can see ‹*harvest*› is a better access term than ‹*wine*›. Combining the two will reduce the list to 6 items. Please note that the t_w ‹*vintage*› is not among them, eventhough it exists in WordNet, which illustrates nicely the fact that storage does not guarantee accessibility (Sinopalnikova and Smrz, 2006).

Looking at figure-1 you will see that the results have improved considerably with Wikipedia. The same input, ‹*wine*› evokes many more words (1845 as opposed to 488). For ‹*harvest*› we get 983 hits instead of 30 (the intersection containing 62 words). Combining the two reduces the set to 45 items among which we will find, of course, the target word.

Input	WordNet		Wikipedia	
	488 words		**1845 words**	
	grape	sweet	alcoholic	country
	serve	france	god	characteristics
wine	small	fruit	regulation	grape
	dry	bottle	appellation	system
	produce	red	bottled	like
	bread	hold	christian	track
	…	…	…	…
	30 words		**983 words**	
	month	fish	produce	grain
	grape	revolutionary	autumn	farms
	calendar	festival	energy	cut
harvest	butterfish	dollar	combine	ground
	person	make	balance	rain
	wine	first	amount	rich
	…	…	…	…
	6 words		**45 words**	
	make	grape	grape	vintage
wine	fish	someone	bottle	produce
+harvest	commemorate	person	fermentation	juice
	…	…	Beaujolais	taste
			viticulture	France
			Bordeaux	vineyard
			…	…

Figure 1: Comparing two corpora (*eXtended WordNet* and *Wikipedia*) with various inputs

We hope that this example is clear enough to convince the reader that it makes sense to use real text as corpus to extract from it the kind of information (associations) people are likely to give when looking for a word.

6 Conclusion and perspectives

We have addressed in this paper the problem of word finding for speakers or writers. Concluding that most dictionaries are not well suited to allow for this kind of reverse access based on meanings (or meaning related elements, associations), we looked at work done by psychologists to get some inspiration. Next we tried to clarify which of these findings could help us build the dictionary of tomorrow, that is, a tool integrating linguistic and encyclopedic knowledge, allowing navigation by taking either or as starting point. While linguistic knowledge is more prominent for analysis (reading), encyclopedic facts are more relevant for production. We've presented then our ideas of how to build a resource, allowing lexical access based

on underspecified, i.e. imperfect input. To achieve this goal we've started building an AM composed of form elements (the words and expressions of a given language) and a_{ws}. The role of the latter being to lead to or to evoke the t_w. In the last part we've described briefly the results obtained by comparing two resources (WN and Wikipedia) and various inputs. Given the fact that the project is still quite young, only preliminary results can be shown at this point.

Our next steps will be to take a closer look at the following work: clustering of similar words (Lin, 1998), topic signatures (Lin and Hovy, 2000) and Kilgariff's sketch engine (Kilgarriff et al., 2004). We plan also to add other lexical functions to enrich our database with a_{ws}. We plan to experiment with corpora, trying to find out which ones are best for our purpose[15] and we will certainly experiment with the window size [16] to see which size is best for which text type. Finally, we plan to insert in our AM the relations holding between the a_w and the t_w. As these links are contained in our corpus, we should be able to identify and type them. The question is, to what extent this can be done automatically.

Obviously, the success of our resource will depend on the quality of the corpus, the quality of the a_{ws}, weights and links, and the representativity of all this for a given population. While we do believe in the justification of our intuitions, more work is needed to reveal the true potential of the approach. The ultimate judge being, of course, the future user.

[15]For example, we could consider a resource like ConceptNet of the Open Mind Common-Sense project (Liuh and Singh, 2004).

[16]For example, it would have been interesting to consider coocurrences beyond the scope of the paragraph, by considering the logical structure of the Wikipedia document. Anyhow, our experiment needs to be redone with more data than just 1000 pages, the size chosen here for lack of time. Indeed one could consider using the entire corpus of Wikipedia or mixed corpora

References

Aitchinson, Jean. 2003. *Words in the Mind: an Introduction to the Mental Lexicon (3d edition)*. Blackwell, Oxford.

Bilac, S., W. Watanabe, T. Hashimoto, T. Tokunaga, and H. Tanaka. 2004. Dictionary search based on the target word description. In *Proc. of the Tenth Annual Meeting of The Association for NLP (NLP2004)*, pages 556–559, Tokyo, Japan.

Boissière, P. 1862. *Dictionnaire analogique de la langue française : répertoire complet des mots par les idées et des idées par les mots*. Larousse et A. Boyer, Paris.

Brown, R. and D. McNeill. 1996. The tip of the tounge phenomenon. *Journal of Verbal Learning and Verbal Behaviour*, 5:325–337.

Cutler, A, editor, 1982. *Slips of the Tongue and Language Production*. Mouton, Amsterdam.

Deese, James. 1965. *The structure of associations in language and thought*. Johns Hopkins Press.

Dong, Zhendong and Qiang Dong. 2006. *HOWNET and the computation of meaning*. World Scientific, London.

Dutoit, Dominique and P. Nugues. 2002. A lexical network and an algorithm to find words from definitions. In van Harmelen, F., editor, *ECAI2002, Proc. of the 15th European Conference on Artificial Intelligence*, pages 450–454, Lyon. IOS Press, Amsterdam.

Fellbaum, Christiane, editor, 1998. *WordNet: An Electronic Lexical Database and some of its Applications*. MIT Press.

Ferret, Olivier and Michael Zock. 2006. Enhancing electronic dictionaries with an index based on associations. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 281–288.

Fontenelle, Thierry. 1997. Using a bilingual dictionary to create semantic networks. *International Journal of Lexicography*, 10(4):275–303.

Goddard, Cliff. 1998. Bad arguments against semantic primitives. *Theoretical Linguistics*, 24(2-3):129–156.

Gross, Maurice. 1984. Lexicon-grammar and the analysis of french. In *Proc. of the 11th COLING*, pages 275–282, Stanford, CA.

Harley, Trevor. 2004. *The psychology of language*. Psychology Press, Taylor and Francis, Hove and New York.

Jung, Carl and F. Riklin. 1906. Experimentelle Untersuchungen über Assoziationen Gesunder. In Jung, C. G., editor, *Diagnostische Assoziationsstudien*, pages 7–145. Barth, Leipzig, Germany.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France.

Levelt, Willem. 1996. A theory of lexical access in speech production. In *Proc. of the 16th Conference on Computational Linguistics*, Copenhagen, Denmark.

Lin, Chin-Yew and Eduard H. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING*, pages 495–501. Morgan Kaufmann.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774, Montreal.

Liuh, H. and P. Singh. 2004. ConceptNet: a practical commonsense reasoning toolkit. *BT Technology Journal*.

Mel'čuk, I., N. Arbatchewsky-Jumarie, L. Iordanskaja, S. Mantha, and A. Polguère. 1999. *Dictionnaire explicatif et combinatoire du français contemporain Recherches lexico-sémantiques IV*. Les Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor. 1996. Lexical functions: A tool for the description of lexical relations in the lexicon. In Wanner, L., editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins, Amsterdam/Philadelphia.

Mihalcea, Rada and Dan Moldovan. 2001. Extended WordNet: progress report. In *NAACL 2001 - Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3(4)*, pages 235–244.

Moerdijk, Fons. 2008. Frames and semagrams; Meaning description in the general dutch dictionary. In *Proceedings of the Thirteenth Euralex International Congress, EURALEX*, Barcelona.

Polguère, Alain. 2006. Structural properties of lexical systems: Monolingual and multilingual perspectives. Sidney. Coling workshop 'Multilingual Language Resources and Interoperability'.

Richardson, S., W. Dolan, and L. Vanderwende. 1998. Mindnet: Acquiring and structuring semantic information from text. In *ACL-COLING'98*, pages 1098–1102.

Robert, Paul, Alain Rey, and J. Rey-Debove. 1993. *Dictionnaire alphabetique et analogique de la Langue Française*. Le Robert, Paris.

Roget, P. 1852. *Thesaurus of English Words and Phrases*. Longman, London.

Schvaneveldt, R., editor, 1989. *Pathfinder Associative Networks: studies in knowledge organization*. Ablex, Norwood, New Jersey, US.

Schwab, Didier and Mathieu Lafourcade. 2007. Modelling, detection and exploitation of lexical functions for analysis. *ECTI Transactions Journal on Computer and Information Technology*, 2(2):97–108.

Sierra, Gerardo. 2000. The onomasiological dictionary: a gap in lexicography. In *Proceedings of the Ninth Euralex International Congress*, pages 223–235, IMS, Universität Stuttgart.

Sinopalnikova, Anna and Pavel Smrz. 2006. Knowing a word vs. accessing a word: Wordnet and word association norms as interfaces to electronic dictionaries. In *Proceedings of the Third International WordNet Conference*, pages 265–272, Korea.

Summers, Della. 1993. *Language Activator: the world's first production dictionary*. Longman, London.

T'ong, Ting-Kü. 1862. *Ying ü tsap ts'ün (The Chinese and English Instructor)*. Canton.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*. Oxford University Press, Oxford.

Wilks, Yorick. 1977. Good and bad arguments about semantic primitives. *Communication and Cognition*, 10(3–4):181–221.

Zock, Michael and Slaven Bilac. 2004. Word lookup on the basis of associations : from an idea to a roadmap. In *Workshop on 'Enhancing and using electronic dictionaries'*, pages 29–35, Geneva. COLING.

Zock, Michael. 2006. Navigational aids, a critical factor for the success of electronic dictionaries. In Rapp, Reinhard, P. Sedlmeier, and G. Zunker-Rapp, editors, *Perspectives on Cognition: A Festschrift for Manfred Wettler*, pages 397–414. Pabst Science Publishers, Lengerich.

Zock, Michael. 2007. If you care to find what you are looking for, make an index: the case of lexical access. *ECTI, Transaction on Computer and Information Technology*, 2(2):71–80.

Accessing the ANW dictionary

Fons Moerdijk, Carole Tiberius, Jan Niestadt
Institute for Dutch Lexicology (INL)
Leiden
{moerdijk,tiberius,niestadt}@inl.nl

Abstract

This paper describes the functional design of an interface for an online scholarly dictionary of contemporary standard Dutch, the ANW. One of the main innovations of the ANW is a twofold meaning description: definitions are accompanied by 'semagrams'. In this paper we focus on the strategies that are available for accessing information in the dictionary and the role semagrams play in the dictionary practice.

1 Introduction

In this paper we discuss the functional design of an interface for a scholarly dictionary of contemporary standard Dutch which is currently being compiled at the institute for Dutch Lexicology in Leiden. The 'Algemeen Nederlands Woordenboek' (General Dutch Dictionary), further abbreviated as ANW, has been set up as an online dictionary from the start. Thus, the ANW is not a clone of an existing printed dictionary, but it truly represents a new generation of electronic dictionaries in the sector of academic and scientific lexicography. A similar dictionary project is undertaken for German at the Institut für Deutsche Sprache in Mannheim, i.e. *elexiko*[1].

The project runs from 2001 till 2019. We have currently finished the functional design of the interface and the first results will be published on the web in 2009.

[1] http://hypermedia.ids-mannheim.de/pls/elexiko/p4_start.portal

The structure of this paper is as follows. First we will provide some background information on the ANW dictionary and we will explain what a semagram is. Then we will discuss the range of search routes that are offered to the user to exploit the information in the dictionary and we will describe the role of the semagram. The ANW dictionary is aimed at the adult Dutch language user ranging from laymen to linguists and other language professionals.

2 The ANW dictionary

The ANW Dictionary is a comprehensive online scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium. Object of description is the general language. Thus words that are specific to a particular region, to a particular group of people or a particular subject field are not included. The dictionary focuses on written Dutch and covers the period from 1970 till 2018. The ANW dictionary is a corpus-based dictionary based on the ANW corpus, a balanced corpus of just over 100 million words, which was compiled specifically for the project at the Institute for Dutch Lexicology. The corpus was completed in 2005[2]. It consists of several subcorpora: a corpus of present-day literary texts, a corpus of neologisms, a corpus of domain dependent texts and a corpus of newspaper texts. The dictionary will contain approximately 80.000 headwords with a complete description and about 250.000 smaller entries.

The ANW is a very informative dictionary. Its abstract entry structure is composed of hundreds of elements and subelements. The reason for this is that special attention is paid to words in context (combinations, collocations, idioms, prov-

[2] For neologisms new corpus material continues to be gathered.

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 18–24
Manchester, August 2008

erbs) and to relations with other words (lexical relations like synonymy, antonymy, hyperonymy, hyponymy), to semantic relations (metaphor, metonymy, generalisation, specialisation) and to morphological patterns, the word structure of derivations and compounds. One of its main innovations is a twofold meaning description: definitions are accompanied by 'semagrams'. As semagrams play a central role in the dictionary (for understanding and production), we provide a short introduction below.

3 The semagram

A semagram is the representation of knowledge associated with a word in a frame of 'slots' and 'fillers'. 'Slots' are conceptual structure elements which characterise the properties and relations of the semantic class of a word meaning. On the basis of these slots specific data is stored ('fillers') for the word in question. In ANW jargon the abstract structure schema is called a 'type template', whereas semagram refers to such a 'type template' populated with concrete word data. Each semantic class has its own predefined type template with its own slots. For instance, the type template for the class of animals contains the slots PARTS, BEHAVIOUR, COLOUR, SOUND, BUILD, SIZE, PLACE, APPEARANCE, FUNCTION and SEX, whereas the type template for beverages has slots for INGREDIENT, PREPARATION, TASTE, COLOUR, TRANSPARANCY, USE, SMELL, SOURCE, FUNCTION, TEMPERATURE and COMPOSITION. So far we have concentrated on semagrams for nouns, those for verbs and adjectives will be different. Below we give an example of a semagram for a member of the animal class, i.e. *koe (cow)* (translated into English for illustration) in its meaning as a 'bovine':

COW

UPPER CATEGORY:	is an animal # animal; mammal; ruminant
CATEGORY:	is a bovine (animal) # bovine; ruminant
SOUND:	moows/lows, makes a sound that we imitate with a low, long-drawn 'boo' # moo; low; boo
COLOUR:	is often black and white spotted, but also brown and white spotted, black, brown or white # black and white; brown and white; red and white; spotted; black; blackspotted; white; brown; rusty brown
SIZE:	is big # big
BUILD:	is big-boned, bony, large-limbed in build # big-boned, bony, large-limbed
PARTS:	has an udder, horns and four stomachs: paunch, reticulum, third stomach, proper stomach # udder; horns: paunch; rumen; honeycomb bag; reticulum; third stomach; omasum; proper stomach; abomasum
FUNCTION:	produces milk and (being slaughtered) meat # milk; flesh; meat; beef; milk production; meat production
PLACE:	is kept on a farm; is in the field and in the winter in the byre # farm; farmhouse; field; pasture; meadow; byre; cow-house; shippon; stable
AGE:	is adult, has calved # adult; calved
PROPERTY:	is useful and tame; is considered as a friendly, lazy, slow, dumb, curious, social animal # tame; domesticated; friendly; lazy; slow; dumb; curious; social
SEX:	is female # female
BEHAVIOUR:	grazes and ruminates # graze; ruminate; chew the cud
TREATMENT:	is milked every day; is slaughtered # milk; slaughter
PRODUCT:	produces milk and meat # milk; meat
VALUE:	is useful # useful

Example 1. Semagram for *koe (cow)*

At present the data in the slots is completed manually by the lexicographers based on information in the ANW corpus, reference works (such as dictionaries and encyclopaedia) and their language and world knowledge. Not all slots in the type template have to be completed in all cases. Only those for which there is a value are shown in the above example. As can be seen from the semagram above, the lexicographers give the characterisation of the slots in terms of short statements about the headword. Such sentences are particularly well suited to get an impression of the meaning starting from the word form, i.e. for 'semasi-

ological' queries. To facilitate the retrieval for queries from content or parts of the content to the matching words, the 'onomasiological queries', those sentences are complemented, after a '#' character (a hash), with one or more keywords and possibly some synonyms or other relevant words. The data after the hash will not be visible to the dictionary user on the screen though and will only be used in searches by the computer to enhance retrieval.

A detailed description of the semagram, including its origin, motivation and the development of the type templates and their slots, can be found in Moerdijk (2008). In this paper we focus on the strategies that are available for accessing information in the dictionary and we discuss the role of the semagrams in this.

4 Accessing the dictionary

As was hinted at in the previous section, semagrams provide an increase and improvement in search and query facilities. This is particularly the case for queries guiding the user from content to form. For instance, a user who cannot think of e.g. the word *apiarist* can find this word through separate content elements (e.g. 'bees', 'keep') that he does know and can use for a search. However, with semagrams it is not only possible to go from content to the appropriate word. It is also possible to retrieve a set of words on the basis of one or more content features. Thus a user can retrieve all names for female animals in Dutch on the basis of a query combining the field CATEGORY with the value 'animal', and a field SEX with the value 'female'. In our online dictionary we wish to make all these possibilities available to the user.

Five search options are distinguished:

a) word → meaning, i.e. search for information about a word;
b) meaning → word, i.e. search for a word starting from its meaning;
c) features → words, i.e. search for words with one or more common features;
d) search for example sentences;
e) search for other dictionary information.

We believe that by presenting the search option this way (rather than using the traditional dichotomy between simple search (a) and advanced search (b, c, d, e)), users have a better overview of what they can actually search for and will be more enticed to explore the various options. Semagrams play a role in the first three search options.

4.1 Word → Meaning

This is the traditional search which allows the user to search for information about a word or phrase in the dictionary. As this is the basic search option, it is offered to the user in a central place on every page of the interface. Some form of fuzzy matching will be incorporated to take care of typing errors and incomplete input.

The ANW contains a wealth of information. To represent this to the user, we use a variation of the two-panel selector model (Tidwell 2005), where two panes are shown next to each other on the screen. (Figure 1)

The left pane contains a tree structure showing all the elements available for the lemma in question in the ANW. These tree structures look like (and work as) Windows Explorer tree structures. Advantage is that users know immediately how to deal with them. Thus the elements are hierarchically structured and can be opened and closed like in Windows Explorer. The meaning structure (the numbered elements in Figure 1) of the lemma remains visible at all times. This way the user keeps an overview and can select the information he likes to see on the right-hand-side of the screen. This is shown for the semagram of the first meaning of *koe* (cow) in Figure 1. The elements are presented in the same order as in the translated semagram in Example 1.[3]

On the article screen, the semagram is presented together with the definition. Its function is to provide, in a systemized, explicit and consistent way, more semantic and encyclopedic information than can be given in the definition. For the lemma *koe* (cow), for instance, it gives the user information on sound, colour and parts, which is not present in the definition.

At the bottom left of the screen, the user is given a direct link to all idioms, proverbs, example sentences and combinations for the lemma *koe* (cow).

[3] Note that the layout is still subject to change during the graphical design of the interface.

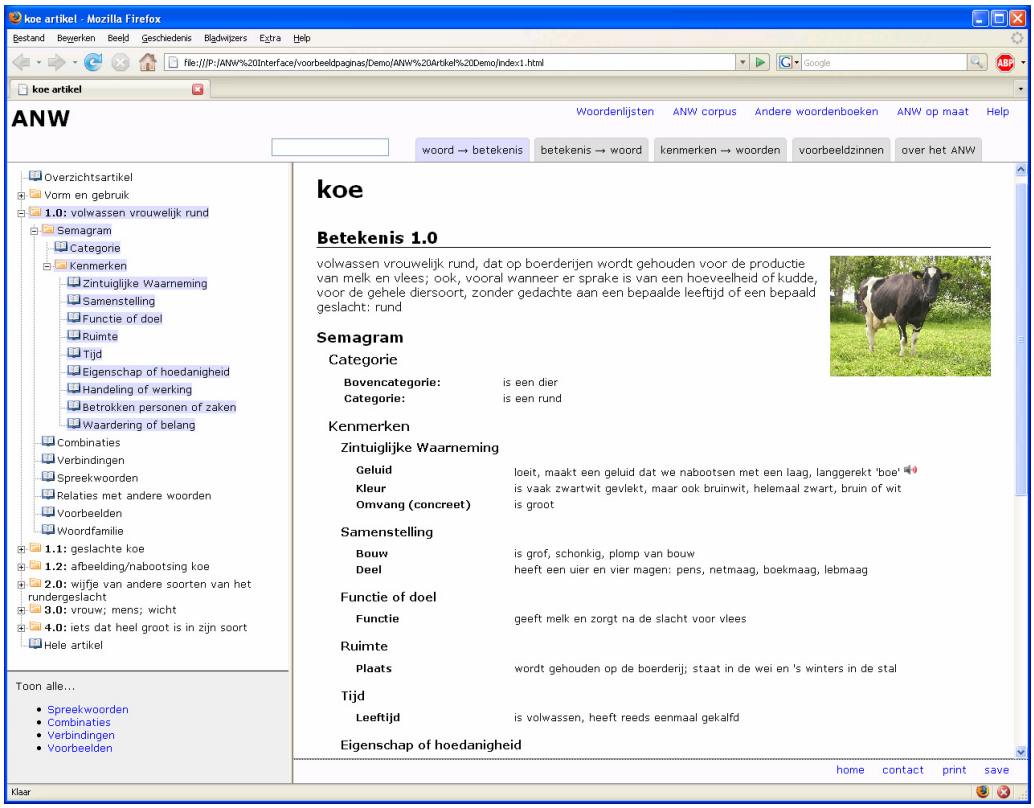


Figure 1 Article screen

4.2 Meaning → Word

By this we mean the onomasiological search where the user is looking for a word that he has forgotten or where he wants to know whether there is a word for a certain concept or not. For instance, a user may want to know whether there is a Dutch equivalent for the English *cradle snatcher* (i.e. one who weds, or is enamoured of, a much younger person (OED)).

Onomasiological searches in electronic dictionaries derived from printed dictionaries have not been very successful so far, mostly because such searches are primarily based on definitions. Going from a definition to a word can only succeed if the words of the user coincide (more or less) with the words in the definition, which is seldom the case (Moerdijk 2002).

As also pointed out by Sierra (2000) the ideal onomasiological search must allow writers to input the concept to be searched for through the ideas they may have, using words in any order. The system must be so constructed that it accepts a wide range of words which it then analyses in order to point the user to the word that most closely approaches the concept he had in mind when he started the search.

Recent work in computational linguistics has therefore looked at the possibility of using associative networks (Zock & Bilac 2004) or a combination of definitions and a resource such as WordNet (El-Kahlout & Oflazer 2004).

It is obvious that the information in the semagrams plays an essential role in the success of onomasiological queries in the ANW. However, rather than just accepting a wide range of words as input, we believe that the format in which the input query is obtained can also help to increase the success rate.

Therefore, we offer the user two alternatives for onomasiological queries. First, the user can search by giving a definition, a description, a paraphrase or by summing up synonyms or other words that he can associate with the word he is looking for. This input will be subject to some linguistic analysis including stemming and removal of stop words. Second, there is a guided search based on the semagram. The user is asked to choose a category (the semantic class or subclass) from a menu (is it a thing, a person, an animal, a vehicle, etc.?). This is a subset of the total number of semantic classes that are distinguished in the ANW. Once the user has selected a category,

the feature slots of the type template for that category appear on the screen and the user is asked to fill in the value(s) that spring to mind. Again we do not present the full list of feature slots of the type template of that particular semantic class, but rather a dozen or so (which have been automatically deduced on the basis of completed semagrams), as we do not want to put off the user with excessively long lists which he needs to complete before he gets an answer. We illustrate this with an example for animals.

Assume the user is looking for the name of a particular breed of dogs, e.g. *borzoi* (*barzoi* in Dutch), but cannot remember the word. In order to find the answer, he selects the category 'animal' from the menu. He is then presented with a list of features that are characteristic for animals (Figure 2). He completes the most prominent ones for the animal he is thinking of e.g. BEHAVIOUR: quiet, intelligent and independent; SOUND: barks; CLASS: greyhound; PLACE: Russia; SIZE: large; BUILD: strong and graceful; APPEARANCE: long-haired; MOVEMENT: sprinter.

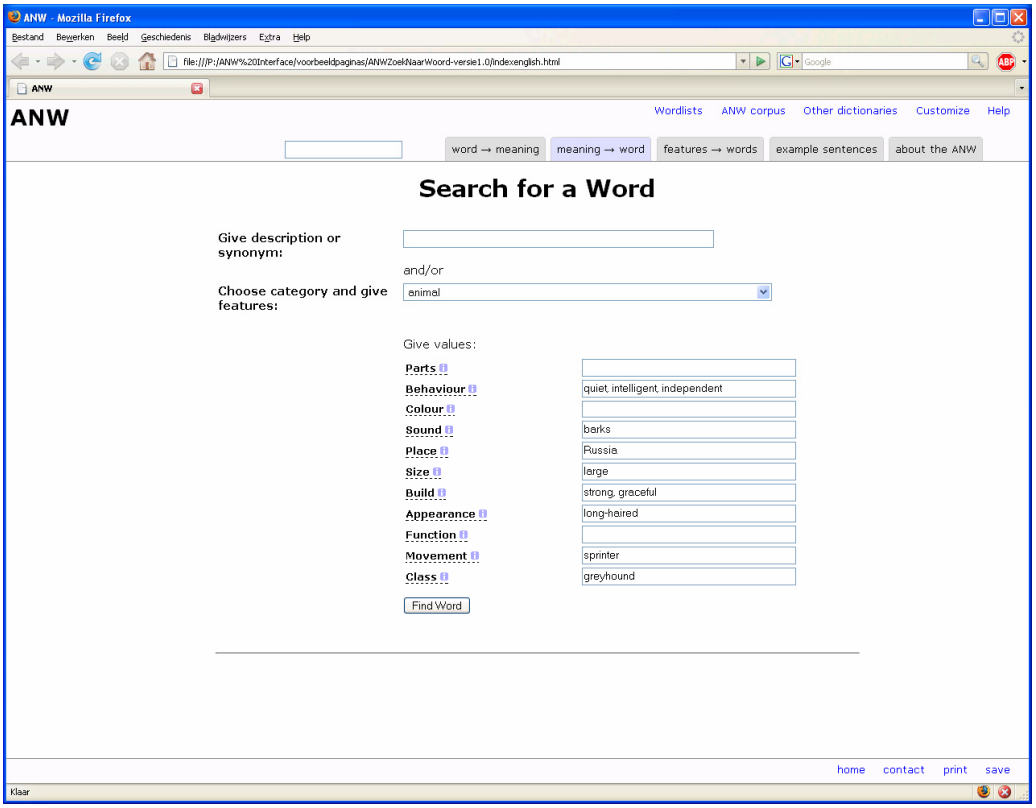


Figure 2 Screenshot Meaning → Word

The input from the user is then compared to the data in the dictionary database (semagrams, definitions, lexical relations and 'contextants'[4]). Now the words behind the hashes are also involved in the retrieval process and the matching cases (in the best scenario just one!) are shown. It is not necessary that the feature-value combinations match exactly one-to-one. For instance, in our example, one of the values given for BEHAVIOUR, i.e. intelligent, matches the value for PROPERTY in the semagram for *barzoi* (borzoi).

The results are then presented in a list, ordered by relevance. Each result is accompanied by a 'mini definition'[5] such that the user can immediately see which word (sense) he is looking for.

[4] We define 'contextants' as words which do not occur in direct combination with the headword, but do occur in a wider context and are semantically relevant for the headword. This is a separate information category in the microstructure of the ANW.

[5] A shortened version of the definition.

4.3 Features → words

This option is particularly relevant for linguists and other language professionals. It enables them to gather words that share one or more identical features within the main dimensions of the ANW, i.e. orthography, pronunciation, morphology, pragmatics, meaning, combinatorics, idioms, etymology. The semagram is of course active in searches in the semantic domain. Its role is to some extent comparable to its role in the search for a word, going from content to form, but users can now search for all the words that belong to a certain semantic class, for all the words that share one or more particular features, or for all the words sharing both class and certain features, instead of searching for a particular word to express a concept. Here the user is presented the full list of feature slots that occur in one or more of the predefined type templates. This means a total of nearly 200 features can be searched for.

To assist the user in finding his way through this forest of criteria, they are presented in a structured way much like the tree structure which is used for navigation on the article screen. We illustrate this with an example query in Figure 3. The user starts from an empty query screen. He is asked to select criteria from the tree structure on the left. By default, the user searches for words, but he can also search for proverbs or idioms which will result in a different feature tree as only a subset of the criteria that can occur in a query for words apply to idioms and proverbs. In our example the user wants to find all words for long-haired animals (semagram) which consist of two syllables and have alternating stress (orthography and pronunciation). Again *barzoi* (borzoi) will be among the results.

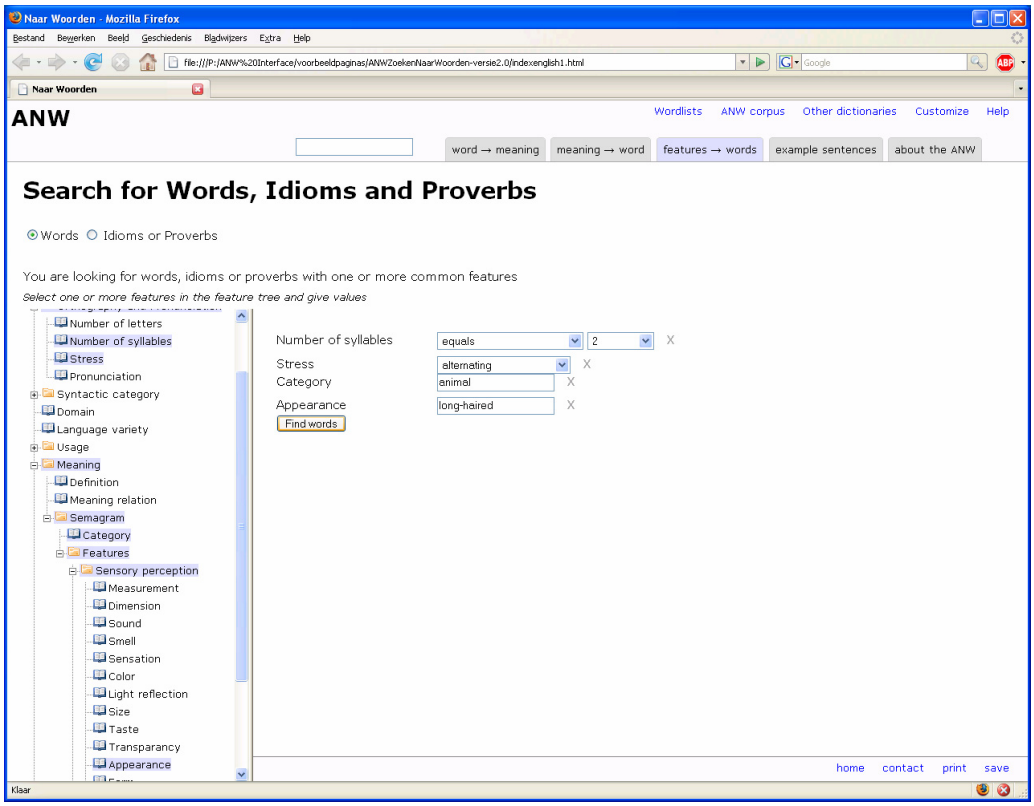


Figure 3 Screenshot Features → Words

This search option can also be used to resolve the so-called tip-of-the-tongue problems where a user is looking for a word which he cannot access in his memory, but where he does know, for instance, what the word looks like (e.g. its beginning, number of syllables) and its part of speech. For example, a user who is unsure whether the particular breed of dogs he is looking for should be called *barzoi* or *borzoi* in Dutch, can find the answer by specifying that the form ends in *–zoi*, the word consist of two syllables, that it is a noun

and that it refers to a breed of dogs (animal category) with long hairs (appearance).

Obviously users will be offered the possibility to save their queries in a kind of 'search templates' to avoid having to reconstruct the same query over and over again.

4.4 Search for examples

This option allows the user to search for example sentences based on a set of 5 criteria, i.e. word(s), author, source, domain and date. For instance, a user could search for all example sentences with the words *koe* (cow) and *schaap* (sheep) in the period from 2000 – 2002 (date). No combo boxes are used for author and source. Although we do not reckon that the user knows which authors and sources are cited in the dictionary, the lists would be excessively long and we assume that the user will only use these criteria in a search to see which other examples are available from a particular author or source he has retrieved in a previous query. Users will also be offered the possibility to link through to more examples of the same source or author by clicking on a particular source or author on the results page.

4.5 Search for information about the ANW

The final search option groups primarily dictionary specific queries and queries of an administrative nature, much like a Frequently Asked Questions page. Here the user will find queries about frequency such as how many lemmas are dedicated to lexicalised phrases? How many names are there in the dictionary? How many nouns? How many semagrams? How many Flemish words? It also comprises questions such as what kind of dictionary is the ANW? How big is the ANW corpus? Which images are included?

5 Conclusion

In this paper we have discussed the functional design of an interface for an electronic dictionary of Dutch, the ANW. We have focused on the access strategies that are offered and the role semagrams play in this. We have shown that semagrams provide an increase in search and query facilities. One the one hand, they lead to a much richer and more consistent semantic description in 'semasiological' queries. On the other hand, they are particularly well-suited to support 'onomasiological' queries by offering a structured way to find words through separate content elements.

References

El-Kahlout, İlknur Durgar & Kemal Oflazer. 2004. Use of Wordnet for Retrieving Words from their Meanings. In *Proceedings of the Global Wordnet Conference (GWC2004)*. 118-123.

Moerdijk, Fons. 2002. *Het woord als doelwit*. Amsterdam, Amsterdam University Press.

Moerdijk, Fons. 2008. Frames and semagrams; Meaning description in the General Dutch Dictionary. In *Proceedings of the Thirteenth Euralex International Congress, EURALEX 2008*. Barcelona.

Sierra, Gerardo. 2000. The onomasiological dictionary: a gap in lexicography. In *Proceedings of the Ninth Euralex International Congress, EURALEX 2000* I, 223-235. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Tidwell Jenifer. 2005. *Designing Interfaces*. O'Reilly.

Zock, Michael & Slaven Bilac. 2004. Word lookup on the basis of association: from an idea to a roadmap. In *Proceedings of the Workshop on Enhancing and using electronic dictionaries, COLING 2004*. 29-35.

ProPOSEL: a human-oriented prosody and PoS English lexicon for machine learning and NLP

Claire Brierley
School of Games Computing & Creative Technologies
University of Bolton
Deane Road
BOLTON
BL3 5AB
cb5@bolton.ac.uk

Eric Atwell
School of Computing
University of Leeds
LEEDS
LS2 9JT
eric@comp.leeds.ac.uk

Abstract

ProPOSEL is a prosody and PoS English lexicon, purpose-built to integrate and leverage domain knowledge from several well-established lexical resources for machine learning and NLP applications. The lexicon of 104049 separate entries is in accessible text file format, is human and machine-readable, and is intended for open source distribution with the Natural Language ToolKit. It is therefore supported by Python software tools which transform ProPOSEL into a Python dictionary or associative array of linguistic concepts mapped to compound lookup keys. Users can also conduct searches on a subset of the lexicon and access entries by word class, phonetic transcription, syllable count and lexical stress pattern. ProPOSEL caters for a range of different cognitive aspects of the lexicon©.

1 Introduction

ProPOSEL (Brierley and Atwell, 2008) is a prosody and part-of-speech (PoS) English lexicon which merges information from respected electronic dictionaries and databases, and which is purpose-built for linkage with corpora; for populating tokenized corpus text with a priori linguistic knowledge; for machine learning tasks which involve the prosodic-syntactic chunking of text; and for open source distribution with NLTK - the Python-based Natural Language Toolkit (Bird *et al*, 2007a).

A pronunciation lexicon like ProPOSEL is an integral part of the front-end natural language processing (NLP) module in a generic text-to-speech (TTS) synthesis system and constitutes a natural way of giving such a system phonetic, prosodic and morpho-syntactic insights into input text. For English, three such resources, originally developed for automatic speech recognition (ASR) and listing words and their phonetic transcriptions, are widely used: CELEX-2 (Baayen *et al*, 1996); PRONLEX (Kingsbury *et al*, 1997); and CMU, the Carnegie-Mellon Pronouncing Dictionary (Carnegie-Mellon University, 1998). The latter is used in Edinburgh's state of the art Festival speech synthesis system (Black *et al*, 1999) and is included as one of the datasets in NLTK.

The starting point for ProPOSEL is CUVPlus[1] (Pedler, 2002), a computer-usable and human-readable dictionary of inflected forms which uniquely identifies word class for each entry via C5 PoS tags, the syntactic annotation scheme used in the BNC or British National Corpus (Burnard, 2000). CUVPlus is an updated version of CUV2 (Mitton, 1992), an electronic dictionary in accessible text file format which in turn derives from the traditional paper-based Oxford Advanced Learner's Dictionary of Current English (Hornby, 1974).

Recently, lexica for thirteen world languages, including US-English, have been created via the European-funded LC-STAR project (Hartinkainen *et al*, 2003) to address the shortage of language resources in the form of wide coverage lexica with detailed morpho-syntactic information that meet the needs of ASR, TTS and speech-to-speech translation (SST) applications. The incorporation of C5 PoS-tags in CUVPlus provides this kind of detail and

[1] http://ota.ahds.ac.uk/textinfo/2469.html

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 25–31
Manchester, August 2008

distinguishes this lexicon from other paper-based and electronic English dictionaries, including CELEX-2, PRONLEX and CMU; it also facilitates linkage with machine-readable corpora like the BNC.

However, CUVPlus entries compact PoS variants for a given word form into a single field as in the following example where *burning* is classified as an adjective, a present participle and a noun in Table 1:

```
burning|AJ0:14,VVG:14,NN1:2|
```

Table 1: Sample from CUVPlus record structure showing PoS variants for the word form *burning*

An early operation during ProPOSEL build was therefore to introduce one-to-one mappings of word form to word class, as defined by C5, to facilitate their use as compound lookup keys when the lexicon is transformed into a Python dictionary or associative array (§4).

2 ProPOSEL: a repository of phonetic, syntactic and prosodic concepts

The current revised version of ProPOSEL[2] is a text file of 104049 separate entries, each comprising 15 pipe-separated fields arranged as follows:

(1) word form; (2) BNC C5 tag; (3) CUV2 capitalisation flag alert for word forms which start with a capital letter; (4) SAM-PA phonetic transcription; (5) CUV2 tag and frequency rating; (6) C5 tag and BNC frequency rating; (7) syllable count; (8) lexical stress pattern; (9) Penn Treebank tag(s); (10) default content or function word tag; (11) LOB tag(s); (12) C7 tag(s); (13) DISC stressed and syllabified phonetic transcription; (14) stressed and unstressed values mapped to DISC syllable transcriptions; (15) consonant-vowel [CV] pattern.

```
sunniest|AJS|0|'sVnIIst|Os%|AJS:0|3|100|JJS
|C|JJT|JJT|'sV-nI-Ist|'sV:1 nI:0 Ist:0|
[CV][CV][VCC]
```

Table 2: Example entry from ProPOSEL textfile

Table 2 shows an example entry showing all fields; subsequent illustrative examples include only a subset of fields. For an explanation of fields 3 to 7, the reader is referred to Pedler (2002) and Mitton (1992). A full account of ProPOSEL build is planned for a subsequent paper, where phonology fields in source lexica (CUVPlus, CELEX-2 and CMU) and new phonology fields in the prosody and PoS English lexicon will be discussed in detail. The rationale for fields displaying syllable count, lexical stress pattern and CFP status is summarised here in section 3.

Four major PoS tagging schemes have been included in ProPOSEL to facilitate linkage with several widely used speech corpora: C5 (field 2) with the BNC as mentioned; Penn Treebank (field 9) with Treebank-3 (Marcus *et al*, 1999); LOB (Johansson *et al,* 1986) (field 11) with MARSEC (Roach *et al*, 1993); and C7 (field 12) with the 2 million-word BNC Sampler Corpus. The lookup mechanism described in section 4 where a match is sought between (token, tag) tuples in incoming corpus text and ProPOSEL's compound dictionary keys, also in the form of (token, tag) tuples, is possible for all four syntactic annotation schemes represented in the lexicon.

3 Accessing the lexicon through sound, syllables and rhythmic structure

One field of particular significance for ProPOSEL's target application of prosodic phrase break prediction (§3) is field (8) for lexical stress patterns, symbolic representations of the rhythmic structure of word forms via a string of numbers. Thus the pattern for the word form *,objec'tivity* - with secondary stress on the first syllable and primary stress on the third syllable - is `20100`. For some homographs, this lexical stress pattern can fluctuate depending on part-of-speech category and meaning. The wordform *present* is a case in point, as demonstrated by fields 1, 2, 4, 7, 8 and 10 for all its entries in ProPOSEL shown in Table 3:

```
present | AJ0 | 'preznt | 2 | 10 | C |
present | NN1 | 'preznt | 2 | 10 | C |
present | VVI | prI'zent | 2 | 01 | C |
present | VVB | prI'zent | 2 | 01 | C |
```

Table 3: Rhythmic structure for the homograph *present* is inverted when it functions as a verb

Two well established phonetic transcription schemes are also represented in ProPOSEL: the original SAM-PA transcriptions in field 4 and DISC stressed and syllabified transcriptions in fields 13 and 14 which, unlike SAM-PA and the International Phonetic Alphabet (IPA), use a single character to represent dipthongs: `/p8R/` for *pair*, for example.

[2] April 2008

Phonology fields in ProPOSEL constitute a range of access routes for users. As an illustration, a search for like candidates to the verb *obliterate* might focus on structure and sound: verbs of 4 syllables (fields 2 and 7), with vowel reduction on the *first* syllable (fields 8 or 14), and primary stress on the *second* syllable (again, a choice of fields as users may wish to use the SAM-PA phonetic transcriptions). This filter retrieves sixty-seven candidates - most but not all of them end in /eIt/ - and includes one oddity among the examples in Table 4. Further examples of live filtered searches are presented in section 5.

```
('affiliate', "@'fIlIeIt")
('caparison', "k@'p&rIs@n")
('corroborate', "k@'r0b@reIt")
('manipulate', "m@'nIpjUleIt")
('originate', "@'rIdZIneIt")
('perpetuate', "p@'petSUeIt")
('subordinate', "s@'bOdIneIt")
('vociferate', "v@'sIf@reIt")
```

Table 4: Sample of 8 candidate verbs retrieved which share requested phonological features with the template verb: *obliterate*

4 ProPOSEL: domain knowledge for machine learning

As previously stated, the rationale for ProPOSEL was to integrate information from different dictionaries and databases into one lexicon, customised for language engineering tasks which involve the prosodic-syntactic chunking of text. One such task is automated phrase break prediction: the classification of junctures (whitespaces) between words in the input text as either breaks (the minority class) or non-breaks. Typically, the machine learner is trained on PoS-tagged and boundary-annotated text - the speech corpus or *gold standard* - and then tested on an unseen reference dataset, *minus* the boundary tags, from the same corpus. Finally, it is evaluated by counting how many of the original boundary locations have been recaptured or *predicted* by the model.

Phrase break classifiers have been trained on additional text-based features besides PoS tags. The CFP status of a token - is it a *content* word (e.g. nouns or adjectives) or *function* word (e.g. prepositions or articles) or *punctuation* mark? - has proved to be a very effective attribute in both deterministic and probabilistic models (Liberman and Church, 1992; Busser *et al*, 2001) and therefore, a default content-word/function-word tag is assigned to each entry in ProPOSEL in field (10). It is anticipated that further research will suggest modifications to this default status when the CFP attribute interacts with other text-based features.

Syllable counts - field (7) in ProPOSEL - have already been used successfully in phrase break models for English (Atterer and Klein, 2002). However, they assume uniformity in terms of duration of syllables whereas we know that in connected speech, an indefinite number of unstressed syllables are packed into the gap between one *stress pulse* (Mortimer, 1985) and another, English being a *stress-timed* language. A lexical stress pattern, where syllables are weighted 0, 1 or 2, has therefore been included in fields (8) and (14) for entries in ProPOSEL because of its potential as a classificatory feature in the machine learning task of phrase break prediction.

The thematic programme for PASCAL [3] in 2008 focuses on approaches to supplementing raw training data (e.g. the speech corpus) with a priori knowledge (e.g. the lexicon) to improve performance in machine learning. The prosody-syntax interface is notoriously complex. Planned research into the phrase break prediction task will attempt to incorporate a dictionary-derived feature such as lexical stress (field 8 in ProPOSEL) into a data-driven model to explore this interface more fully.

5 Implementing ProPOSEL as a Python dictionary

The Python programming language has a dictionary mapping object with entries in the form of (key, value) pairs. Each key must be unique and immutable (e.g. a string or tuple), while the values can be any type (e.g. a list). This data structure can be exploited by transforming ProPOSEL into a *live* Python dictionary, where the recommended access strategy is via compound keys (word form and C5 PoS tag) which uniquely identify each lexical entry. Thus, using a sample of 4 entries to represent ProPOSEL and version 0.8 of NLTK, we can use the code in Listing 1 (§next page) to convert this mini lexicon into the new formalism. The Python dictionary method returns an as yet unsorted dictionary, where the data structure itself is represented by

[3] Pattern Analysis, Statistical Modelling and Computational Learning research network
http://www.cs.man.ac.uk/~neill/thematic08.html

squigs { } and where ***key : value*** pairs are separated by a colon. Table 5 displays the output from Listing 1 (below), demonstrating how multiple values representing a series of linguistic observations on syllable count, lexical stress pattern and content/function word status have now been mapped to compound keys (cf. Bird *et al*, 2007b, chapter 6; Martelli *et al*, 2005 pp. 173-5).

```
{
('cascaded', 'VVD') : ['3', '010', 'C'],
('cascaded', 'VVN') : ['3', '010', 'C'],
('cascading', 'VVG') : ['3', '010', 'C'],
('cascading', 'AJ0') : ['3', '010', 'C']
}
```

Table 5: Output from Listing 1

```
from nltk.book import * # In NLTK 0.9, the import statement would be: import nltk, re, pprint
lexicon = """
cascaded|VVD|0|k&'skeIdId|Ic%,Id%|VVD:1|3|010|VBD|C|VVD|VBD
cascaded|VVN|0|k&'skeIdId|Ic%,Id%|VVN:0|3|010|VBN|C|VVN,VVNK|VBN
cascading|VVG|0|k&'skeIdIN|Ib%|VVG:1|3|010|VBG|C|VVG,VVGK|VBG
cascading|AJ0|0|k&'skeIdIN|Ib%|AJ0:0|3|010|JJ|C|JJ,JK|JJ,JJB,JNP
"""
lexicon = [line.split('|') for line in list(tokenize.line(lexicon))]
lexKeys = [(index[0], index[1]) for index in lexicon]
lexValues = [[index[6], index[7], index[9]] for index in lexicon]
proPOSEL = dict(zip(lexKeys, lexValues))
```

Listing 1: Code snippet using Python list comprehensions and built-ins to transform the prosody-PoS English Lexicon into an associative array

For linkage with corpora and for annotating a corpus with the prior knowledge of phonology contained in ProPOSEL, a match is sought between incoming corpus text in the familiar (token, tag) format and the dictionary keys (§Table 5). Thus intersection enables corpus text to accumulate additional values which have the potential to become features for machine learning tasks. This lookup mechanism is relatively straightforward for corpora tagged with C5, the basic tagset used in the BNC. For corpora tagged with alternative schemes (i.e. Penn, LOB, and C7), incoming tokens and tags can either be matched against word forms and PoS tokens in the corresponding tagset field in the lexicon, or C5 tags can be appended to each item in the input text such that lookup can proceed in the normal way.

6 Filtered searches and having fun with ProPOSEL

ProPOSEL will be supported by a tutorial, offering a range of Python software compatible with NLTK, to enable users to prepare the text file for NLP; to implement ProPOSEL as a Python dictionary; to cross-reference linguistic data in the lexicon and corpus text; and to customise searches via multiple criteria.

The previous section demonstrated how fine-grained grammatical distinctions in the PoS tag field(s) in ProPOSEL are integral to linkage with corpora. It also demonstrated how an electronic dictionary in the form of a simple text file can be reconceived and reconstituted as a computational data structure known as an associative memory or array. When ProPOSEL is thus transformed, filtered searches can be performed on the text itself.

Brierley and Atwell (ibid.) present automatic corpus annotations achieved via intersection of two parallel iterables: ProPOSEL's keys and a LOB-tagged corpus extract (this is a short extract of 153 tokens just for demonstration) which also carries equivalent C5 tags generated from the lexicon. A successful match between C5 tags in both lists results in a corpus sequence object where word tokens and syntactic annotations have now been complemented with prosodic information from selected fields in ProPOSEL, as in Table 6:

```
[["aren't",   'BER+XNOT',   'VBB+XX0',
  ['1', '1', 'CF', "'#nt:1"]]
```

Table 6: Entry index of length 3, with word token mapped to LOB and C5 tags plus syllable count, lexical stress pattern, CFP status and syllable-stress mapping

The corpus sequence object can now be queried. Suppose, for instance, we wanted to find all bi-syllabic prepositions and particles in

this extract. By specifying part-of-speech and syllable count, we unearth just one candidate matching our search criteria, as shown in Table 7:

```
['between', 'IN', 'PRP', ['2', '01',
'F', "bI:0 'twin:1"]]
```

Table 7: There is one candidate in the 153 word extract which meets the condition: PoS equals preposition or particle and syllable count is 2

It is not always necessary to transform ProPOSEL into a Python dictionary, however. Users can also read in the lexicon textfile, apply Python's splitlines() method to process the text as a list of lines, and then apply the split() method, with the *pipe* field separator as argument, to tokenize each field. Listing 2 presents this much more succinctly:

```
lexicon = open('filepath', 'rU').read()
lexicon = lexicon.splitlines()
lexicon = [line.split('|') for line in
lexicon]
```

Listing2: Reading in ProPOSEL as a nested structure

Users can then perform a search on a defined subset of the lexicon. For example, users may wish to retrieve all entries with seven syllables from the lexicon. As well as returning items like: *industrialisation*, *operating-theatre*, and *radioactivity*, Listing 3 discovers the rather intriguing *sir roger de coverley*!

```
for index in lexicon:
   if index[6] == '7': # look in the subset
     print index[0] # return word form(s)
```

Listing 3: Searching a subset of the lexicon

Another illustration would be finding words which rhyme. If we wanted to find all the words which rhyme with *corpus* in the lexicon, we could search field (4), for example, the SAM-PA phonetic transcriptions, for similar strings to `/'kOp@s/`. One way of doing this would be to compile a regular expression, substituting the metacharacter **.** for the '*c*' in *corpus* and then seek a match in the SAM-PA field[4]. We might also look for minimal pairs, replacing the phoneme `/s/` with the phoneme `/z/` as in `/'.Op@z/`. Retaining the apostrophe as diacritic for primary stress before the wildcard here imitates the lexical stress pattern for *corpus* and is part of the rhyme. It transpires there is only one candidate which rhymes with *corpus* in the lexicon and two half rhymes. Listing4 gives us *porpoise* `/'pOp@s/` and then *paupers* `/'pOp@z/` and *torpors* `/'tOp@z/`.

```
p1 = re.compile("'.Op@s")
p2 = re.compile("'.Op@z")
sampa = [index[3] for index in lexicon]
rhymes1 = p1.findall(' '.join(sampa))
rhymes2 = p2.findall(' '.join(sampa))
```

Listing 4: Using regular expressions to retrieve bi-syllabic words with primary stress on the first syllable that rhyme with *corpus*

7 Cognitive Aspects of the Lexicon

ProPOSEL and associated access tools are presented to the CogALex workshop audience to illustrate our approach to enhancing the structure, indexing and entry points of electronic dictionaries. As the Call for Papers notes, "Access strategies vary with the task (text understanding vs. text production) and the knowledge available at the moment of consultation (word, concept, sound). Unlike readers who look for meanings, writers start from them, searching for the corresponding words. While paper dictionaries are static, permitting only limited strategies for accessing information, their electronic counterparts promise dynamic, proactive search via multiple criteria (meaning, sound, related word) and via diverse access routes. … The goal of this workshop is to perform the groundwork for the next generation of electronic dictionaries, that is, to study the possibility of integrating the different resources …" ProPOSEL integrates a range of different resources, and enables a variety of access strategies, with consultation based on various combinations of partial syntactic and prosodic knowledge of the target words. It addresses the main themes of the workshop:

[4] Note that Python lists start at index 0, hence in Listing 4, the SAM-PA field is at position [3] in the inner list of tokenized list fields for each entry.

7.1 Conceptual input of a dictionary user

Human users of electronic dictionaries can start from partial concepts or patterns when they are generating a message or looking for a (target) word. Other papers in the workshop focus on semantic cues, such as conceptual primitives, semantically related words, some type of partial definition, something like *synsets* etc; but speakers/writers may also be searching for a word which matches syntactic, phonetic or prosodic partial patterns, for example seeking a matching rhythm or rhyme.

7.2 Access, navigation and search strategies

The Call for Papers notes that "we would like to be able to access entries by word form but also by meaning and sounds (syllables) …Even if input is given in an incomplete, imprecise or degraded form." Meaning is clearly the main focus of many lexicography researchers, but access by sound, rhythm, prosody, and also syntactic similarity may also prove useful complementary strategies for some users.

7.3 Indexing words and organizing the lexicon

Another key issue for discussion in the Call for Papers is robust yet flexible organization of lexical resources: "Indexing must robustly allow for multiple ways of navigation and access… ". By building on and integrating with Python and the NLTK Natural Language Tool Kit, ProPOSEL can be accessed by other NLP tools or via the standard Python interface for direct browsing and search. ProPOSEL is also a potential exemplar for lexical entry standardization. Many lexicographers focus on standardization of semantics or definitions, but standardization of syntactic, phonetic and prosodic information is also an issue. Our pragmatic approach is to integrate lexical entries from a range of resources into a standardized Python dictionary format.

7.4 NLP Applications

We initially developed ProPOSEL in the context of research in linking lexical, syntactic and prosodic markup in English corpus text, and specifically as a resource for prosodic phrase break prediction (Brierley and Atwell, 2007a,b,c). The software developed within the NLTK architecture has been able to utilize existing NLTK tools for PoS-tagging, phrase-chunking and partial parsing; in turn, other researchers in these fields may want to use the syntactic information in ProPOSEL in their future NLP applications, particularly in research which attempts to compare or map between alternative tagsets or labeling systems, eg (Nancarrow and Atwell 2007), Atwell and Roberts 2006), (Atwell et al 2000), (Teufel 1995).

8 Conclusions

The English lexicon presented in this paper, - a revised version to that reported in (Brierley and Atwell, 2008), - is an assembly of domain knowledge of phonology and syntax from several widely used lexical resources. Linkage with corpora is facilitated by the inclusion of four variant PoS tagging schemes in the lexicon and by re-thinking and reconstituting the lexicon as a Python dictionary or associative array. A successful match between (token, tag) pairings in input text and new linguistic annotations mapped to ProPOSEL's compound keys will in turn embed a priori knowledge from the lexicon in data-driven models derived from a corpus and enhance performance in machine learning. The lexicon is also *human-oriented* (de Schryver, 2003). ProPOSEL's software tools are compatible with NLTK and enable users to define and search a subset of the lexicon and access entries by word class, phonetic transcription, syllable count and rhythmic structure. ProPOSEL was initially developed as a language engineering resource for use in our own research, but in the process of development we have also addressed several more general issues relating to cognitive aspects of the lexicon: the partial patterns in the mind of a dictionary user; the need for access and search by sound, rhythm, prosody, and also syntactic similarity; robust and standardised organization of lexical entries from different sources; and ease of integration into NLP applications.

References

Atterer M., and E. Klein. 2002. Integrating Linguistic and Performance-Based Constraints for Assigning Phrase Breaks. In *Proceedings of Coling 2002*:29-35.

Atwell, E., G. Demetriou, J. Hughes, A. Schriffin, C. Souter, S. Wilcock. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal*, vol. 24, pp. 7-23.

Atwell, E. and A. Roberts. 2006. Combinatory hybrid elementary analysis of text. In Kurimo, M, Creutz, M & Lagus, K (editors) *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*. Venice.

Baayen, R. H., R. Piepenbrock, and L. Gulikers 1996. *CELEX2* Linguistic Data Consortium, Philadelphia

Bird, S., E. Loper, and E. Klein 2007a. *NLTK-lite 0.8 beta* [June 2007] Available online from: http://nltk.sourceforge.net/index.php/Main_Page (accessed: 21/06/07).

Bird, S., E. Klein, and E. Loper 2007b. *Natural Language Processing* Available online from: http://nltk.sourceforge.net/index.php/Book (accessed: 21/09/07).

Black A.W., P. Taylor, and R. Caley. 1999. *The Festival Speech Synthesis System: System Documentation Festival version 1.4* Available online from: http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html (Accessed: 07/03/08)

Brierley, C. and E. Atwell. 2007a. Corpus-based evaluation of prosodic phrase break prediction in: *Proceedings of Corpus Linguistics 2007*, Birmingham University.

Brierley, C. and E. Atwell. 2007b. An approach for detecting prosodic phrase boundaries in spoken English. *ACM Crossroads journal*, vol. 14.1.

Brierley, C. and E. Atwell. 2007c. Prosodic phrase break prediction: problems in the evaluation of models against a gold standard. *Traitement Automatique des Langues,* vol. 48.1.

Brierley, C. and E. Atwell. 2008 ProPOSEL: a Prosody and POS English Lexicon for Language Engineering. In *Proceedings of LREC'08 Language Resources and Evaluation Conference,* Marrakech, Morocco. May 2008.

Burnard, L. (ed.) 2000. *Reference Guide for the British National Corpus (World Edition)* Available online from: http://www.natcorp.ox.ac.uk/docs/userManual/ (accessed: 20/05/07).

Busser, B. W. Daelemans, and A. van den Bosch 2001. Predicting phrase breaks with memory-based learning. *4th ISCA Tutorial and Research Workshop on Speech Synthesis*. Edinburgh, 2001.

Carnegie-Mellon University 1998. *The CMU Pronouncing Dictionary (v. 0.6)* Available online from: http://www.speech.cs.cmu.edu/cgi-bin/cmudict (accessed: 21/06/07).

Hartinkainen, E., G. Maltese, A. Moreno, S. Shammass, U. Ziegenhain 2003. Large Lexica for Speech-to-Speech Translation: frm specification to creation. *EUROSPEECH-2003*:1529-1532.

Hornby, A.S. 1974. *Oxford Advanced Learner's Dictionary of Current English* (third edition) Oxford: Oxford University Press

Johansson, S; Atwell, E S; Garside, R; Leech, G. 1986. *The Tagged LOB Corpus - User Manual,* 160pp, Bergen, Norwegian Computing Centre for the Humanities.

Kingsbury, P., S. Strassel, C. McLemore, and R. MacIntyre 1997. *CALLHOME American English Lexicon (PRONLEX)* Linguistic Data Consortium, Philadelphia

Liberman, M.Y., and K.W. Church 1992. Text Analysis and Word Pronunciation in Text-to-Speech Synthesis. In Furui, S., and Sondhi, M.M., (eds.) *Advances in Speech Signal Processing* New York, Marcel Dekker, Inc.

Marcus, M.P., B. Santorini, M.A. Marcinkiewicz, and A. Taylor 1999. *TREEBANK-3* Linguistic Data Consortium, Philadelphia

Martelli, A., A. Martelli Ravenscroft, and D. Ascher 2005. *Python Cookbook* (second edition) Sebastopol: O'Reilly Media, Inc.

Mitton, R. 1992. *A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English* Available online and accessed (22/03/08) from: http://comp.lin.msu.edu/stabler-notes/1850/ascii_0710-2.txt

Mortimer, C. 1985. *Elements of Pronunciation.* Cambridge: Cambridge University Press

Nancarrow, O. and E. Atwell. 2007.A comparative study of the tagging of adverbs in modern English corpora *Proceedings of Corpus Linguistics 2007.* Birmingham University.

Pedler, J. 2002. *CUVPlus* [Electronic Resource] Oxford Text Archive Available online from: http://ota.ahds.ac.uk/textinfo/2469.html (accessed: 21/06/07)

Roach P., G. Knowles, T. Varadi and S.C. Arnfield. 1993. Marsec: A machine-readable spoken English corpus *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 47—53.

Schryver, G. M. de. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 2003 16(2):143-199

Teufel, S.. 1995. A support tool for tagset mapping. Proceedings *of SIGDAT 1995. Workshop in cooperation with EACL 95*, Dublin

Natural Language Searching in Onomasiological Dictionaries

Gerardo Sierra
Instituto de Ingeniería
Universidad Nacional Autónoma de México
gsierram@ii.unam.mx

Abstract

When consulting a dictionary, people can find the meaning of a word via the definition, which usually contains the relevant information to fulfil their requirement. Lexicographers produce dictionaries and their work consists in presenting information essential for grasping the meaning of words. However, when people need to find a word it is likely that they do not obtain the information they are looking for. There is a gap between dictionary definitions and the information being available in peoples' mind. This paper attempts to present the conceptualisation people engage in, in order to arrive at a word from its meaning. The insights of an experiment conducted show us the differences between the knowledge available in peoples' minds and in dictionary definitions.

1 Introduction

Many lexicographers recognise users need dictionaries to look for a word that has escaped their memory although they remember the concept. From a semantic point of view, Baldinger (1980) takes user needs into account and thus distinguishes dictionaries that serve as aids in encoding from those that help with decoding. The best known dictionaries of this type allow users to find the meaning of a word they already know. Such dictionaries are semasiological: they associate meanings with expressions/words, i.e. within entries we move from word to meaning. The second kind of dictionary helps those users who have an idea to convey and want to find a word to designate it. Such dictionaries are onomasiological: they connect names to concepts, i.e. within entries we move from meaning or concept to name or word.

Sierra (2000) confirmed the well known observation that the organisation of the world varies from author to author, by contrasting some recognized onomasiological dictionaries, such as Roget's Thesaurus of English Words and Phrases (1852), Bernstein's Reverse Dictionary (1975), and WordNet (Miller et al., 2008).

In order to build a system that maps natural language descriptions of concepts to the terms corresponding to those concepts, Sierra and McNaught (2000) outlined the design of an Onomasiological Search System. They described the principles of the system, whereas the architecture and its components are presented as part of the design. This also includes an idealised user interface, with a discussion of the organisation of the probable terms and additional information that can help the user to identify precisely the term he is looking for.

As cognitive issues for the design of such system, this paper attempts to present the conceptualisation people engage in, in order to arrive at a word from its meaning. In this sense, it breaks the traditional lexicographic assumption that one should utilise a semasiological approach to provide formal representations to describe the meaning of a word. In contrast, in the onomasiological approach, the user can formulate a concept in several ways and use a variety of words in order to find a particular word.

Our starting point is to understand what conceptualisation is (section 2), and the process of designation (section 3). To validate our approach in practice, section 4 presents the results of an experiment, which is compared with other stud-

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 32–38
Manchester, August 2008

ies on conceptual analysis. Finally, the conclusions are stated.

2 Conceptualisation

The concept is a mental representation of an object which is formed in the mind of individuals through a process of abstraction. We call this process of constructing an internal representation of external things a *conceptualisation*. Conceptualisation is a mental activity of grouping the data of common properties according to external factors, and then concepts are internalised and form part of each individual's knowledge.

2.1 Properties

The terms *property*, *characteristic* or *feature* have been used for the knowledge necessary to describe and classify a concept. The identification of properties is crucial to concept analysis since it helps to define concepts and identify their interrelationships.

According to Sager (1990) some properties are necessary and sufficient to distinguish a concept from any other, and these properties reflect the essential characteristics of a concept. Conversely, other properties are inessential, merely observable in an individual thing, so that they are accidental, may change with time and may not even be necessarily true in a scientific sense (Petöfi, 1982).

The dichotomy of these two opposing types of properties has been discussed from a psycholinguistic point of view. Aitchison (1994) does not consider that it is obvious that experts and ordinary people distinguish between essential and inessential characteristics. Despite the fact that experts might be able to specify the true nature of things, they sometimes provide information which is irrelevant to the mental lexicon. Conversely, ordinary people disagree and change their minds.

As the comparative analysis in section 4 shows, the essential characteristics are not necessarily present in the mental lexicon of a person; each one describes different properties. Nevertheless, even a description of the inessential characteristics, given together, provides enough information to identify the term (Wierzbicka, 1985).

2.2 Social conceptualisation

People acquire knowledge about things on the basis of cultural, geographical and social factors. The environment conditions the conceptualisation of reality and the use of language. In order to communicate effectively, people will try to use language in a similar way to that of the collective view of the community, in agreement with the social norm. In fact, because of the social norm, there is an idealised knowledge structure which makes it possible to use the same names for the same things. In the contrary case, when the designation of a concept is outside that norm, people assume that the individual's knowledge is wrong. However, as we will see in the final analysis, we must accept that an individual's knowledge cannot cover the whole knowledge of the community norm.

2.3 Individual conceptualisation

Reality goes beyond the perception of individuals. Our knowledge about reality has increased throughout human history. No one – human, computer or even the biggest library– possesses the whole knowledge about reality.

The knowledge structure of things varies from one person to the other, so that their description of concepts will be different. As Fugman (1993) states, since the number of properties is virtually unlimited, people concentrate on those characteristics which appear essential, according to their personal or professional view. As an example, he points out the different essential properties for the concept "benzene" given by a physicist, a biologist, an engineer, a fire-fighter and a chemist.

Even the same person can demonstrate different conceptualisations of simple things, such as "dog" or "apple", depending on the contextual situation. For example, a dog seen in different domains, such as a conference, a zoology lecture, a road or a house, may be described as canine, mammal, animal or pet.

3 Designation

The process of designation is the opposite of signification. Signification is the identification of the meaning of a word, and the result of finding a meaning is a definition. Designation is the identification of a term for a concept and the result is a word.

To retrieve a term, one can use a terminological definition, which provides the information necessary to identify and differentiate a concept within a system of concepts, so that it sometimes comprises encyclopaedic information, not usually necessary in a lexicographic definition.

3.1 Properties

Just as a word may have many meanings (semasiological approach), a concept, which is described by a set of properties, may be designated by more than one word (onomasiological approach). Within the onomasiological approach, all the properties together provide the necessary and sufficient information to identify the concept. However, since the description of concepts in natural language does not incorporate all knowledge or ideas associated with each concept, it can happen that the projection of a concept, i.e., the query formulation of the user, will retrieve a set of terms. For example, the concept "strong winds", consisting of two properties, can retrieve, in the domain of weather terminology, a variety of terms, such as: "gale", "tornado", "hurricane", "typhoon", etc.

The concept a user has in mind when looking for a target word is expressed by a sentence. When a person hears this sentence, he translates each word into his own language and easily identifies the context. A person may understand the expression "that which determines air pressure" and get a mental representation of "that" for "thing" and then for "instrument"; or that the speaker might have said "measures" instead of "determines". From the context, at the same time, he may differentiate whether the word "air" refers to the atmosphere or the air of a tyre.

Equally, either the lack or change of any one property may result in the identification of a different concept. For example, take the following definition for "barometer":

- A device to measure air pressure.

Each of the four keywords yields a property. Then, we can change one property at a time and get a different concept. If we change

- "Device" to "unit", the result is the concepts "inch" or "millibar".
- "Measure" to "provide", the result is "air scoop".
- "Air" to "blood", the result is "sphygmomanometer".
- "Pressure" to "humidity", the result is the concept "hygrometer".

4 Comparative analysis on conceptualization

In order to verify some of the ideas presented above, an experiment was carried out with a small group of twenty undergraduate students. Although a small group is unrepresentative for any generalisation to be made from a statistical point of view, it has been sufficient for our purpose to demonstrate that the conceptualisation used by a random set of students is far from the definitions found in a dictionary.

From two sets of five words, each student was asked to take a set and write on a blank sheet of paper, similar to an onomasiological search, a concept, a definition or the ideas suggested to them by each word. After interchanging the sheets, the other students participating in the experiment wrote the word or words designating the concepts identified or written on the blank sheets by the previous student.

The sets of word given contained three general language words and two terms.

- Set A: water, squirrel, bench, euthanasia, hurricane.
- Set B: lemon, bucket, clothing, monopoly, barometer.

The general words were chosen because they permit us, as can be observed from the following sections, to compare the results with the words analysed by other researchers as well.

We will next introduce our definitions by comparing with four studies on conceptualisation.

4.1 Putnam

Putnam (1975) proposes the representation of the meaning of a word as a finite sequence of at least four properties:

The *syntactic markers*, which are the category-indicators used in a host of contexts to classify words.

The *semantic markers*, which are the most central properties, form part of a widely used classification system and very may be affected by any change in the knowledge about the thing.

A description of the features of the *stereotype*, which is a conventional idea of what the object looks like or acts like or is, regardless whether it is true or not for all the objects. For example, "yellow" is a stereotype of "gold", even when gold is white by nature.

A description of the *extension*, i.e., the set of things of which a term is true. The extension is determined socially depending upon the nature of the particular things, rather than on the concept of the individual speaker.

The first three properties belong to the individual competence of speakers. The extension

does not necessarily have to be known to every member of a linguistic community.

According to Petöfi (1982), the semantic markers and the stereotype may be compared with Ullman's concept of meaning. From the perspective of the lexicographic definition, they resemble genus and differentia, respectively.

The description of the meaning of "water", as a particular natural kind, following these components, is given in table 1.

Syntactic markers	**Semantic markers**	**Stereotype**	**Extension**
mass noun	natural kind	colourless	H_2O
concrete	liquid	transparent	(give or take
		tasteless	impurities)
		thirst-quenching	

Table 1. Properties for the natural kind "water"

In order to permit comparison of the definitions given in our experiment with his meaning of "water", the same four kinds of properties are used. Our definitions, as shown in Table 2, include the property "fluid", which easily can be classified as a semantic marker.

n.	***Concept***
1	It's a clear liquid that you get from a tap
2	The colourless transparent liquid occurring on rivers
3	A clear, neutral liquid that surrounds us everywhere
4	Liquid, clear, drinkable – constituents are hydrogen and oxygen
5	Liquid, clear, H_2O
6	Liquid form, scientific term H_2O
7	Liquid, freezes at 0°C
8	Liquid, clear, boils at 100°C, freezes at 0°C
9	Fluid, clear, tasteless, colourless
10	Wash with it; drink it; used for dilution; H_2O; found in springs, rivers, lakes, seas, oceans

Table 2. Conceptualisation of water

The properties referring to the boiling and freezing points of water, given in definitions 7 and 8 in our experiment, may be considered as part of the concept's extension, since these properties depend upon the nature of the water.

Therefore, definitions one to nine include the semantic marker "liquid", beside the particular features of water, the stereotypes, and/or the description by extension. The definition ten, which does not include the semantic marker, describes water by extension.

4.2 Wiegand

Wiegand (1984) tries to identify the properties of a definition by means of a *scale of usability* obtained statistically from a questionnaire to 100 students. He suggests 21 properties and asks the students to judge which of them describe a lemon. Each property is evaluated in three categories according to the sum of ticks it received as good, not so good and not good (Table 3).

GOOD	**NOT SO GOOD**	**NOT GOOD**
oval	yellow	tapers at both ends
juicy pulp	thick rind	oblong
sour pulp	citrus fruit	thin rind
yellow rind	green rind	used to make pectin
fruit of the lemon tree		pulp containing approx. 3.5-8% citric acid
		pulp rich in vitamin C
		variable protuberant tip
		pulp rich in vitamins
		many uses in cooking
		used to make drinks
		used to make citric acid

Table 3. Properties of lemon using a scale of usability

Even although a test with ten students is not a representative sample from which one can generalise the *scale of usability* of the properties of a concept, our experiment, as shows in table 4, allows us to challenge the values identified by Wiegand.

n.	***Concept***
1	It's a yellow fruit, like limes. Citrus. Used in cooking for sharpness
2	A yellow citrus fruit. Sour tasting. Often used as an accompaniment to drinks
3	a yellow citrus fruit with a bitter taste often sliced and put in drinks
4	It's a citrus fruit, yellow, used with sugar on pancakes
5	It's a yellow citrus fruit. Tastes bitter. Oval shaped
6	A yellow sour fruit

n.	*Concept*
7	A yellow citrus fruit
8	Yellow, citrus, fruit
9	Citrus fruit which is yellow
10	Yellow citrus fruit

Table 4. Conceptualisation of lemon

As observed in table 5, there is no match between the values for the seven properties extracted by Wiegand and our own experiment from the definitions.

Property	**Our experiment**	**Wiegand**
yellow	good	not so good
citrus	good	not so good
oval	not good	good
sour pulp	not good	good
many uses in cooking and drinks	not good	not good
variable protuberant tip	not good	not good
similar to limes	not good	---

Table 5. Comparative analysis of the properties of lemon

The reasons why these values differ are not obvious. Probably this comparison means statistical methods from a group of students are not reliable to assess the properties of concepts.

4.3 Ayto

In order to define the meaning of words, Ayto (1983) adapts the componential analysis introduced by Pottier to semantic fields. He also identifies the semantic features that characterise various sorts of seats, but analyses these characteristics to compose an analytical definition. The definition of a word is determined by the semantic features that differentiate it from other words rather than by the sum of the individual characteristics. The genus for each word in the semantic field is "seat", as it presents the only common characteristic for the rest of the set.

The differentia is determined by comparing the other characteristics and checking those which are different. The characteristics are: For several people, not upholstered, for outdoors, functional.

Then the definition for "bench" is, for example, "a seat for two or more people that has a back, is typically used outdoors, and may be fixed in position".

For a comparative analysis, it is possible to find the semantic features of the definitions in our experiment (Table 6) and try to match them with those given by Ayto.

n.	*Concept*
1	You can sit on it in the street or a park and they are made of wood
2	A long hard seat for several persons on which the players on a sport team sit
3	An object for sitting on, usually long which can seat many people
4	Sit on it (a few people can) in parks, made of wood or iron
5	Object used for sitting on. Often found in public places such as parks and gardens. Used to seat 1 or more people at a time
6	Something you seat on, is longer than a chair, usually made of wood
7	Long platform for sitting on (fit many people on one)
8	Apparatus for sitting on, designed for more than one person, often found in parks
9	A kind of seat found in parks, made of wood
10	A type of chair

Table 6. Conceptualisation of bench

For this purpose, we should assume that:

1. For several people = longer than a chair.
2. Not upholstered = made of wood or iron, hard seat, platform.
3. Outdoors = street, park.
4. Functional = for a sports team.

Table 7 presents the four semantic features used in our experiment to define "bench".

	Char. 1	**Char. 2**	**Char. 3**	**Char. 4**
Bench 1		+	+	
Bench 2	+	+		+
Bench 3	+			
Bench 4	+	+	+	
Bench 5	+		+	
Bench 6	+	+		

Bench 7	+	+	
Bench 8	+		+
Bench 9		+	+
Bench 10			

Table 7. Semantic features of bench

In the light of this contrastive analysis, it is clear that each semantic feature (for example "outdoor") can be expressed by a set of equivalent alternatives (for example, "public places", "parks").

4.4 Wierzbicka

Wierzbicka (1985) considers that a good lexicographic definition must be exhaustive, i.e., providing all the properties of the concept. Her view of a definition differs from an encyclopaedic definition because the latter conveys knowledge about the object, while the lexicographic definition does not include specialised knowledge, unless it is part of the concept. Her demand for exhaustiveness is contrary to traditional semasiological lexicography, where, through a genus and the differentia, the definition provides the essential properties to identify a concept and distinguish it from others. However, when there is a full description, we may be sure that a user will retrieve the word in an onomasiological search.

Wierzbicka uses five general properties to reach a definition of animals, e.g. squirrels, namely: habitat, size, appearance, behaviour and relation to people. Table 8 presents an example of a description for each general property.

General property	**Description**
Habitat	They live in places where there are many trees.
Size	They are not too big for a person to be able to hold one easily in both hands.
Appearance	They have a big bushy tail. Their fur is reddish or greyish.
Behaviour	They collect and eat small hard things which grow on trees of certain kinds.
Relation to people	People think of them as nice and amusing little creatures.

Table 8. Examples of full description of "squirrel"

As observed in the definitions of our experiment (Table 9), the sum of properties in our definitions agrees with the description of the five kinds of properties of Wierzbicka, although she does not consider that squirrels build nests, as one of our definitions does.

n.	***Concept***
1	It's a little rodent and can be red or grey, it has a big bushy tail
2	A small rodent living in trees with a long bushy tail
3	A small rodent which lives in trees, collects nuts and has a bushy tail
4	Animal, grey/red, bushy tail, lives in trees, buries nuts
5	Small animal, lives in trees, eats acorns, has a bushy tail
6	Animal, bushy tail, eats nuts, builds nests in trees called dreys
7	Small funny animal with big, bushy tail, likes nuts, likes trees
8	Animal that lives in trees and collects acorns, has a long tail
9	A small-sized animal, habitat in trees
10	Small grey mammal, relative to the rodent, found in both countryside and town

Table 9 Conceptualisation of lemon

5 Conclusions

The distinction between semasiology and onomasiology permits us to consider a new perspective in lexicography. In the semasiological approach, the perspective is from the dictionary to the user. Dictionaries are a lexicographer's product and definitions provide the necessary and sufficient elements in order to know the meaning of a word.

Conversely, the onomasiological approach is from the user to the dictionary. The user should provide the concept, while the dictionary interprets that concept in order to find the most appropriate word. The user can formulate the concept by several methods and may use a variety of words that in a certain context are similar. According to the user's social, cultural and geographical background, the description of the concept may differ in multiple properties.

With regard to the preceding analysis, it is worthwhile to note that even the most complete description of a concept can lack "essential" properties from the point of view of a user. None

of the methods of componential analysis, even the most open ones, has been sufficient to foresee the properties used by a small set of students. That gap should be filled with the aid of a good onomasiological retrieval system.

This does not mean that it is unlikely that we shall be able to design a complete and efficient onomasiological dictionary. In our context, efficiency means that a dictionary has to satisfy the requirements of a particular kind of user, in a certain domain of a terminology with a specific background. Therefore, the design of an onomasiological dictionary must first foresee a multiplicity of properties for each concept and secondly the diversity of words that can be used to name them. Then, the task consists in the accurate interpretation of the description of the concept and providing the word or probable words the user is looking for.

The core of such onomasiological dictionary, as reported by Sierra and McNaught (2000), is the lexical knowledge base (LKB), which should provide all the necessary knowledge to be manipulated in order to enable onomasiological search. In principle, it must represent what a person knows about both concepts and their corresponding terms. Such LKB consists then of a set of terms, a set of definitions for each term, a set of keywords associated with the definitions and a set of lexical paradigms that group keywords with the same meaning. It not only includes the databases that constitute these sets of data, but the interrelationships among all the sets.

References

Aitchison, Jean. 1994. *Words in the mind: an introduction to the mental lexicon*. Blackwell Publishers, Oxford.

Ayto, John R. 1983. "On specifying meaning: semantic analysis and dictionary definitions". *Lexicography: principles and practice*. R.R.K. Hartmann (ed). Academic Press, London: 89-98.

Baldinger, Kurt. 1980. *Semantic theory: towards a modern semantics*. Basil Blackwell, Oxford.

Bernstein, Theodore M. 1975. *Bernstein's reverse dictionary*. Routledge & Kegan Paul, London.

Fugman, Robert. 1993. *Subject analysis and indexing: theoretical foundation and practical advice*. INDEKS Verlag, Frankfurt.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. "Introduction to WordNet: An on-line lexical database". *International Journal of Lexicography*, 3(4): 235-244.

Petöfi, János S. 1982. "Exploration in semantics: analysis and representation of concept systems". *The Cocta Conference*. F.W. Riggs (ed).

Putnam, Hilary. 1975. *Mind, language and reality: philosophical papers*, Volume 2. Cambridge University Press, New York.

Roget, Peter. 1852. *Thesaurus of English Words and Phrases*. Longman, London.

Sager, Juan C 1990. *A practical course in terminology processing*. John Benjamins, Amsterdam.

Sierra, Gerardo. 2000. "The onomasiological dictionary: a gap in lexicography". *Proceedings of the Ninth EuralexInternational Congress.* Stuttgart.

Sierra, Gerardo and John McNaught. 2000. "Design of an Onomasiological Search System: a Concept-Oriented Tool for Terminology". *Terminology* 6(1).

Wiegand, Herbert E. 1984. "On the structure and contents of a general theory of lexicography". *Proceedings of the International Conference on Lexicography*. M. Niemeyer, Tübingen, 13-30.

Wierzbicka, Anna 1985. *Lexicography and conceptual analysis*. Karoma Publishers.

First ideas of user-adapted views of lexicographic data exemplified on OWID and *elexiko*

Carolin Müller-Spitzer
Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
mueller-spitzer@ids-mannheim.de

Christine Möhrs
Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
moehrs@lexik.ids-mannheim.de

Abstract

This paper is a project report of the lexicographic Internet portal OWID, an Online Vocabulary Information System of German which is being built at the Institute of German Language in Mannheim (IDS). Overall, the contents of the portal and its technical approaches will be presented. The lexical database is structured in a granular way which allows to extend possible search options for lexicographers. Against the background of current research on using electronic dictionaries, the project OWID is also working on first ideas of user-adapted access and user-adapted views of the lexicographic data. Due to the fact that the portal OWID comprises dictionaries which are available online it is possible to change the design and functions of the website easily (in comparison to printed dictionaries). Ideas of implementing user-adapted views of the lexicographic data will be demonstrated by using an example taken from one of the dictionaries of the portal, namely *elexiko*.

1 Project report

The *Online-Wortschatz-Informationssystem Deutsch* (OWID; Online Vocabulary Information System of German), a project of the *Institut für Deutsche Sprache* (IDS; Institute of German Language) in Mannheim is a lexicographic Internet portal containing both, various electronic dictionary resources that are currently being compiled at the IDS on the one hand and external resources on the other hand which will be included additionally in the near future (cf. www.owid.de). Originally, the project had its roots based in the IDS project *elexiko*, a lexicographic enterprise, which develops a new corpus-based dictionary of contemporary German. It formed the basis of a lexicographic information portal for the IDS (cf. Klosa et al. 2006). The main emphasis of OWID is on the integration of different academic lexicographic resources with the focus on contemporary German. Presently, the following dictionaries are included in OWID:

- *elexiko*: This electronic dictionary consists of an index of about 300.000 short entries with information on spelling and syllabication, including information about inflection (from www.canoo.net). In the near future, further information (e.g. on word formation) and corpus samples will be added for all lexemes. Furthermore, *elexiko* comprises over 900 fully elaborated entries of headwords which are highly frequent in the underlying corpus. These contain extensive semantic-pragmatic descriptions of lexical items in actual language use. The dictionary is being extended continuously by further elaborated entries (cf. Klosa et al. 2006).
- *Neologismenwörterbuch* (Dictionary of Neologisms): This electronic dictionary describes about 800 new words and new meanings of established words in detail which emerged in the German vocabulary during the 1990s. This dictionary is also being upgraded constantly.
- *Wortverbindungen online* (Collocations Online): This resource of OWID publishes the research results of the project *Usuelle*

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 39–46
Manchester, August 2008

Wortverbindungen. These concern different fixed multiword combinations. Currently, 25 detailed entries for fixed multiword combinations and 100 shorter entries dealing with collocations are available to users.

- *Diskurswörterbuch 1945-55* (Discourse Dictionary 1945-55): This dictionary is a reference work resulting from a larger study of lexemes that establish the notional area of "guilt" in the early post-war era (1945-55), published in 2005.

In the near future, the "Handbuch Deutscher Kommunikationsverben" (Handbook of German Communication Verbs) with approximately 350 paradigms of communication verbs as well as the "VALBU – Valenzwörterbuch deutscher Verben" (Valency Dictionary of German Verbs) will be published in OWID.

It has always been an explicit goal of OWID not to present a random collection of unrelated dictionary resources but to build a network of interrelated lexicographic products. Therefore it was necessary to maintain the independence of each individual dictionary project while, at the same time, to ensure the integration of all the different data. Even though, the different lexicographic resources may appear to be very diverse at first glance, they share some of their data modelling features. Both, the common intergration and the individual independence of each project are reflected in the current online presentation of the portal. On the welcome page of OWID the user can choose which dictionary s/he wants to use. If s/he looks up a word in all dictionaries of the portal there is a coloured marker indicating the corresponding dictionary resource (black = *elexiko*, blue = Neologism, green = Discourse dictionary, red = Collocations). In addition, there are links and cross-references between the products (see for example the interrelation between the entry "Liebe macht blind" in the dictionary "Collocations Online" and the entries "Liebe" / "blind" in *elexiko*). This kind of interrelation will be expanded in the future.

Another goal is to provide a basis for user-adapted access to the lexicographic data. "It is one thing to be able to store ever more data, but another thing entirely to present just the data users want in response to a particular look-up" (de Schryver 2003: 178). Hence, the core of the project is the design of an innovative concept of data modelling and structuring.

2 Data Modelling

As emphasised before, the contents of the individual participating projects and their compiled lexicographic resources in OWID are independent of each other. However, it has been obvious from the very beginning that the value of OWID would be increased, if more common access structures for the different contents were to be developed and if the lexicographic data had been

OWID DTD-library				
modules for the whole OWID portal		allg-entities.dtd (DTD for general entities)	allg-elemente.dtd (DTD for general elements and attributes)	
modules for cross-dictionary object groups	ewl-objekte.dtd (DTD for objects of single-word-items)	mwl-objekte.dtd (DTD for objects of multi-word-items)	ewl_mwl-objekte.dtd (DTD for objects of single-word-and multi-word-tems)	ewl-grammatik.dtd (DTD for grammatical objects)
modules for object groups of specific dictionaries		elexiko-allgobj.dtd (DTD for general objects of *elexiko*)	neo-allgobj.dtd (DTD for general objects of the neologism-dictionary)	
Head-DTDs for each dictionary	elexiko-ewl.dtd (Head-DTD for *elexiko*)	neo-ewl.dtd (Head-DTD for single-word-items of the neologism-dictionary) neo-mwl.dtd (Head-DTD for multi-word-items of the neologism-dictionary)	mwl.dtd (Head-DTD for multi-word-items of the project "Usuelle Wortverbindungen")	zeitreflexion1945-55.dtd (Head-DTD for the discourse-dictionary 1945-55)

Table 1. OWID DTD-library

interlinked even more adequately. So on the one hand, in order to guarantee a basis for a common access structure to the all contents, consistent principles for modelling and structuring the contents were applied to all integrated products. On the other hand, OWID is also kept open for the possible integration of externally developed lexicographic resources, namely reference works that are written outside the IDS. However, externally compiled data has to be structured in accordance to the OWID modelling concept.

The approach chosen here not only guarantees to connect different lexicographic products under the management of OWID on the macro structure level – which means the level of the headwords – but also makes it possible to access the dictionaries on a more granular level. OWID attempts to harmonise modelling on the level of the content structure, that is, the level of the individual lexicographic information unit rather than organizing the different lexicographic processes independently.

OWID uses a single modelling process for all projects: For each individual resource, a specifically-tailored XML-DTD and XML-schema were developed respectively. Each individual information unit is granularly tagged in all entry structures, so that automatic access to each content unit is ensured. The dictionary entries are then written in an XML editor and stored in an Oracle database system. For presentation purposes, the XML data are transformed by an XSLT stylesheet to HTML (cf. Müller-Spitzer 2007).

A DTD library was created for OWID where specific DTDs contain all entities, elements, or attributes that are shared by all entry structures in order to provide a uniform structure for lexicographic information of the same type which is contained in the different dictionaries (cf. Tab. 1). The modelling shows which information is accessible across the different dictionaries (the results from the different dictionaries are marked in different colours). This type of data modelling – a singular specificially-tailored but explicitly synchronised modelling for diverse lexicographic resources – can be considered to be an innovative approach of a new kind, as Schlaps (2007) and Kunze / Lemnitzer (2007) have recently explained.

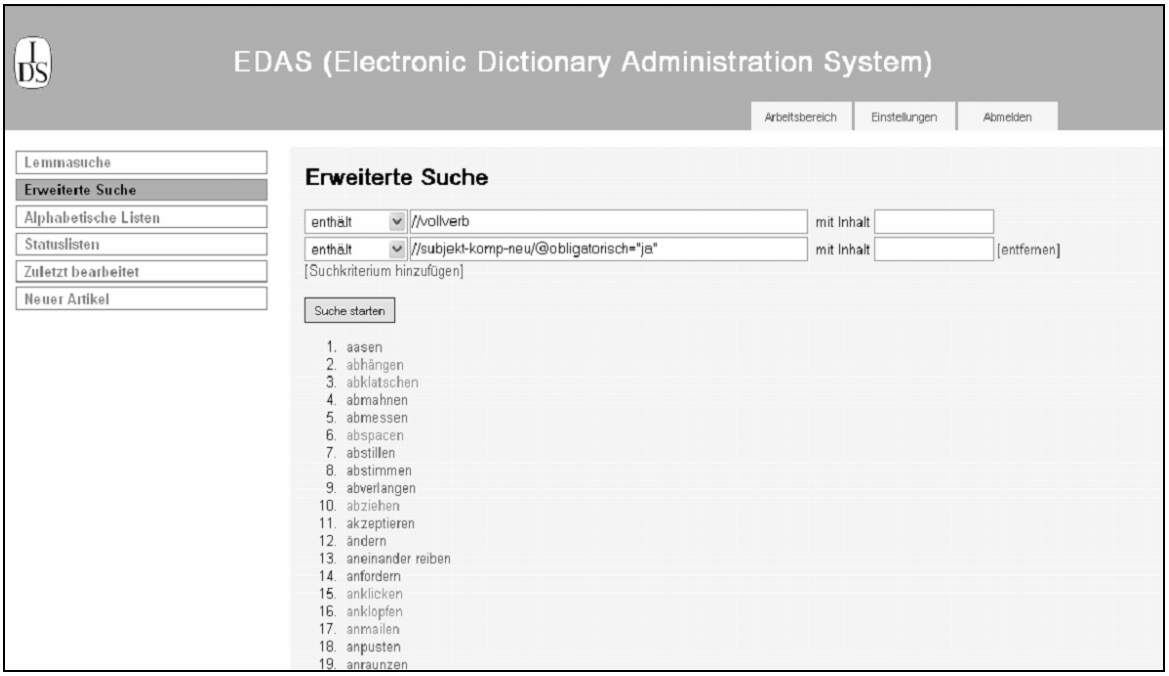


Figure 1: Advanced search options for lexicographers

We decided to use a specifically-tailored modelling because the XML-structure also serves as a model for compiling the lexicographic entries in the XML-Editor. What this means for lexicographers is that the more individually customised the XML-structure is, the less one needs an additional manual for comply with the entry structure. However, one could easily transform this structure into a specific standard such as LMF or TEI because the structure is very fine-grained. The following XML detail of the entry “emailen” from the Dictionary of Neologisms illustrating the tagging of information on valency gives an example for the overall granularity of tagging.

```
   <vb-valenz-neu>

   <satzbauplan>
   <satzbauplanA>jemand emailt (jemandem) (et-
was)</satzbauplanA>

   </satzbauplan>
   <satzbauplan>
   <satzbauplanA> jemand emailt (etwas) an je-
manden</satzbauplanA>
   </satzbauplan>

   <satzbauplan>
   <satzbauplanA>jemand      emailt,      dass
[...]</satzbauplanA>
   </satzbauplan>

   <vb-komplemente-neu>

   <subjekt-komp-neu obligatorisch="ja">
   <nom-nominalphrase-neu/>
   </subjekt-komp-neu>

   <objekt-komp-vb obligatorisch="nein">
   <dat-nominalphrase-vb/>
   </objekt-komp-vb>

   <objekt-komp-vb obligatorisch="nein">
   <akk-nominalphrase-vb/>
   <dass-satz-vb/>
   </objekt-komp-vb>

   <objekt-komp-vb obligatorisch="ja">
   <praepositionalphrase-vb        praeposi-
on="an"/>
   </objekt-komp-vb>

   </vb-komplemente-neu>
   </vb-valenz-neu>
```

Within our internal editorial system, lexicographers are able to use this structure for advanced searches (with XPath expressions). For example, one can search for all regular verbs `(//vollverb)` which have obligatory object complements `(//objekt-komp-vb/@obligatorisch="ja"` which are realised as a dative NP `(//dat-nominalphrase-vb)`. In this example, the search results are entries from *elexiko* as well as from the neologism-dictionary (cf. Fig. 1). We are planning to provide these extended search options also for users.[2]

Moreover, it would be possible to involve the user in the process of deciding which information should be presented on the website. As explained, every information unit in the dictionaries is encoded separately. Against this background, we can think of customizing the microstructure by the users themselves (in addition to the extended search for example in *elexiko*). So the user could select the type of information s/he wants to use individually. Fig. 2 shows what such a presentation could look like. At the top of the page, the user is able to select the type of information which s/he wants to see directly underneath. If s/he wants to change the options s/he can use the update button in order to modulate the desktop view. In this example, the two different senses of the entry 'Meer' are shown side by side with the chosen kind of information (here the definition together with typical uses of the headword). This kind of presentation enables the users to compare this information given for the two senses at one sight.

3 Research on using electronic dictionaries

Research on using dictionaries is a core field of study in lexicography (cf. Wang 2001 or Atkins 1998). Fortunately, in the last two decades, research on using printed dictionaries has attracted the attention of more researchers. Although Engelberg and Lemnitzer had noticed in 2001 that there are only little inquiries about influences on the users' behaviour in relation to innovations in the field of electronic lexicography (cf. Engelberg and Lemnitzer 2001), in the last few years research on electronic dictionaries has grown.

Such metalexicographic research plays a major role with regard to monitoring the dictionary user on the Internet – for example in the analysis of log-files. At the moment, there are not many research reports about the analysis of log-files. "Although the proposal to draw upon log files in order to improve dictionaries was already expressed in the mid-1980s […], and although numerous researchers have reiterated this idea in recent years […], very few reports have been published of *real-world* dictionaries actually marking use of this strategy" (de Schryver and Joffe 2004, 187). The studies and methods mentioned here are interesting for research on using electronic dictionaries especially because an electronic dictionary is a product which can be modulated and updated immediately. Log-files can show what the user has inserted into the search box and how the user has navigated (cf. de Schryver and Joffe 2004). However, good results are only seen with this method if the database of the dictionary is created with a flat structure. In the actual log- files we only see which word the user has typed in the search box. We can not easily detect in which way and how comfortly the user navigates through the entry or

[2] The development of the *Electronic Dictionary Administration System* (cf. Fig. 1) is a work of Roman Schneider, a researcher of the IDS.

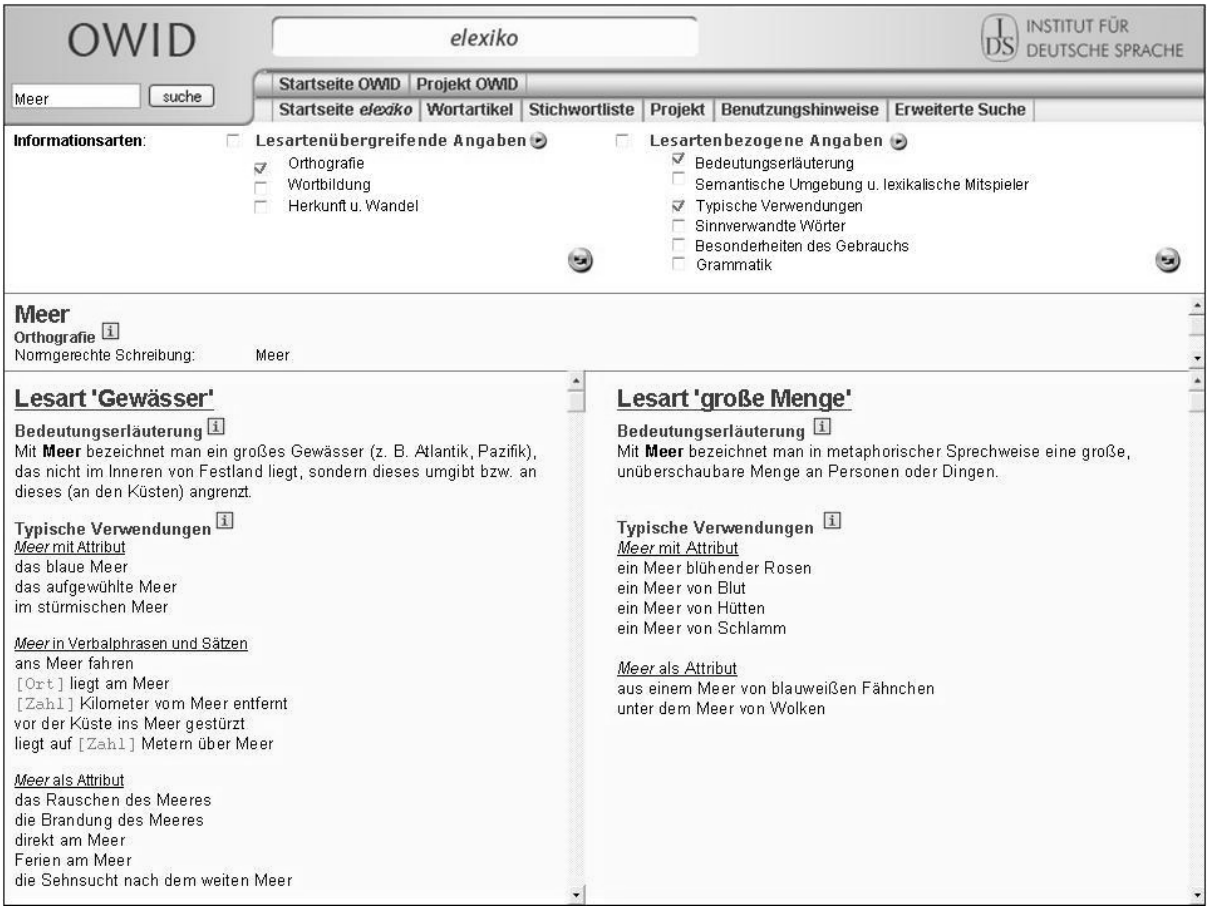


Figure 2. Online view of *elexiko* with an information display for customizing the microstructure dynamically

which information s/he has looked at more closely. However, this is exactly the type of information we are looking for. Therefore, other methods like standardised evaluation, interviews etc. also have to be taken into account. Analysing log-files can not substitute these methods alone.

OWID is also gradually putting user research into practice: Firstly, OWID has been making use of the analysis of log-files for some time. Secondly, a standardised online survey was conducted in the context of an MA thesis (cf. Scherer 2008). Finally, a short study based on interviews of OWID and in particular of *elexiko*, one of the dictionaries of the portal, was carried out.

Although currently the modelling is used mainly in the lexicographic process there is still a lot of room for further development of the abilities to present the structured information. The capability of data modelling in OWID should be visible for lexicographers as well as for users (cf. Müller-Spitzer 2007). Involving the user and his/her requirements in searching and navigating through OWID is the starting point for defining user-adapted views of the lexicographic data.

4 Defining user-adapted Views

As shown above, the lexicographic contents are structured granulary and strictly content-based. This technology allows to define user-adapted views of the lexicographic data. Printed dictionaries cannot offer this option. A printed dictionary is designed for a specific user type and for specific situations of use as a whole. In OWID, the data for electronic dictionaries is initially organised independently of its users. In a second step, lexicographic information can be used as the foundation of the definition of user-specific layers (e.g. based on the technology of XSLT-stylesheets) in order to filter relevant data for a specific situation of use “on demand”. Knowlegde on what users prototypically look for in printed dictionaries is established by numerous research works. For example someone who uses a dictionary to understand a text wants to get a short overview on the meaning of a word. If someone has to produce a text it is more helpful to get word information about correct spelling, grammar, typical uses, collocations or sense-related items. Furthermore lexicographers of

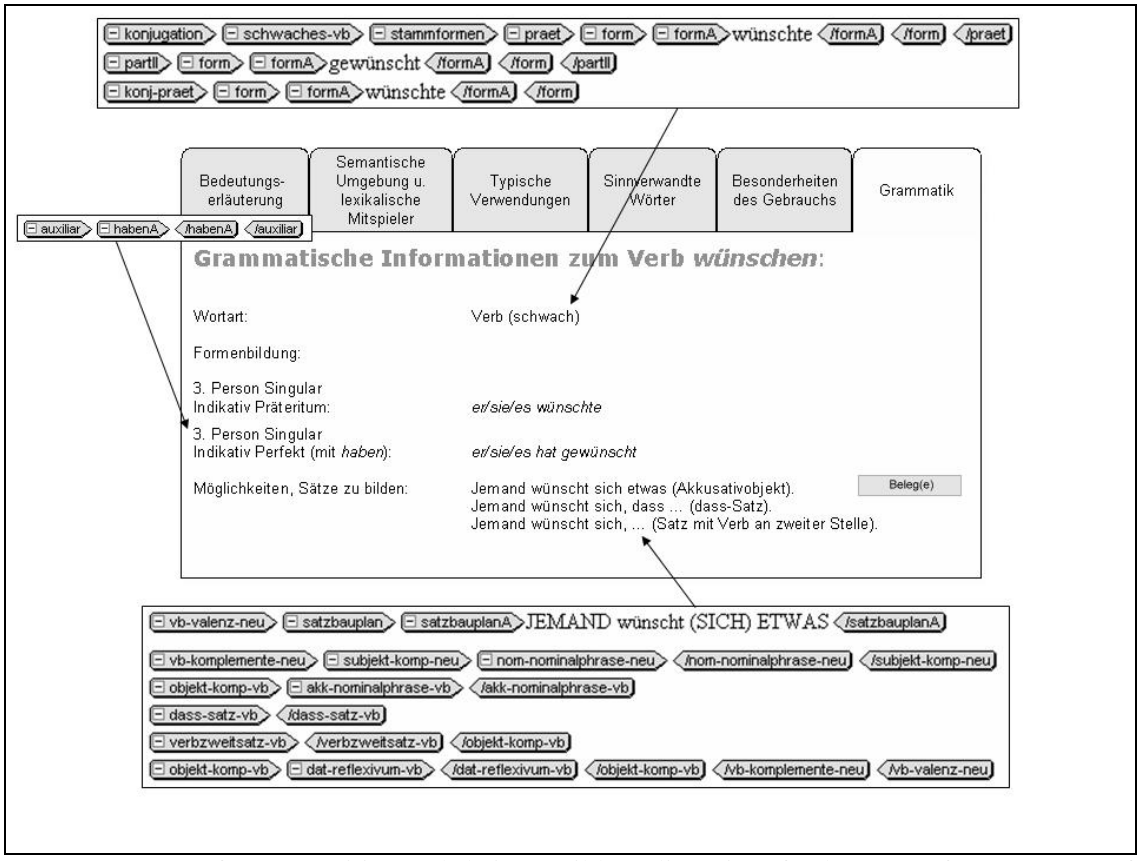


Figure 3. Extracts of XML-entities and their possible online view for learners of German as a foreign language (entry **wünschen**, part "Grammar" for the meaning 'ersehnen')

electronic dictionaries can go into detail about the demands of learners of German as foreign language (L2-Learners) resp. German native speakers. By taking this into consideration, one can think of developing different profiles for different user situations. According to a chosen profile the lexicographic information is then presented in a specialised way. This would be another form of a user-adapted view (besides customizing the microstructure dynamically as it is shown in Fig. 2). In *elexiko*, one of the dictionaries of OWID, the online view presents the lexicographic data in one standardised view. However, the technical conditions can also allow to show the same XML-data of an entry in different ways for different user groups. As an example one can see the part "Grammar" in *elexiko* in Figure 3 and 4 differing in comprehensiveness. Detailed information on inflection and word order are very important for L2-Learners. Therefore such information is presented more extensively in Fig. 3. In comparison native speakers know intuitionally the inflection of words or the realization of different sentence constructions. In Fig. 4 one can see a shortened presentation of grammatical information of the same XML-data. This example illustrates the general principle of defining different user-adapted views of one lexicographic data. It is important that the different user-adapted presentations of the part "Grammar" in *elexiko* or every other part of word information in *elexiko* can be realised without changing the data. The only change happens in the stylesheet. Other views completely different from the actually used stylesheet can be imagined easily. We will discuss further examples in the talk.

For a printed dictionary it is sufficient to define the types of information that shall be included for the intendend user. Questions of presentation are discussed on this basis and along the strong tradition for the layout of printed dictionaries. When compiling user-adapted views of a general lexicographic data for an electronic medium we have to consider:

How do users navigate in electronic dictionaries especially in a dictionary portal? How do they use the search options? Which form of nesting the specific word information is user friendly and when does clearness suffer? (Cf. Almind 2005) More specifically we need to ask: Should a user (i.e. while using a dictionary) create a profile at

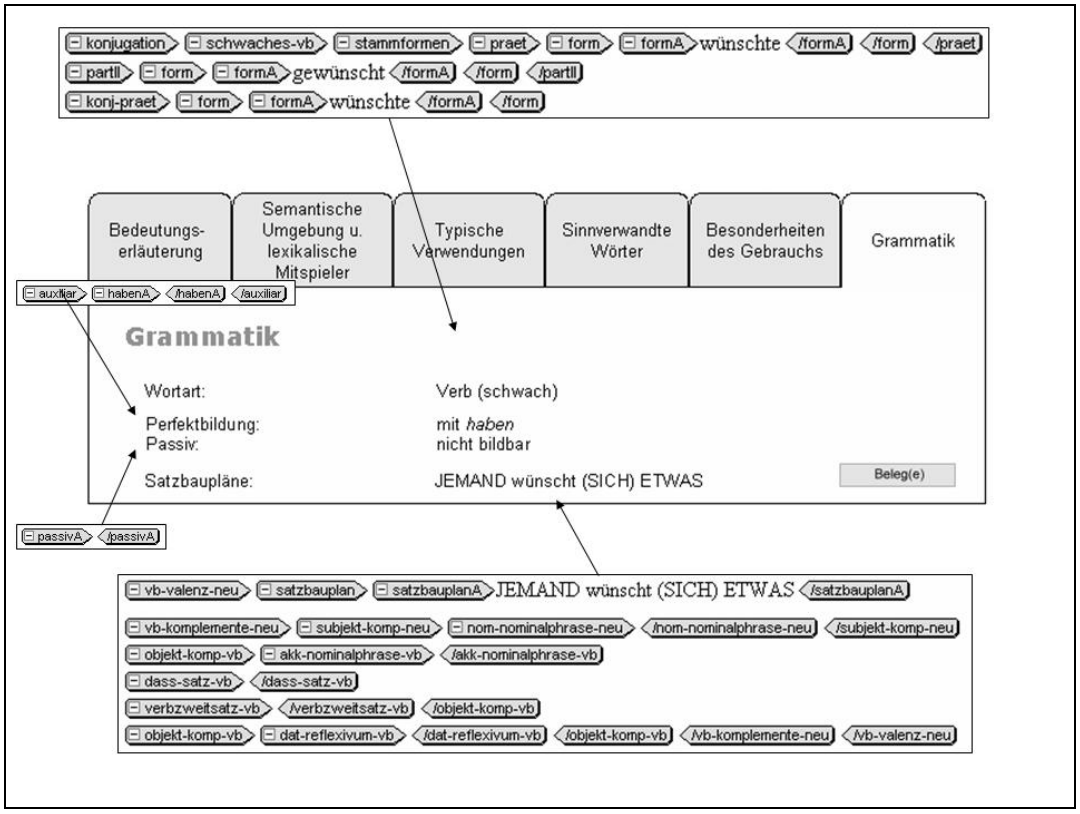


Figure 4. Extracts of XML-entities and their possible online view for German native speakers

the beginning of a session (e.g. user type: non-native speaker, situation of use: reception of a text) and should s/he navigate in all articles with this profile? Or is it more user friendly to being able to change ones profile and look at the same entry with different profiles which means customizing the microstructure dynamically?

As OWID fulfills all technical requirements for a user-adapted presentation, as shown above, this project will be able to realise innovative forms of access to the lexicographic data. Research on the use of the dictionaries published in OWID will be the basis on which different forms of presentation will be developed.

References

Almind, Richard. 2005. *Designing Internet Dictionaries*, in: Hermes 34:37-54.

Atkins, B. T. Sue (Ed.) (1998): *Using dictonaries. Studies of dictionary use by language learners and translators.* (= Lexicographica. Series maior 88), Tübingen.

De Schryver, Gilles-Maurice. 2003. *Lexicographer's Dreams in the Electronic-Dictionary Age*, in: International Journal of Lexicography 16 (2):143-199.

De Schryver, Gilles Maurice / Joffe, David. 2004. *On How Electronic Dictionaries are Really Used*, in: Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France. Vol. I, ed. by Geoffrey Williams / Sandra Vessier:187–196.

Engelberg, Stefan / Lemnitzer, Lothar. 2001. *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Klosa, Annette / Schnörch, Ulrich / Storjohann, Petra. 2006. *ELEXIKO - A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache*, Mannheim, in: Proceedings of the 12th EURALEX International Congress (Atti del XII Congresso Internazionale di Lessicografia), EURALEX 2006, Turin, Italy, September 6th-9th, 2006. Vol. 1, ed. by Carla Marello et al., Alessandria:425-430.

Kunze, Claudia / Lemnitzer, Lothar. 2007. *Computerlexikographie. Eine Einführung*. Tübingen: Narr.

Müller-Spitzer, Carolin (2007): *Das elexiko-Portal: Ein neuer Zugang zu lexikografischen Arbeiten am Institut für Deutsche Sprache*, in: Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Proceedings of the Biennial GLDV Conference 2007 (April 11-13, 2007, Eberhard Karls Universität Tübingen), ed. by Georg Rehm / Andreas Witt / Lothar Lemnitzer:179-188.

Scherer, Tanja. 2008. *Umsetzung von Zugriffsstrukturen bei Online-Wörterbüchern.* Unveröffentlichte Magisterarbeit an der Universität Mannheim, Phi-

losophische Fakultät, Seminar für Deutsche Philologie, Germanistische Linguistik (Prof. Dr. L. M. Eichinger).

Schlaps, Christiane. 2007. *Grundfragen der elektronischen Lexikographie. Elexiko – das Online-Informationssystem zum deutschen Wortschatz.* Ed. by Ulrike Hass. Berlin, New York: de Gruyter 2005. Short review". *Lexicographica* 22:311-314.

Wang, Weiwei. 2001. *Zweisprachige Fachlexikographie. Benutzungsforschung, Typologie und mikrostrukturelle Konzeption*, Frankfurt a.M. (= Angewandte Sprachwissenschaft 8).

Multilingual Conceptual Access to Lexicon based on Shared Orthography: An ontology-driven study of Chinese and Japanese

Chu-Ren Huang
Institute of Linguistics,
Academia Sinica
Nanking, Taipei,
Taiwan 115
churen@sinica.edu.tw

Ya-Min Chou
Ming Chuan University
250 Zhong Shan N. Rd., Sec. 5,
Taipei 111, Taiwan
milesymchou@yahoo.com.tw

Chiyo Hotani
Department of Linguistics
University of Tuebingen
Wilhelmstr. 19
72074 Tübingen, Deutschland
chiyo.hotani@student.uni-tuebingen.de

Sheng-Yi Chen
Institute of Linguistics,
Academia Sinica
Nanking, Taipei,
Taiwan 115
eagles@gate.sinica.edu.tw

Wan-Ying Lin
Institute of Linguistics,
Academia Sinica
Nanking, Taipei,
Taiwan 115
waiin@gate.sinica.edu.tw

Abstract

In this paper we propose a model for conceptual access to multilingual lexicon based on shared orthography. Our proposal relies crucially on two facts: That both Chinese and Japanese conventionally use Chinese orthography in their respective writing systems, and that the Chinese orthography is anchored on a system of radical parts which encodes basic concepts. Each orthographic unit, called hanzi and kanji respectively, contains a radical which indicates the broad semantic class of the meaning of that unit. Our study utilizes the homomorphism between the Chinese hanzi and Japanese kanji systems to ide[1]ntify bilingual word correspondences. We use bilingual dictionaries, including WordNet, to verify semantic relation between the cross-lingual pairs. These bilingual pairs are then mapped to an ontology constructed based on relations to the relation between the meaning of each character and the basic concept of their radical parts. The conceptual structure of the radical ontology is proposed as a model for simultaneous conceptual access to both languages. A study based on words containing characters composed of the "□(mouth)" radical is given to illustrate the proposal and the actual model. The fact that this model works for two typologically very different languages and that the model contains generative lexicon like coersive links suggests that this model has the conceptual robustness to be applied to other languages.

1 Motivation

Computational conceptual access to multilingual lexicon can be achieved through the use of ontology or WordNet as interlingual links. Some languages do conventionally encode semantic classification information, such as the linguistic system of classifiers or the orthographic system of characters. We attempt to make use of these implicitly encoded linguistic knowledge for conceptual access to lexical information.

On the other hand, even though ontology seems to be a natural choice for conceptual framework to access multilingual lexical information, there is no large-scale implementation nor is there any

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 47–54
Manchester, August 2008

direct evidence for psychological reality of the frameworks of ontology. Hence, we hope that using a conventionalized semantic classification system will mitigate some of the problems and provide the constructed ontology some motivation since they are the shared and implicit conceptual systems.

2 Background

2.1. Hanzi and kanji: Shared Orthography of Two Typologically Different Languages

Chinese and Japanese are two typologically different languages sharing the same orthography since they both use Chinese characters in written text. What makes this sharing of orthography unique among languages in the world is that Chinese characters (*kanji* in Japanese and *hanzi* in Chinese) explicitly encode information of semantic classification (Xyu 121, Chou and Huang 2005). This partially explains the process of Japanese adopting Chinese orthography even though the two languages are not related. The adaptation is supposed to be based on meaning and not on cognates sharing some linguistic forms. However, this meaning-based view of kanji/hanzi orthography faces a great challenge given the fact that Japanese and Chinese form-meaning pair do not have strict one-to-one mapping. There are meanings instantiated with different forms, as well as same forms representing different meanings. The character 湯 is one of most famous *faux amis.* It stands for 'hot soup' in Chinese and 'hot spring' in Japanese. In sum, these are two languages where their forms are supposed to be organized according to meanings, but show inconsistencies.

It is important to note that WordNet and the Chinese character orthography are not so different as they appear. WordNet assumes that there are some generalizations in how concepts are clustered and lexically organized in languages and propose an explicit lexical level representation framework which can be applied to all languages in the world. Chinese character orthography intuited that there are some conceptual bases for how meanings are lexicalized and organized, hence devised a sub-lexical level representation to represent semantic clusters. Based on this observation, the study of cross-lingual homo-forms between Japanese and Chinese in the context of WordNet offers an unique window for different approaches to lexical conceptualization. Since Japanese and Chinese use the same character set with the same semantic primitives (i.e. radicals), we can compare their conceptual systems with the same atoms when there are variations in meanings of the same word-forms. When this is overlaid over WordNet, we get to compare the ontology of the two represent systems.

2.2. Hantology and the Ontologization of the Semantic Classification of the Radicals

The design of Hantology differs from other word-based ontology. A typical word-based ontology is WordNet which describes the different relations among synonyms. All of the relations among synonyms are based on the senses of words. Therefore, WordNet only needs to take senses into consideration. Hantology is more complicated than WordNet because it describes orthographic forms, pronunciations, senses, variants, lexicalization, the spread of Chinese characters and Japanese kanji.This approach can systematically illustrate the development of Chinese writing system (Chou et al. 2007).

Hantology also provides mapping with Sinica BOW(Academia Sinica Bilingual Ontological WordNet). Sinica BOW is a Chinese-English Ontology and have mapping with WordNet. Therefore, character-based and word-based ontologies are integrated to provide resources from character to word for Chinese language processing.

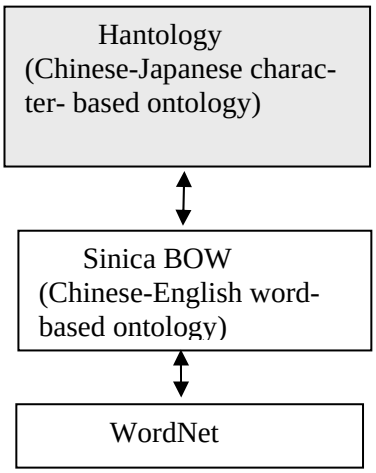


Figure 1. The Mapping among Hantology, Sinica BOW and WordNet

The structure of Hantology is divided into three parts: orthography, pronunciation, and lexicalization.

The orthographic part of Hantology describes the structure of characters, the principles of formatting characters, the evolution of script,

glyph expression, the relation of variant and the spread of Chinese characters.

(1) The structure of characters describes the components of each hanzi/kanji, including semantic and phonetic symbols.

(2) The principles of formatting Chinese characters encode the classification of the relation used to compose the character from its components: The pictographic characters were formed by reformatting the pictures of concrete objects. The ideographic (zhi3shi4, refer-event) characters are formed by abstract representation of an concept. The compound ideographic characters are formed by combining two (ore more) semantic symbols. The semantic-phonetic (xing2sheng1) characters, representing over 90 percent of Chinese character, are formed by combining a semantic symbol and a phonetic symbol.

(3) The evolution of script illustrates the different scripts of Chinese characters. The script is a kind of writing style. Because Chinese characters have been used for thousands years, the scripts have changed.The orthographic forms do not change with different scripts. Hantology provides Bronze, Lesser Seal, Kaishu scripts to illustrate evolution of Chinese scripts used from 3000 years ago.

(4) Variants are the characters with different orthographic forms with identical pronunciation and meaning. For example, Chinese characters台and 臺are variants. Variants relations are an important feature in Hantology, similar to WordNet synset relations.

(5) The contrasts between kanji and hanzi glyphs are also encoded. The Japanese language continues to evolve and change after the adoption of Chinese characters. Hence the kanji system includes both historical changes and cross-lingual variations. The kanji system has its own variants which are not necessarily the same set of variants in the hanzi system. Most of Chinese characters adopted by simplified kanji are the variants already used in Chinese. For example, '国' is a simplified kanji of traditional kanji '國'. In addition, Chinese character '国' is also the variant of Chinese character'國'. So, '國'and'国' both are variants in Chinese and Japanese. But, some simplified kanji are not variants used in Chinese. For example, new kanji '欠' is the variant of old kanji '缺' in Japan. However, '欠' is not the variant of '缺' in Chinese.

The second reason of the kanji orthographic form to to be changed is that Japanese not only adopted Chinese characters but also have created hundreds kanji known as Kokuji (国字). Most Kokuji characters have only Japanese pronunciations. Some of Kokuji have been adopted in Chinese. For example, Kokuji '癌'is also borrowed by Chinese. The meaning of '癌' is the same both in Japanese and Chinese.

3. Preliminaries: Orthography based Mapping of Chinese and Japanese Words

3.1 EDR Japanese-English Dictionary

The Japanese-English dictionary of *EDR Electronic Dictionary* is a machine-tractable dictionary that contains the lexical knowledge of Japanese and English.[1]It contains list of 165,695 Japanese words (jwd) and each of their related information.

In this experiment, the English synset, definition and the Part-of-Speech category (POS) of each jwd are used to determine the semantic relations.

We assume that the concept, synonyms, near-synonyms, and paraphrases are the synset of each jwd.In the case when there is no English definition for the word, we assume that there is no equivalent term in English, therefore we use the concept definition of the jwd as its definition.

3.2 SinicaBow

In the previous experiment, the CWN, which contains a list of 8,624 Chinese word (cwd) entries, was used as the cwd data, however since the number of cwds was too small, many jwds were not mapped, even when there is actually a corresponding J-C word pairs exists.

This time we adopt SinicaBow, which contains 9,9642 entries, hoping to find more valid corresponding J-C word pairs.In SinicaBow, each entry is a definition and it contains one or more cwds corresponds to the definition.

In this experiment, the English synset, definition and the POS of each cwd are used to determine the semantic relations.

3.3List of Kanji Variants

List of 125 pairs of manually matched Chinese and Japanese characters with variant glyph forms provided by Kyoto University.

[1] http://www2.nict.go.jp/r/r312/EDR/index.html

Some Japanese kanji and Chinese hanzi have identical property but have different font and Unicode.This resource contains list of Japanese kanji and Chinese hanzi pairs that the kanji properties are exactly the same but the forms and the Unicode are different.
During the mapping procedure, whenever a Japanese kanji and a Chinese hanzi being compared are in the variant list and are the variants of each other, they are considered to be the identical hanzi.

3.4 Procedure

3.4.1Kanji Mapping

Each jwd is mapped to the corresponding cwd according to their kanji similarity.Such mapping pairs are divided in to the following three groups:
(1) *Identical Kanji Sequence Pairs,* where the numbers of kanji in the jwd and cwd are identical and the n^{th} characters in the two words are also identical.

E.g. 頭, 歌手

(2) *Different Kanji Order Pairs,* where the numbers of kanji in the jwd and cwd are identical, and the kanji appear in the two words are identical, but the order is different.

E.g.	Japanese	Chinese
	制限	限制
	律法	法律

(3) *Partially Identical Pairs,* where at least half kanji in the shorter word matches with the part of the longer word.In the case when the shorter word has 4 or less kanji, 2 of the kanji have to be in the longer word.In the case when the shorter word is only 1 kanji, the pair is not considered.jwd matches with a kanji in the cwd.

E.g.,	Japanese	Chinese
	浅黄色	棕黄色
		蛋黄色的
		黄色的
	宇宙飛行体	飛行
		飛行的
		etc…

In the case no corresponding pair relation (one of the three groups explained above) is found for a jwd or a cwd, each word is classified to one of the following group
(4) unmapped jwd is classified to an independent Japanese
(5) unmapped cwd is classified to an independent Chinese
J-C word pairs in such mapping groups are classified in the following manner: (1) A jwd and a cwd are compared.If the words are identical, then they are an identical kanji sequence pair.(2) If the pair is found to be not an identical kanji sequence pair, check if the pair has identical kanji in different order (equal length).If so, then they are a different kanji order pair.(3) If the pair is found to be not a different kanji order pair, then check the partial identity of the pair.Meanwhile, if they are partially identical (according to the characteristics of partially identical pairs described above), the pair is classified to a partially identical pair.

After the mapping process, if the jwd is not mapped to any of the cwd, the jwd is classified to (4) independent Japanese group. If a cwd is not mapped by any of the jwd, it is classified to (5) independent Chinese group.

The number of Japanese kanji- Chinese hanzi pairs' similarity distribution is shown in Table1.

	Number of Words	Number of J-C Word Pairs
(1) Identical hanzi Sequence Pairs	2815 jwds	20199
(2) Different hanzi Order Pairs	204 jwds	473
(3) Partly Identical Pairs	264917 jwds	8438099
(4) Independent Japanese	57518 jwds	-
(5) Independent Chinese	851 cwds	-

Table1. J-C Hanzi Similarity Distribution (Huang et al. 2008).

3.4.2Finding Synonymous Relation (Word Relation)

After the kanji mapping, each of (1) identical kanji sequence pairs, (2) different kanji order pairs and (3) partially identical pairs is divided into three subgroups;
(1-1, 2-1, 3-1) Synonym pairs with identical POS: words in a pair are synonym with identical POS.

E.g. (1-1) 歌手: singer (noun)
(2-1) 藍紫色 (Japanese) and
紫藍色 (Chinese):
blue-violet color (noun)

(3-1) 赤砂糖 (Japanese) and
紅砂糖 (Chinese):
brown sugar (noun)

(1-2, 2-2, 3-2) Synonym pairs with unmatched POS: words in a pair are synonym with different POS or POS of at least one of the words in the pair is missing.

E.g. (1-2) 包:
(Japanese) action of wrapping (noun)
(Chinese) to wrap (verb)
(2-2) 嗽咳 (Japanese): a cough (noun)
咳嗽 (Chinese): cough (verb)

(1-3, 2-3, 3-3) Relation Unidentified: the relation is not determinable by machine processing with the given information at this point.

E.g.	Japanese	Chinese
(1-3)	湯: hot spring (noun)	湯: soup (noun)
(2-3)	生花: arrangement (noun)	花生: flower peanut (noun)
(3-3)	青葡萄: blue grapes (noun)	葡萄牙: Portugal (noun)

In order to find the semantic relation of J-C word pairs by machine analysis, the jwd and the cwd in a pair are compared according to the following information:

Jwd: English synset (jsyn), definition (jdef) and POS

Cwd: English synset (csyn), definition (cdef) and POS

The process of checking the synonymy of each pair is done in the following manner:

If any of the following conditions meets, we assume that the pair is a synonym pair:

at least any one of the synonym from each of jsyn and csyn are identical

at least one of the word definition contains a synonym of the other word

If any synonym pair was found, check if the POS are identical.If the POS are identical, the pair is classified to a synonym pair with identical POS.Otherwise the pair is classified to a synonym pair with non-identical POS.If the pair is not a synonym pair then they are classified to a relation-unidentified pair.

After the process, each of the subgroups is manually examined to check the actual semantic relations of each word pair.

4. Result

4.1 Word Family as Domain Ontology Headed by a Basic Concept

Chinese radical (*yi4fu2*, ideographs; semantic symbols) system offers a unique opportunity for systematic and comprehensive comparison between formal and linguistic ontologies. Chou and Huang (2005) suggests that the family of Chinese characters sharing the same radical can be linked to a basic concept by Qualia relations. Based on Pustejovsky's Quilia Structure [Pustejovsky, 1995] and the original analysis of "ShuoWen-JieXi"[Xyu, 121], each radical group can be as domain ontology headed by one basic concept.

Chou and Huang (2005) assume that 540 radicals in "ShuoWenJieXi" can each represent a basic concept and that all derivative characters are conceptually dependent on that basic concept. Also, they hypothesis that a radical can be classified into six main types: formal, constitutive, telic, participating, descriptive (state, manner) and agentive. Modes of conceptual extension capture the generative nature of radical creativity. All derived characters are conceptually dependent on the basic concept. In their preliminary studies, word family could be headed by a basic concept and also could be represented ontologies in OWL format.

4.2Data Analysis: Japanese and Chinese Words with Identical Orthography

4.2.1 Kanji Mapping

We present our study over Japanese and Chinese lexical semantic relation based on the kanji sequences and their semantic relations.We compared Japanese-English dictionary of Electric Dictionary Research (EDR) with the SinicaBow in order to examine the nature of cross-lingual lexical semantic relations.

	Identical	Different Order	Part Identical
Synonym (Identical POS)	(1-1) 13610 pairs	(2-1) 567 pairs	(3-1) 37466 pairs
Synonym (Unmatched POS)	(1-2) 2265 pairs	(2-2) 214 pairs	(3-2) 22734 pairs
Relation Unidentified	(1-3) 21154 pairs	(2-3) 2336 pairs	(3-3) 1116141 pairs
Total	(1) 37029 pairs	(2) 3117 pairs	(3) 1176341 pairs
	16950 jwds	1497 jwds	39821 jwds

(4) Unmapped Japanese: 107427 jwds

(5) Unmapped Chinese: 41417 entries
Table 1.J-C Kanji Similarity Distribution

The next step is to find Synonymous Relation. (Word Relation).

	Number of 1-to-1 Form-Meaning Pairs Found by Machine Analysis	% in (1)
(1-1) Synonym (Identical POS)	13610	36.8%
(1-2) Synonym (Unmatched POS)	2265	6.1%
(1-3) Relation Unidentified	21154	57.1%

Table 2. Identical Kanji Sequence Pairs (37029 pairs) Synonymous Relation Distribution

	Number of 1-to-1 Form-Meaning Pairs Found by Machine Analysis	% in (2)
(2-1) Synonym (Identical POS)	567	18.2%
(2-2) Synonym (Unmatched POS)	214	6.9%
(2-3) Relation Unidentified	2336	74.9%

Table 3.Identical Kanji But Different Order Pairs (3117 pairs) Synonymous Relation Distribution

	Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing	% in (3)
(3-1) Synonym (Identical POS)	37466	3.2%
(3-2) Synonym (Unmatched POS)	22734	1.9%
(3-3) Relation Unidentified	1116141	94.9%

Table 4. Partially Identical Pairs (1176341 pairs) Synonymous Relation Distribution

The following tables are summarized tables showing the Japanese-Chinese form-meaning relation distribution examined in our preliminary study.

	Pairs Found to be Synonym	% in (1)	Relation Unidentified	% in (1)
Machine Analysis	15875	42.9 %	21154	57.1%

Table 5. Identical kanji Sequence Pairs (37029 pairs) Lexical Semantic Relation

	Pairs Found to be Synonym	% in (2)	Relation Unidentified	% in (2)
Machine Analysis	781	25.1%	2336	74.9%

Table 6. Identical kanji But Different Order Pairs (3117 pairs) Lexical Semantic Relation

	Pairs Found to be Synonym	% in (3)	Relation Unidentified	% in (3)
Machine Analysis	60200	5.1%	1116141	94.9%

Table7. Partially Identical Pairs (1176341 pairs) Lexical Semantic Relation

Since each entry in SinicaBow corresponds to a definition and each jwd has at least a definition or a concept definition, no pairs with insufficient information to check the semantic relation was found.The data shows that as the word forms of the two languages are closer, the more synonyms are found.In order to confirm this observation and to see the actual semantic relation of each pairs, we will continue with more detailed analysis.In addition, in order to pursue the further details of the Japanese-Chinese words relation, we will also analyze the semantic relations (not only synonymous relation) of the relation-unidentified pairs.

4.2.2 "口(mouth)"Analysis Procedure:

In our experiment, we select the identical kanji Sequence Pairs (POS) as our main resources. Characters with the radical"口(mouth)"are selected. In addition, if any character of the words owns the radical "口(mouth)", then it would be included here for anaylysing the detailed semantic relation between jwd and cwd..

Second, we would like to define the semantic relations of J-C word pairs in more details. We examined the actual semantic relation of J-C word pairs by by classifying into 8 semantic relations and marked the relation into [ ] remark.

1.[SYN](Synonym)
2.[NSN](Near-Synonym)
3.[HYP](Hypernym)
4.[HPO](Hyponym)
5.[HOL](Holonym)
6.[MER](Meronym)
7.[/](No Corresponding Semantic Relation)
8.[??](unable to decide)
The pattern is as follows.
[(JWD>jsyn>詞類>jdef>)-[Semantic Relation]-(CWD)>csyn>詞類>cdef]]

Sample:
[(J)-[HYP]-(C)]@
(J is the hypernym of C)

The examples are shown here. In each pair, we define the semantic relation between the jwd and the cwd. The mapping process would be as follows.

E.g
1. [(**咂**> JWD0028646> N> a condition of being incapable of speaking using the voice>)-[SYN]-(啞> 10137481N> N> paralysis of the vocal cords resulting in an inability to speak> alalia,)]@
2. [(**嘴**> JWD0378514> N> of a bird, a bill> bill)-[SYN]-(嘴> 01278388N> N> horny projecting jaws of a bird> nib,neb,bill,beak,)]@
3. [(**咽喉**> JWD0161758> N> part of an animal called a throat>)-[SYN]-(咽喉> 04296952N> N> the passage to the stomach and lungs; in the front part of the neck below the chin and above the collarbone> pharynx,throat,)]@
4. [(**啄**木鳥> JWD0398785> N> a bird that is related to the picidae, called woodpecker> woodpecker)-[SYN]-(啄木鳥> 01355454N> N> bird with strong claws and a stiff tail adapted for climbing and a hard chisel-like bill for boring into wood for insects> woodpecker,)]@
5. [(人工**呼吸**器> JWD0401642> N> a medical instrument with which a patient can breathe artificially> respirator)-[SYN]-(人工呼吸器> 03233384N> N> a device for administering long-term artificial respiration> inhalator,respirator,)]@

According to our observation, we notice that most of the Japanese kanji can get their synonyms or near-synonyms in Chinese hanzi and the percentage for this relation is about 63 % in characters with the radical"口(mouth) selected from Identical Synonym POS data. Please refer to table1. The distributions of Semantic Relations comparing jwd to cwd in characters with the radical"口(mouth) chosen from Identical Syno PO-Sare as follows.

Semantic Relations between J-C word	Distribution in Characters with the radical□(mouth)	% in Characters with the Radical □(mouth), 486 total pairs
[SYN]	190	39%
[NSN]	129	27%
[HYP]	16	4%
[HPO]	7	2%
[HOL]	11	3%
[MER]	12	3%
[/]	118	25%
[??]	1	1%

Table8. Semantic Relation Distribution in Characters with the radical"口 Mouth"

4.3 Conceptual Access: A Preliminary Model

In this part, we try to apply dimension of conceptual extension of "□(mouth)" radical into the data we have chosen from the Identical Synonym POS data comparing with Japanese kanji and Chinese hanzi.(Please refer to the Appendix A.) A study based on words containing characters composed of the "□(mouth) " radical is given for illustration in this preliminary study. It shows that the conceptual robustness can also be applied to other languages, such as Japanese kanji.

Categories in "口 mouth conceptual extension"	Examples in "口 mouth conceptual extension"	Japanese kanji-Chinese hanziExample
Formal -Sense-Vision&Size	喡	
Formal -Sense-Hearing	叫	
Constitutive	吻、嚨、喉	吻、口吻、嘴、咽喉、喉頭、喉頭炎、喉頭鏡
Descriptive-Active	吐、叫	嘔吐
Descriptive-State	含	含量、含意、含糊、嗜好
Participating-Action	咳、啞、呼、吸	啞、咳嗽、吸血鬼、呼吸、吸盤
Participating-others	哼、嚏	
Participating-instrument	右	左右、右側、右手，周到
Metaphor	启	入口、門口、出入口、出口
TELIC- Subordinate Concept1& Subordinate Concept2		
Subordinate Concept1(Speaking)		
Formal-Property	唐	
Formal-Sense-Hearing	呷	
Constitutive	名、吾	匿名、名詞、名言、名人、物質名詞
Descriptive-Active	吃、哽	吃、吃水線
Participator	吠、喔	狗吠、唯我論、唯

		心論
Participating-Action-Way	呻、吟	唱歌
Participating-others	君，命	君、命令、革命、生命、命運
Subordinate Concept2 (Eating)		
Formal-Sense-Taste	味、啜	味、趣味
Descriptive-Active	噎	
Participating-Action	啜	
Participating-State	嚥	
Participator	啄	啄木鳥、啄木鳥目

Table 9.Jwd Correspondence to“口(mouth) Conceptual Extension” Graph (口(mouth), Basic Concept: the body part which used mainly in Language & Food)

5. Conclusion

The result of the experiment comparing the Japanese and Chinese words is to see their form-meaning similarities.Since the Japanese and the Chinese writing system (kanji) and its semantic meanings are near-related, analyzing such relation may contribute to the future research related to Hantology.In this paper, we examine and analyze the form of kanji and the semantic relations between Japanese and Chinese.This paper describes the structure of Hantology which is a character-based bilingual ontology for Chinese and Japanese. Hantology represents orthographic forms, pronunciations, senses, variants, lexicalization, the spread and relation between Chinese characters and Japanese kanji. The results show Hantology has two implications. First, Hantology provides the resources needed by Chinese language processing for computers.Second, Hantology provides a platform to analyze the variation and comparison of Chinese characters and kanji use.

References

Chou, Ya-Min and Chu-Ren Huang. 2005. *Hantology: An Ontology based on Conventionalized Conceptualization.* Proceedings of the Fourth OntoLex Workshop. A workshop held in conjunction with the second IJCNLP. October 15. Jeju, Korea.

Chou,Ya-Min, Shu-Kai Hsieh and Chu-Ren Huang. 2007. HanziGrid: Toward a knowledge infrastructure for Chinese characters-based cultures. In: Ishida, T., Fussell, S.R., Vossen, P.T.J.M. Eds.: Intercultural Collaboration I. Lecture Notes in Computer Science, State-of-the-Art Survey. Springer-Verlag

Fellbaum Christiane.1998. *WordNet: An Electronic Lexical Database*. Cambridge : MIT Press.

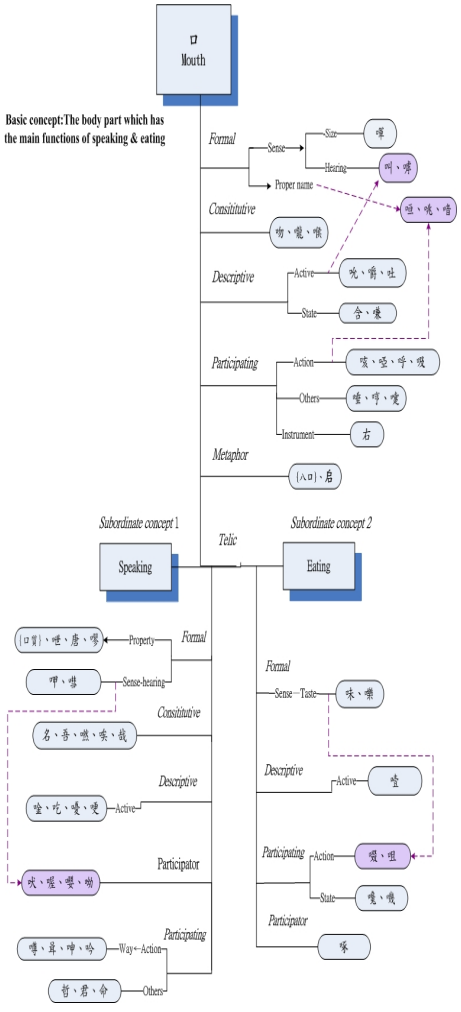
Hsieh, Ching-Chun and Lin, Shih. *A Survey of Full-text Data Bases and Related Techniques for Chinese Ancient Documents in Academia Sinica*, International Journal of Computational Linguistics and Chinese Language Processing, Vol. 2, No. 1, Feb. 1997. (in Chinese)

Huang, Chu-Ren, Chiyo Hotani, Tzu-Yi Kuo, I-Li Su, and Shu-kai Hsieh. 2008. WordNet-anchored Comparison of Chinese-Japanese kanji Word. Proceedings of the 4th Global WordNet Conference. Szeged, Hungary. January 22-25

Pustejovsky, James. 1995. *The Generative Lexicon*, The MIT Press.

Xyu, Sheng. 121/2004. *'The Explanation of Words and the Parsing of Characters' ShuoWenJieZi*. This edition. Beijing: ZhongHua.

Appendix A. The Dimension of “口(mouth) Conceptual extension”.



Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing

Neculai Curteanu
Institute for Computer Science,
Romanian Academy, Iaşi Branch
ncurteanu@yahoo.com

Alex Moruz
Institute for Computer Science,
Romanian Academy, Iaşi Branch
Faculty of Computer Science,
"Al. I. Cuza" University, Iaşi
mmoruz@info.uaic.ro

Diana Trandabăț
Institute for Computer Science,
Romanian Academy, Iaşi Branch
Faculty of Computer Science, "Al.
I. Cuza" University, Iaşi
dtrandabat@info.uaic.ro

Abstract

This paper aims to introduce a new parsing strategy for large dictionary (thesauri) parsing, called *Dictionary Sense Segmentation & Dependency* (DSSD), devoted to obtain the sense tree, *i.e.* the hierarchy of the defined meanings, for a dictionary entry. The real novelty of the proposed approach is that, contrary to dictionary 'standard' parsing, DSSD looks for and succeeds to separate the two essential processes within a dictionary entry parsing: sense tree construction and sense definition parsing. The key tools to accomplish the task of (autonomous) sense tree building consist in defining the dictionary sense marker classes, establishing a tree-like hierarchy of these classes, and using a proper searching procedure of sense markers within the DSSD parsing algorithm. A similar but more general approach, using the same techniques and data structures for (Romanian) free text parsing is SCD (Segmentation-Cohesion-Dependency) (Curteanu; 1988, 2006), which DSSD is inspired from. A DSSD-based parser is implemented in Java, building currently 91% correct sense trees from DTLR (Dicționarul Tezaur al Limbii Române – Romanian Language Thesaurus) entries, with significant resources to improve and enlarge the DTLR lexical semantics analysis.

1 Introduction

Since the last decade, researchers have proven the need for machine readable dictionaries. The idea behind parsing a dictionary entry is the creation of a lexical-semantic tree of senses corresponding to the meanings that define the dictionary lexical entry. The aim of this paper is to introduce a new parsing strategy for thesauri shallow parsing, called *Dictionary Sense Segmentation & Dependency* (DSSD), devoted to the task of extracting the *sense tree*, *i.e.* the hierarchy of the lexical-semantics defined meanings for a dictionary entry. The concrete task which DSSD algorithm was used for is to obtain the sense tree from an entry of the Romanian Language Thesaurus (DTLR – Dicționarul Tezaur al Limbii Române) within the eDTLR research project (Cristea et al., 2007) devised for DTLR electronic acquisition and processing (Curteanu et al., 2007).

In order to obtain the sense tree for a head word, the dictionary entry is divided into primary and secondary senses, respecting a sense hierarchy introduced by sense markers. For the DTLR dictionary, the sense markers hierarchy (presented in Section 3) includes 5 levels. Those are, from the topmost level: *capital letter* markers (**A.**, **B.**, etc.), *Roman numeral* markers (**I.**, **II.**, etc.), *Arabic numeral* markers (**1.**, **2.**, etc.), *filled diamond* ♦ and *empty diamond* ◊. Besides the

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 55–63
Manchester, August 2008

five levels, there exists also a special marker category, the so-called *literal enumeration*, consisting of *lowercase letter* markers **a)**, **b)**, **c)**, etc. The literal enumeration can appear at any of the 5 levels, as presented in Section 3.

Thus, using the sense markers, any dictionary entry is represented as a tree of senses, the lower levels being more specific instances of the higher levels.

For example, for the dictionary entry *verb*, the sense tree contains 3 senses corresponding to level 3, one of them having a sub-sense corresponding to level 5. Each sense/sub-sense can have its own definition (gloss) or examples.

```
<entry>
  <hw>VERB</hw>
  <senses>
    <marker level="3">1.
      <definition>…</definition>
      <marker level="5">◊
      <definition>…</definition>
      </marker>
    </marker>
    <marker level="3">2.
      <definition>…</definition>
    </marker>
    <marker level="3">3.
      <definition>…</definition>
    </marker>
  </senses>
</entry>
```

The presented method can be applied to any dictionary, provided that a hierarchy of the sense markers of the dictionary is established.

The paper is organized as follows: Section 2 points out the characteristic features of DSSD strategy, discussing the special relationship between DSSD and SCD *parsing strategy for general text*, on one hand, and between DSSD and the standard *dictionary entry parsing* (DEP), on the other hand. Section 3 presents the main components of the DSSD strategy: DTLR sense marker classes, their dependency hyper-tree structure, and the DSSD parsing algorithm. The final Section 4 discusses the current stage implementation (in Java) of the DSSD algorithm, exposing several parsed examples. Possible sources of error and ambiguity in the DSSD parsing process are discussed, and further developments of DSSD analysis software are outlined.

2 DSSD compared to Free Text Parsing and to Dictionary 'Standard' Parsing

This section outlines the origins of the DSSD idea, pointing out the connections between DSSD and free text parsing based on the SCD linguistic strategy (Curteanu 2006), on one hand, and between DSSD and dictionary *standard* parsing, *e.g.* (Neff, Boguraev; 1989), (Lemnitzer, Kunze; 2005), (Hauser, Storrer; 1993), on the other hand. The main difference (and positive feature) of the DSSD strategy compared to the standard approach to *dictionary entry parsing* (DEP), *e.g. LexParse* system in (Hauser, Storrer; 1993), (Kammerer; 2000), (Lemnitzer, Kunze; 2005), or JavaCC grammar-based parsing in (Curteanu, Amihaesei; 2004), is that DSSD *detached completely* the process of *sense tree building* from the process of *sense definition parsing*, within the DEP general task. This fact is clearly reflected in Fig. 2, which compares, at the macro-code level, the main four DEP operations for standard DEP and DSSD strategies.

2.1 SCD Marker Classes, Hierarchy, and Parsing Algorithms

DSSD parsing strategy involves a configuration of components that is similar (but less general) to the SCD (Segmentation-Cohesion-Dependency) parsing strategy, developed and applied to (Romanian) free text analysis (and generation) (Curteanu; 2006). The process of solving the parsing of DTLR entries have been inspired by the resemblance between the classes of DTLR sense markers and the SCD marker classes on one side, and between the *sense trees* of (DTLR) dictionary entries and the *discourse trees* of finite-clause dependency trees at sentence or paragraph levels on the other side. While discourse trees provide a formal similarity to the sense trees, nucleus–satellite rhetorical relations among discourse segments is quite different to the *subsumption relation* of *lexical semantics nature* among the sub-sense definitions (sub-senses) of a dictionary entry.

The *subsumption* relation is defined as follows: $sense_1$ *subsumes* $sense_2$ if (informally) $sense_1$ is *less informative* (or, more general) than $sense_2$, or if (formally) the sense tree of $sense_1$ is a (*proper*) *subtree* of $sense_2$. DSSD parsing of an

Parsing Strategy	SCD markers & DSSD markers	Semantics to be applied on the parsed textual spans	Resulted structures of the parsing process
SCD	**M4-class** (discourse) markers	rhetorical discourse semantics , *i.e.* RST discourse (high-level cohesion) dependencies	*discourse tree* (of RST-based discourse segments)
	M3-class (inter-clause) markers	inter-clause predicational semantics, *i.e.* Predicate-Argument (global-level cohesion) dependencies among finite clauses	clause-level dependency trees based on syntactic or semantic relations
	M2-class (clause) markers	single finite-clause predicational semantics, *i.e.* Predicate-Argument (local-level cohesion) dependencies among VG-NGs (Verbal Group – Noun Groups)	single finite clause(s)
	M1-class intra-clausal (phrase) markers	non-finite predicational semantics, *i.e.* (local-level cohesion) dependencies *inside* VG and NGs (Verbal Group – Noun Groups)	simple and complex VGs; simple and complex (predication-related) NGs
	M0-class flexionary markers of lexical categories	lexical semantics categories	lexical textual words = inflected words
SCD - DSSD	**M(–1)-class** of lemmatization markers for DTLR lexical entries	semantic description at the *lexicon* level	lexical lemmatized words = dictionary entries
DSSD	**sense and subsense definition markers** of a DTLR lexical entry	*subsumption* relations between the *subsenses* of a DTLR lexical entry (*cohesion-free* semantics)	*sense trees* and (XCES-TEI 2007 codification-based) *sense definitons* of DTLR entries

Fig. 1. DSSD *vs.* SCD marker classes, the corresponding semantics and textual structures

entry sense tree works in an akin *Breadth-First*, *Top-Down* manner as SCD does, for those classes of markers that produce only *segmentation* and *binary dependency* between discourse segments or finite clauses, ignoring the more complex "*cohesion*" relationship. Thus one can rightly say that DSSD approach is *derived* from the SCD parsing strategy (Fig. 1).

SCD parsing strategy is exposed at large in (Curteanu 2006). SCD-based discourse parsing presents a special interest for DSSD because of their (formal) algorithmic analogy. The method proposed by the SCD strategy includes building the discourse tree by the intensive use of discourse markers, while discourse segments are obtained by clause parsing. Employing the results of the SCD clausal parsing and a database which contains information about the discourse markers, one can obtain the discourse structure of a text. The outcome is represented as a *discourse tree* whose terminal nodes are clause-like structures, having specified on the arcs the name of the involved *rhetorical relations*.

The SCD segmentation / parsing algorithm in (Curteanu 2006) may have the same shape of a *Breadth-First* (or sequential-linear) processing form as DSSD does, using as input a morphologically tagged text, obtaining the finite clauses and sub-clausal phrase (XG-)structures. Data representation is in standard XML and the implementation of the SCD algorithm for free text parsing is made in Java. (Curteanu 2006) presents *recursive Breadth-First* (and *Depth-First*), or *parallel Breadth-First* shapes of the SCD segmentation-parsing algorithms.

The relationship between SCD and DSSD parsing strategies, the former devoted to the free text parsing and the latter to be used for DEP, could be summarized as follows: the two strategies work formally with the same technology, using very similar analysis tools and data structures, including the same *Breadth-First* search strategy. The clear *distinction* between SCD and DSSD consists in the quite different kind of texts to be analyzed (free text *vs.* dictionary entry text), and the two different (but complementary) semantics that *drive* the corresponding parsing structures: *predicational* and *rhetorical* (cohesion-proper) semantics for SCD, and *lexical semantics* (cohesion-free) for DSSD. The table in Fig. 1 gives a detailed comparison between the two parsing strategies. The SCD parsing technology, especially with its presently discovered DSSD sub-sort, evolves (at least) three features: *generality* (different text structures), *flexibility* (different underlying semantics), and *adequacy* (proper text markers and their corresponding hierarchies).

Dictionary Classical Parsing Strategy	DSSD Parsing Strategy
For *i* from 0 to *MarkerNumber* ① **Sense-*i* Marker Recognition;** ② **Sense-*i* Definition Parsing;** *If*(Success) ③ **Attach (Parsed) Sense-*i* Definition to Node-*i*;** ④ **Add Node-*i* to EntrySenseTree;** *Else* Fail and Stop. EndFor **Output: EntrySenseTree** *with* **Parsed Sense Definitions** (only if all sense definitions are parsed). **Notice:** *MarkerNumber* is the number of the input marker sequence.	For *i* from 0 to *MarkerNumber* ① **Sense-*i* Marker Recognition;** **Assign (Unparsed) Sense-*i* Definition** to **Node-*i*;** ④ **Add Node-*i* to EntrySenseTree;** **Standby** on **Sense-*i* Definition Parsing;** EndFor **Output: EntrySenseTree.** **Node-*k*** = Root(**EntrySenseTree**); *While* not all nodes in **EntrySenseTree** are visited ② **Sense-*k* Definition Parsing;** *If*(Success) ③ **Attach Sense-*k* Definition to Node-*k*;** *Else* **Attach Sense-*k* Parsing Result to Node-*k*;** **Node-*k*** = getNextDepthFirstNode(**EntrySenseTree**) Continue EndWhile. **Output: EntrySenseTree** *with* **Parsed** *or* **Unparsed Sense Definitions**

Fig. 2. A macro-code comparison of classical and DSSD parsing strategies

2.2 DSSD Approach *vs.* Standard DEP

Another perspective on DSSD is outlined in this section: the novelties of DSSD approach fetched to the standard DEP, *e.g.* (Neff, Boguraev; 1989), (Lemnitzer, Kunze; 2005), (Kammerer, 2000). DSSD applies the same "technology" as SCD strategy does, *i.e. marker classes*, *specific hierarchies*, and adequate *searching procedures* embedded and governing the parsing algorithms. Most important, DSSD parse and *construct* the *sense tree* of a (DTLR) dictionary entry, *independently of*, and possibly *lacking the*, DTLR sense definition parsing process.

In the standard DEP, including the Java-grammar based construction of parsers in the JavaCC environment (Curteanu, Amihaesei, 2004; Curteanu et al., 2007), building the *sense tree* for an entry is inherently embedded into the general process of parsing *all* the sense and sub-sense definitions enclosed into the dictionary entry. In the same typically (standard) DEP way works also the parser in (Neff, Boguraev; 1989) or *LexParse*, (Kammerer; 2000: 10-11) specifying that the *LexParse* recognition strategy is a *Depth-First*, *Top-Down* one.

The advantage of the proposed DSSD approach is that it "ignores", at least in the beginning, the "details" of sense definitions, concentrating only on the *sense marker* discovery and their dependency establishing. The result is that DSSD parsing concentrates on and obtains, in the first place, the *sense tree* of a DTLR entry. Of course, parsing of a dictionary entry does not means *only* its sense tree, but the entry sense tree represents the essential, indispensable structure for *any* kind of DEP.

Based on different types of DTD standards for dictionary text representation, such as CONCEDE-TEI (Erjavec et al. 2000; Kilgarriff 1999, Tufis 2001) or (XCES-TEI; 2007), the parsing process may continue "in depth" for identifying the (other important) fields of sense and sub-sense definitions. DSSD strategy has the quality of being able to compute independently the entry sense tree, prior to the process of sense definition parsing. Subsequently, the process of parsing the sense definitions can be performed separately, one by one, avoiding the current situation when the general parsing of an entry may be stopped simply because of a single (even if the last one) unparsable sense definition.

The procedural *pseudo-code* in Fig. 2 shows clearly the important difference between *standard* DEP and DSSD *parsing*, with the essential advantage provided by DSSD: standard DEP is based on *Depth-First* search, while DSSD works with *Breadth-First* one. Specifically, the procedural running of the four operations that are compared for the standard DEP and DSSD strategies, labeled with ①, ②, ③, ④, are organized in quite different cycles: in the table left-side (standard DEP), there is a single, large running cycle, ① + ②, under ② being embedded (and strictly depending) the sub-cycle ③ + ④. The DSSD parsing exhibits two distinct (and in-

dependently) running cycles: ① + ④, for constructing the (DTLR) sense trees, and ② + ③, devoted to parse the sense definitions and to attach the parsed or unparsed sense definitions to their corresponding nodes in the sense tree(s).

We emphasize firstly, that the second procedural cycle is optional, and secondly, that the first cycle is working on the sense marker sequence of the entry (either correct or not), the DSSD output being an entry sense tree in any case (either correct or not). This is why the DSSD algorithm never returns on FAIL, regardless whether the obtained sense tree is correct or not.

3 DTLR Marker Classes, their Dependency Structure, and the DSSD Parsing Algorithm

As already pointed out, DSSD can be viewed as a simplified version of SCD, since only the *segmentation* and *dependency* aspects are involved, the (local) *cohesion* matters being without object for the (one-word) lexical semantics of DSSD. As in the case of SCD, the DSSD parsing strategy requires a set of *marker classes* (in our case, DTLR sense markers), arranged in a *hierarchy* illustrated in Fig. 3, and described below:

The *capital letter* marker class (**A.**, **B.**, etc.) is the topmost level on the sense hierarchy of DTLR markers (see Fig. 3) for any given dictionary entry. When it appears, this marker designates the (largest-grained meaning) *primary senses* of the lexical word defined. If the top level marker has only one element of this kind, then the marker is not explicitly represented.

The *Roman numeral* marker class (**I.**, **II.**, etc.) is the *second-level* of sense analysis for a given DTLR entry. It is subsumed by a capital letter marker if some exists for the head word; if a capital letter marker does not exist (it is not explicitly represented), the Roman numeral marker appears on the topmost level of the sense tree. If the lexical entry has only one sense value for this analysis level, the marker is not explicitly represented.

The *Arabic numeral* marker class (**1.**, **2.**, etc.) is the *third-level* of sense analysis for a DTLR entry. It is subsumed by a Roman numeral marker if there exists some for the entry; if a Roman numeral marker is not explicitly represented, it is subsumed by the first explicit marker on a higher level. If the entry has only one sense value for this level of sense analysis, the marker is not explicitly represented. These first *three levels* encode the *primary senses* of a DTLR lexical entry.

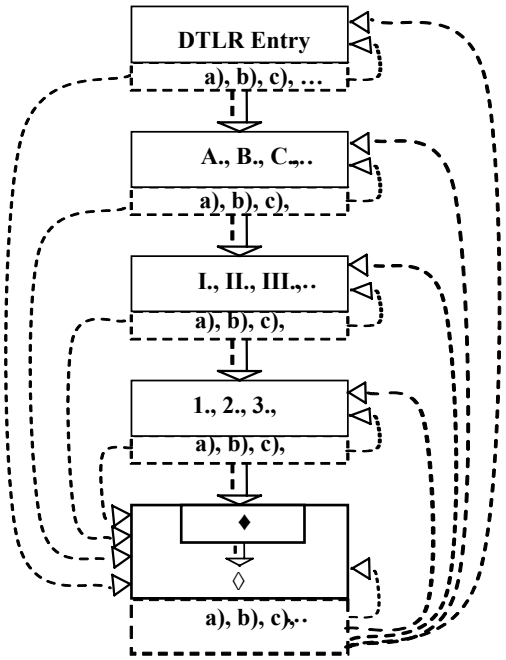


Fig. 3. The tree-like dependency structure for the classes of DTLR markers

The *filled diamond* marker class is the *fourth-level* of sense analysis and it is used for enumerating *secondary* (finer-grained) *senses* of a DTLR entry. It is generally subsumed by any explicit DTLR sense marker on a higher level, *i.e.* any of the primary sense markers.

The *empty diamond* marker class is the *fifth-level* of sense analysis and it is used for enumerating expressions for a given, *secondary sub-sense*. It is generally subsumed by a filled diamond marker or by any primary sense marker.

The *lowercase letter* markers **a)**, **b)**, **c)**, etc. are not an actual class of sense markers, but rather a *procedure* used to refine, through *literal enumeration*, a semantic paradigm of a DTLR entry sense or sub-sense. A lowercase letter marker does not have a specific level on the marker class tree-like hierarchy since it belongs to the sense marker level (of either primary or secondary sense) that is its parent. The important rules of the *literal enumeration* procedure in DTLR are: **(a)** it associates with the hierarchy level of the sense marker class to which is assigned (in Fig. 3), and **(b)** it can embed lower (than its parent level) senses, provided that each literal enumeration is closed finally on the sense level to which it belongs.

Fig. 3 is a *hyper-tree hierarchy* of the DTLR sense marker classes since (at least) the lowest hyper-node contains recursively embedded dia-

mond-marker nodes. The *dashed arrows* point to the upper or lower levels of DTLR sense marker hierarchy, from the *literal enumeration* layer-embedded level. The *continuous-dashed arrows* in Fig. 3 point downwards from the higher to the lower priority levels of DTLR marker class hyper-tree. Because of its special representation characteristics, the literal enumeration is illustrated on a layer *attached* to the hierarchy level to which it belongs, on each of the sense levels. Some examples supporting the marker hierarchy in Fig. 3, including the literal enumeration that can appear at any DTLR sense level, are presented below:

I. Literal enumeration under a filled diamond (secondary sense):

```
<entry>
  <hw>VÍȚĂ[2]</hw>
  <pos>s. f.</pos>
  <senses>
    <marker>I.
      <marker>1.
      <definition> (De obicei determinat prin „de vie") Arbust din familia vitaceelor, cu rădăcina puternică, cu tulpina scurtă, …</definition>
        <marker>♦
        <definition> C o m p u s e: viță-albă = </definition>
          <marker>a)
          <definition> arbust agățător din familia ranunculaceelor, cu tulpina subțire, cu frunze penate...; </definition>
          </marker>
          <marker>b)
          <definition>(regional) luminoasă (Clematis recta). Cf. CONV. LIT. XXIII, 571, BORZA, D. 49, 301; </definition>
          </marker>
          <marker>c)
          <definition>(învechit) împărătească (Bryonia alba).....</definition>
          </marker>
        </marker>
      </marker>
    </marker>
  </senses>
</entry>
```

II. Literal enumeration under an Arabic numeral (primary sense):

```
<entry>
  <hw>VERIGÚȚĂ</hw>
  <pos>s. f.</pos>
  <senses>
    <definition>Diminutiv al lui v e r i g ă. Cf. LB, POLIZU, DDRF, BARCIANU, ALEXI, W., TDRG, CADE, SCRIBAN, D., DL, DM, DEX.</definition>
    <marker>1.
      <marker>a)
      <definition> (Prin Transilv. şi prin sudul Mold.) Cf. v e r i g ă (2 c). Cf. ALR II 6 653/95, 192, 605.
</definition>
      </marker>
      <marker> b)
      <definition>Cf. v e r i g ă (2 b). Şi am dat cercel în narea ta şi veriguţe în urechile tale. BIBLIA (1688), 543[1]/25. La ferestre spînzurau nişte perdeluţe de adamască, aninate în nişte veriguţe ce se înşirau pe o vargă de fier. GANE, N. II, 160. </definition>
      </marker>
    </marker>
    <marker>2.
    <definition> (Popular) Verighetă. Cf. SCRIBAN, D., ȚIPLEA, P. P., BUD, P. P. Mi-o dat o veriguță Şi-ntr-on an i-am fost drăguță. BÎRLEA, C. P. 143. </definition>
    </marker>
  </senses>
</entry>
```

III. Literal enumeration directly under the entry root:

```
<entry>
  <hw>VENTRICÉA</hw>
  <pos>s. f.</pos>
  <senses>
    <definition> Numele mai multor specii de plante erbacee (folosite în medicină): </definition>
    <marker>a)
    <definition> ventrilică (c) (Veronica persica). Cf. GRECESCU, FL. 442, PANȚU, PL., CADE. Un gorun negru şi singuratic… e năpădit la poale de ventricele cu spicuri albăstrii....; </definition>
    </marker>
    <marker>b)
    <definition> ventrilică (a) (Veronica officinalis). Cf. TDRG, BORZA, D. 179, 300; </definition>
    </marker>
    <marker>c)
    <definition>bobornic (Veronica prostrata). Cf. BORZA, D. 179, 300. </definition>
    </marker>
  </senses>
</entry>
```

The DSSD algorithm for the construction of the DTLR sense tree, according to the marker hierarchy described in Fig. 3, is the following:

```
Stack S
Tree T
S.push(root)
while article has more markers
  crt = get_next_marker()
  while crt > S.top() – get to the
first higher rank marker in the
stack
    S.pop()
  if(crt = lowercaseLetter)
    S.top.addPart(crt) – add a low-
ercase marker as a subset of the
higher level sense value
```

```
      crt.level=S.top.level+1 - the
  lowercase letter maker is given a
  level in accordance to the level of
  its parent
      S.push(crt)
    else
      S.top.add_son(crt) - add the
  son to the higher level marker in
  the stack
      S.push(crt) - add the current
  marker to the stack
```

The DSSD parsing algorithm was implemented in Java and running examples of its application on DTLR entries are presented in Section 4. While the DTLR sense marker recognition in DSSD is achieved with a *Breadth-First* search, the marker sequence analysis for sense tree construction is based on a *Depth-First* parsing of the sense marker sequence input, as it uses a stack to keep track of previous unfinished (in terms of attaching subsenses) sense markers.

4 DTLR Parsing with DSSD Algorithm: Examples and Developments

4.1 DSSD Parser Applied on DTLR Entries

The enclosed Fig. 4 shows the result of applying the DSSD Java parser described in Section 3 on a DTLR entry. We notice that the presented input example (*VENIT*[2]) represents just sequences of DTLR sense markers. The entry for which the parsing was conducted is given only as tags, in part below (the entire entry spans for more than two dictionary pages):

```
<entry>
  <hw><VENÍT², -Ă </hw>
  <pos>adj. </pos>
  <senses>
    <definition>...</definition>
    <marker>1.
    <definition>...</definition>
    <marker>2.
    <definition>...</definition>
      <marker>◊
        <marker> a)
        <definition>...</definition>
        </marker>
        <marker> b)
        <definition>...</definition>
        </marker>
        <marker> c)
        <definition>...</definition>
        </marker>
      </marker>
      <marker>◊
        <marker> a)
        <definition>...</definition>
        </marker>
        <marker> b)
        <definition>...</definition>
        </marker>
      </marker>
    </marker>
  </senses>
</entry>
```

```
− <entry>
    <list>VENIT2, -Ă 1. 2. ◊ a) b) c) ◊ a) b) n-11</list>
  − <node value="VENIT2, -Ă" class="0">
      <node value="1." class="6"> </node>
    − <node value="2." class="6">
      − <node value="◊" class="10">
        − <parts>
            <node value="a)" class="11"> </node>
            <node value="b)" class="11"> </node>
            <node value="c)" class="11"> </node>
          </parts>
        </node>
      − <node value="◊" class="10">
        − <parts>
            <node value="a)" class="11"> </node>
            <node value="b)" class="11"> </node>
          </parts>
        </node>
      </node>
    </node>
  </entry>
− <entry>
```

Fig. 4. DSSD parsing for the sense tree building of DTLR entry **VENIT²**

As one can see, the input of the sense tree parser is the DSSD marker sequence of the considered DTLR entry (the <list> tag in Figure 4). The output of the parsing is much less verbose than the original dictionary entry, since the sense definitions and the entire example text is not depicted, in order to better observe the sense tree of the entry. Also, this representation proves that the understanding of the sense definitions is not strictly necessary for building the sense tree, a task for which the marker hierarchy discussed in Section 3 is sufficient.

Fig. 5 presents the sense tree for the dictionary entry "VIÉRME" (En: *worm*). It can be seen that this particular entry is quite large, with the original dictionary text spanning for more than *six* pages of DTLR thesaurus.

After its completion, the DSSD parser was tested on more than 500 dictionary entries (of medium and large sizes), the only ones already in electronic format to which we had access to at the moment (the vast majority of dictionary volumes is only available in printed form). The success rate was determined to be 91.18%, being

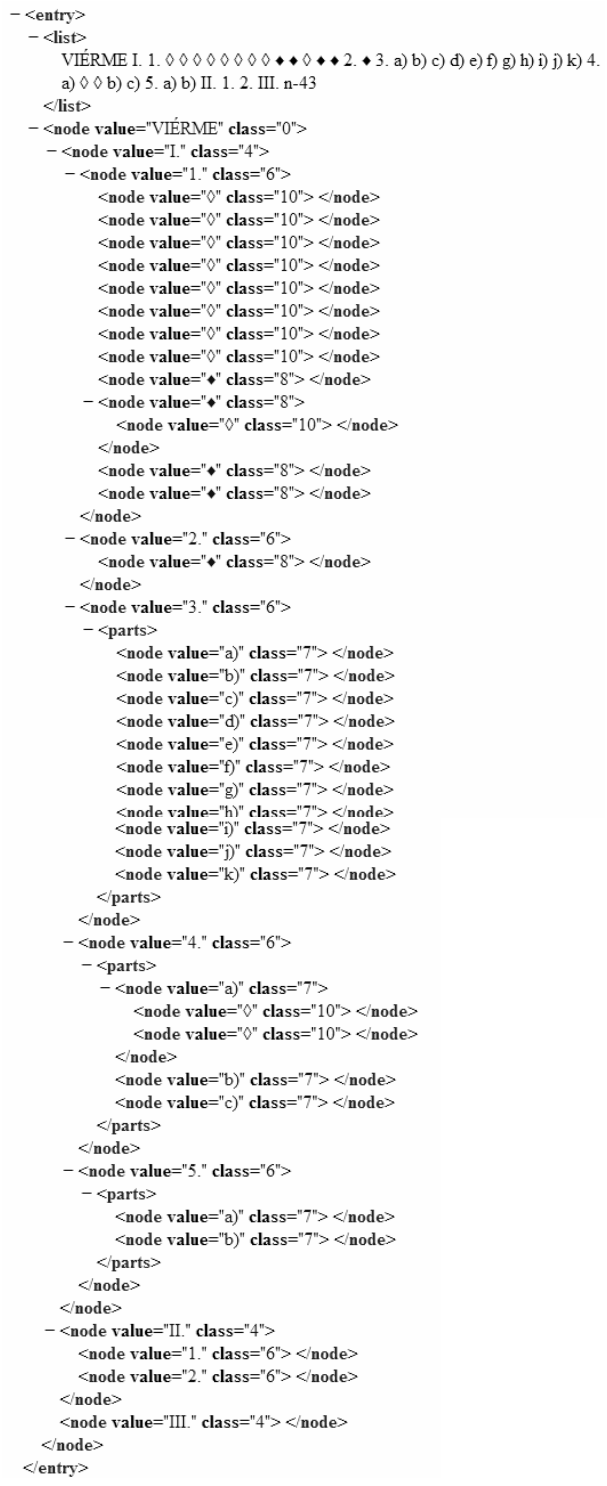


Fig. 5. Sense tree for the dictionary entry "VIÉRME"

computed as a perfect match between the output of the program and the gold standard.

Furthermore, it is worth noting that an article with only *one* incorrect parse (*i.e.* one node in the sense tree attached incorrectly) was considered to be erroneously parsed in its entirety, an approach which disregards all the other correctly attached nodes in that entry. This fact shows also significant improvement resources of the DSSD parsing process.

4.2 Error and Ambiguity Sources in DTLR Parsing

It is worth to mention some sources of error and ambiguity found in DSSD parsing for DTLR sense tree computing. We grouped the error sources in three main classes:

I. Inconsistencies in writing the original DTLR article

A first source of parsing errors is the non-monotony of the marker values on the same level of sense marker hierarchy (Fig. 3):

Ex.1. **A.** [**B.** missing] … **C.** etc.;

Ex.2. **2.** [instead of **1.**]... **2.** etc.;

Ex.3. **a)**… **b)** … **c)** … **b)** [instead of **d)**]etc.

The tree structure returned by the parser does not consider the consistency of each marker level. Thus, in **Ex.1**, it will place the two markers **A.** and **C.** as brother nodes in the sense tree. A (partial but feasible) solution for the parser is to check the *strict monotony* of the marker neighbors, an operation which is useful also when sense markers interfere with literal enumeration.

A validity of the marker succession at each level will be checked after the completion of the sense tree parsing.

II. Ambiguity in deciding which is the regent and which is the dependent (sub)sense

An inherently ambiguity was found for the following sequences of DTLR sense markers:

Ex.4. **1. a) b) c)** ◊ [◊]

The problem occurs since one can not discern between attaching the first (and / or second) "◊" as depending on **c)** or on the upper level marker (**1.)**. Solving these ambiguities is a problem related on syntactic and / or semantic contexts of the involved (multiple) pairs of markers. Namely, if "**c)**" is the last small letter in the literal enumeration, then "◊" is attached to the "**1.**" marker (and sense), while if "**c)**" in the literal enumeration, followed by "◊", has a continuation "**d)**" in the literal enumeration, then "◊" depends on its small letter "**c)**" regent sense.

III. More refined subsense classification

A third source of errors when creating the sense tree is met within the following marker sequence **I. 1. ♦ a) b) c) a) b)**. Even if at a quick

look the problem with the inconsistent literal enumeration is similar to the problems presented in the first class, at a closer inspection we realized that under the full diamond ♦ there are three subsenses (three expressions), two of them having literal enumeration: (1) **viță-albă = a)... b)... c)**; (2) **viță-neagră = ...**; (3) **vița-evreilor = a)...b)**. To solution this problem makes necessary a more refined subsense classification within the sense definition and adding possible new markers to the hierarchy. Working to solve these problems is in good progress, as it concerns types of sense structure closely related to various sense definition parsing, the next step in the development of the DSSD dictionary parser.

We already identified *seven* definition types, encoded as follows, together with the most important *dependency conditions* among the definitions below, within DTLR senses and subsenses:

1. *MorfDef* (Morphological Definitions);
2. *SpecDef* (Specification-based Definitions);
3. *SpSpec* (Spaced-character Definitions);
4. *RegDef* (Regular-font Definitions);
5. *BoldDef* (Bold-font Definitions);
6. *ItalDef* (Italic-font Definitions);
7. *ExemDef* (Example-based Definitions),

The 4, 5, 6, definition types are possibly followed by the *literal enumeration* scheme of sense codification.

Further developments of DSSD analysis software are meant to be achieved: **(a)** The complete parsing of a DTLR entry entails the natural extension of DSSD approach towards sense definition parsing and representation within the XCES TEI P5 (2007) standard set of tags. **(b)** A specialized subset of TEI P5 tags for representing all the types of definitions met within the primary and secondary senses of a DTLR entry is necessary. **(c)** Resolution of all the references within a DTLR entry is necessary: references to the excerpt sources (sigles), reference to a sense within the same entry (internal reference), or to a (sub)sense within another entry (external reference). **(d)** Verification of the sense-tree correctness can be achieved by restoring the linear structure of a DTLR entry from its parsed sense-tree representation, and comparing it with the DTLR original entry.

Acknowledgement. The present research was financed within the **eDTLR** grant, PNCDI II Project No. 91_013/18.09.2007.

References

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007): *The Digital Form of the Thesaurus Dictionary of the Romanian Language*. In Proceedings of the 4th International IEEE Conference SpeD 2007.

Curteanu, Neculai (1988): *Augmented X-bar Schemes*. COLING'88 Proceedings, Budapest, pp. 130-132.

Curteanu, N., E. Amihăesei (2004): *Grammar-based Java Parsers for DEX and DTLR Romanian Dictionaries*. ECIT-2004 Conference, Iasi, Romania.

Curteanu, N. (2006): *Local and Global Parsing with Functional* (F)X-*bar Theory and SCD Linguistic Strategy*. (I.+II.), Computer Science Journal of Moldova, Academy of Science of Moldova, Vol. 14 no. 1 (40):74-102 and no. 2 (41):155-182.

Curteanu, N., D. Trandabăț, G. Pavel, C. Vereştiuc, C. Bolea (2007): *eDTLR – Thesaurus Dictionary of the Romanian Language in electronic form.* Research Report at the PNCDI II Project No. 91_013/18.09.2007, Phase 2007, and (D. Cristea, D. Tufiș, Eds.) *eDTLR Parsing – The Current Stage, Problems, and Development Solutions*, Romanian Academy Editorial House (in Romanian – to appear).

DLR Revision Group (1952): Codification Rules for the Dictionary (Thesaurus) of the Romanian Language. Institute of Philology, Bucharest, Romanian Academy.

Erjavec, T, Evans, R., Ide, N., Kilgariff A., (2000): The CONCEDE Model for Lexical Databases. Research Report on TEI-CONCEDE LDB Project, Univ. of Ljubljana, Slovenia.

Hauser, R., Storrer, A. (1993). *Dictionary Entry Parsing Using the LexParse System*. Lexikographica 9 (1993), 174-219

Kammerer, M. (2000): *Wöterbuchparsing Grundsätzliche Überlegungen und ein Kurzbericht über praktische Erfahrungen*, http://www.matthias-kammerer.de/content/WBParsing.pdf

Lemnitzer, L., Kunze, C. (2005): *Dictionary Entry Parsing*, ESSLLI 2005

Neff, M., Boguraev, B. (1989) *Dictionaries, Dictionary Grammars and Dictionary Entry Parsing*, Proc. of the 27th annual meeting on Association for Computational Linguistics Vancouver, British Columbia, Canada Pages: 91 - 101

Tufiș, Dan (2001): From Machine Readable Dictionaries to Lexical Databases, RACAI, Romanian Academy, Bucharest, Romania.

XCES TEI Standard, Variant P5 (2007): http://www.tei-c.org/Guidelines/P5/

Lexical-Functional Correspondences and Their Use in the System of Machine Translation ETAP-3

A.S. Andreyeva
Moscow State University, Philological Faculty,
Department of Theoretical and Applied Linguistics
Moscow, Vorobjevi Gori, 1st Building of the Humanities
andreyevs@mtu-net.ru

Abstract

ETAP-3 is a system of machine translation consisting of various types of rules and dictionaries. Those dictionaries, being created especially for NLP system, provide for every lexeme not only data about its characteristics as a separate item, but also different types of information about its syntactic and semantic links to other lexemes.

The paper shows how the information about certain types of semantic links between lexemes represented in the dictionaries can be used in a machine translation system. The paper deals with correspondences between lexical-functional constructions of different types in the Russian and the English languages.

Lexical-functional construction is a word-combination consisting of an argument of a lexical function and a value of this lexical function for this argument.

The paper describes the cases when a lexical functional construction in one of these languages corresponds to a lexical-functional construction in the other language, but lexical functions represented by these two constructions are different. The paper lists different types of correspondences and gives the reasons for their existence. It also shows how the information about these correspondences can be used to improve the work of the linguistic component of the machine translation system ETAP-3.

1 Introduction

The concept of lexical function (LF) was proposed in Igor Mel'čuk's "Meaning ⇔ Text Theory" (Mel'čuk, 1974; Mel'čuk & Zholkovsky, 1984; Mel'čuk et al., 1984, 1988, 1992) as the means of description of certain types of lexeme meaning correlations. "Lexical function f describes the dependence that determines for the certain word or word-combination such a multitude of words or word-combinations $\{Y_i\}=f(X)$, that for every X_1, X_2 the following statement is true: if $f(X_1)$ and $f(X_2)$ exist, then there is always the same semantic correlation between $f(X_1)$ and X_1, on the one hand, and between $f(X_2)$ and X_2, on the other hand". (Mel'čuk, 1974)

Soon lexical-functional description turned out to be of great value for the systems of natural language processing. Different ways the LF description can be used in NLP system are described in (Apresjan et al., 2003). As far as machine translation is concerned, lexical functions play an important role in it, being used, in particular, for providing translation equivalents.

The mechanism of their usage is the following: if in one language (L1) X1 is an argument of the lexical function lf1, and lf1(X1)=Y1, and X1 has a translation equivalent X2 in another language (L2), and X2 is an argument of the same lexical function lf1, and lf1(X2)=Y2, then if in the process of translation from L1 to L2 a word-combination "X1+Y1" turns out to be a lexical-functional construction[2] representing lf1, X1 is replaced with X2, and Y1 is replaced with Y2

[2] Lexical-functional construction, or lexical-functional word-combination, is a word-combination consisting of an argument of a lexical function and a value of given lexical function for the same argument.

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 64–72
Manchester, August 2008

irrespective of the fact what trivial translation equivalent[3] it has.

(1) IncepReal1 (bus) = take
IncepReal1 (avtobus) = sadit'sja na
(*avtobus* – bus)
(sadit'sja na – sit on, *brat'* - take)

To take a bus is translated as *sadit'sja na avtobus*, not *brat' avtobus*.

In the system of machine translation ETAP-3 information about LF links between the words is stored in the dictionaries[4]. If a lexeme is an argument of one or several lexical functions, the list of these LFs is written in the dictionary entry of this lexeme along with the values of these LFs for this argument. Thus, dictionary entry of the word *bus* includes the following fragment:

IncepReal1: take

Such a way of storage allows the information about LF links between words to be easily used in the process of translation.

The described above mechanism of usage of lexical functions in the process of translation is very useful, but it can be implemented only if X1 and X2 are arguments of the same lexical function, and Russian and English do not provide such a correspondence in 100% of cases.

If such a direct correspondence between two languages does not exist, information about lexical functions can still be used for providing proper translation equivalents. In many cases "X1+Y1" and its translation equivalent are both LF constructions but representing different lexical functions.

The goal of this paper is to describe different types of such correspondences, to explain the reasons of their existence, and to show the ways they can be used in a machine translation system.

2 Translation and false homonymy

The first type of lexical-functional correspondences I would like to mention is described in an earlier paper (Andreyeva, 2007).

That paper is devoted to homonymous word-combinations which are lexical-functional at least in one of its meanings. It describes different types of homonymy, and one of them is so-called false homonymy.

This type of homonymy characterizes LF constructions which are not actually homonymous having only one meaning each, but every such construction can be described with the help of at least two lexical functions.

(2) *to conclude an agreement*

As I note in (Andreyeva, 2007), the word-combination from (2) can be described with the help of two LFs (IncepOper1 and CausFunc0), but the way of description does not change the meaning: "to begin to have an agreement" (IncepOper1) or "to cause an agreement to take place" (CausFunc0) are two descriptions of the same situation, not descriptions of two different situations.

In (Andreyeva, 2007) I show that (2) is not unique, there exist quite big groups of non-homonymous word-combinations which can also be described with the help of the same pair of lexical functions. These are, for example, arguments of LF IncepOper1 with the value *begin* (*to begin an argument, a battle, a struggle* and so on).

There exists also a much larger group of words denoting different objects that can be created this or that way (*to grow plants, to write music* etc). In this case the word-combinations can be described with the help of both CausFunc0 and Oper1.

The work also shows one more pair of lexical functions describing the same constructions - FinOper1 and LiquFunc0 (*to stop the battle*, for example).

In (Andreyeva, 2007) it is claimed that there are several reasons of the existence of false homonymy.

First, the descriptions of lexical functions are quite general and approximate. The creators of the system of LF did not have an aim to divide all the possible situations into non-crossing classes, the aim was to describe the main prototypical semantic correspondences.

Second, lexical functions were initially created for the description of situations. Their usage for the description of objects produced additional cases of false homonymy.

To sum it up, it is possible to list three pairs of lexical functions, and each pair can be used for the description of a non-homonymous word-combination:

[3] Trivial translation equivalent of a word from L1 is its default translations equivalent in L2, or translation equivalent this word from L1 has as a separate word.

[4] The system has a separate dictionary for every language.

№	English	↔	English
1.	IncepOper1 (X)	↔	CausFunc0 (X)
2.	Oper1 (X)	↔	CausFunc0 (X)
3.	FinOper1 (X)	↔	LiquFunc0 (X)

Table 1. False homonymy correspondences in the English language

All the examples in this section were given for the English language. The fact is that for the majority of their translation equivalents in Russian the situation is the same – being non-homonymous they can be described with the help of two lexical functions (the same ones as their English equivalents).

So, table 1 can be transformed into the following one:

№	L1	↔	L2
4.	IncepOper1 (X)	↔	CausFunc0 (X)
5.	Oper1 (X)	↔	CausFunc0 (X)
6.	FinOper1 (X)	↔	LiquFunc0 (X)

Table 2. False homonymy correspondences[5]

So, there are a lot of word-combinations in both languages that have only one meaning but can be described with the help of two different lexical functions. This fact is interesting from the point of view of theoretical semantics, but it causes difficulties for a machine translation system.

For simplification of the situation in (Andreyeva, 2007) I propose to use only one of these lexical-functional descriptions in every case. It is a good decision for every word-combination in particular, but being implemented for the system in general it still causes difficulties.

The fact is that the number of these word-combinations is quite big. The ETAP-3 system is developed by many linguists and it is changed all the time. In the cases described in this section there is no or nearly no difference between the meanings of lexical functions in the pairs, so each of these LFs can be chosen for the description of a word-combination. It is impossible to guarantee that the same functions will be chosen to describe translation equivalents. Besides, there can be a slight difference between the meanings of equivalents in two languages, and different functions can seem preferable for their description. If different lexical functions are used to describe translation equivalents, the information about their equivalency will be lost.

To avoid such a situation and to be free to choose the best corresponding lexical function for the description of an LF construction without taking into consideration the material of the other language it seems reasonable to add special rules to the ETAP-3 system. These rules allow to replace the value of a lexical function from the pair not only with the value of the same function but also with the value of the other LF from the same pair. The technique of this replacement is described in section 4.

3 Real lexical-functional correspondences

In the introduction there was given the general definition of lexical function given by Mel'čuk. According to it, to be regarded as an argument of the LF and a value of given LF for the same argument, two words must have a certain correlation between their meanings. But the majority of definitions of concrete lexical functions include not only semantic, but also syntactic conditions: to have a right to be called an argument of the LF and a value of given LF for the same argument, two words must also be connected by a certain syntactic link.

All the pairs of corresponding lexical functions in section 2 have differences in the semantic parts of their definitions, but the syntactic parts are absolutely identical. If they were not, it would be impossible to use these pairs for the description of the same constructions.

But it has already been mentioned that the reasons of adding these correspondences to the system of machine translation are mainly technical. What is really important for the system is the possibility to establish correspondences between different word-combinations, among which one can be described only with the help of lexical function lf1, and the other one represents only lexical function lf2.

Actually, such correspondences have already been described and implemented in the system of machine translation ETAP-3. See, for example, (Apresjan, Tsinman, 2002), where several dozens of such correspondences are listed, including quite rare ones. But the majority of these correspondences were used only in the paraphrasing block of the system, i.e. a block responsible for paraphrasing of sentences of one language only. As far as translation is concerned,

[5] L1 and L2 can be both the Russian and the English languages.

only a few LF correspondences were implemented in the translation process.

This section shows which real LF correspondences can be found between the Russian and the English languages and how they can be used to improve translation process.

3.1 LF correspondences types

It happens quite often that in one of the described languages X1+Y1 form an LF construction (with X1 as an argument and Y1 as a value of the LF lf1), and the translation equivalent of X1+Y1 in the other language represents some other lexical function. Sometimes it is X2+Y2 (with X2 as a translation equivalent of X1 and Y2 as a value of LF lf2 for X2), but it also happens that it is Y2 only.

In some cases such correspondences represent system differences in strategies two languages use. In these cases the same correspondence describes large groups of word-combinations. In other cases the correspondences are not caused by system differences, but nevertheless can be used for the processing of big groups of constructions. There are also situations when LF correspondences are specific for small groups of constructions.

3.2 System differences

Func - Oper

This subsection describes not a pair, but a whole class of lexical functions dealing with the idea of possession (in its widest meaning, of course).

It is common knowledge that Russian belongs to so-called "be-languages", and English is a "have-language". These characteristics of the languages could not help influencing the sphere of lexical functions describing possession.

Of course, both Russian and English have Func and Oper LF constructions of different types. But in the Russian language the number of LF constructions of Func type is approximately two times bigger than in the English language. In the majority of cases, if a Russian LF construction of the Func type cannot be translated into English with the help of the same LF, an Oper correspondence can be found.

This principle can be illustrated by the following list of corresponding pairs of lexical functions:

№	Russian	↔	English
1.	Func1 (X)	↔	Oper1 (X)
2.	IncepFunc1 (X)	↔	IncepOper1 (X)
3.	FinFunc1 (X)	↔	FinOper1 (X)
4.	Func2 (X)	↔	Oper2 (X)

Table 3. Some correspondences of Func type and Oper type lexical functions

Here are some illustrations for LF correspondences from table 3.

(3) Oper1 (boredom) = feel
Func1 (toska) = glodat'
(*toska* – boredom, *glodat'* – gnaw)

(4) IncepOper1 (impression) = gain
IncepFunc1 (vpechtlenije) = skladivat'sja u
(*vpechtlenije* – impression,
(*skladivat'sja u* – form itself at)

(5) FinOper1 (cold[6]) = shake off
FinFunc1 (nasmork) = prohodit' u
(*nasmork* – cold, *prohodit' u* – be over at)

(6) Oper2 (threat) = bear
Func2 (ugroza) = navisat' nad
(*ugroza* – treat, *navisat' nad* – hang over)

Examples (3)-(6) show pairs of translation equivalents among which the Russian one is an argument of the Func type lexical function from the corresponding pair of LFs and is not an argument of the LF of Oper type; on the other hand, the English word is an argument of the Oper type LF and is not an argument of the Func type one.

In the process of translation word-combinations formed by these arguments of LFs and values of these LFs for these arguments are replaced with each other. Translation equivalents for the material from examples (3)-(6) are the following:

(3a) Ego glojet toska. ↔ He is feeling boredom.
(4a) U nego skladivajetsja vpechatlenije, chto... ↔ He is gaining an impression that…
(5a) Nasmork u nego proshel. ↔ He shook off the cold.
(6a) Nad nim navisla ugroza iskluchenija. ↔ He bore the threat of exclusion.

[6] cold3 – a disease

Technical aspects of implementation of LF correspondences described in this section are given in section 4.

Nouns and gerunds

Another system difference between the English and the Russian languages that turns out to be important for the use of lexical functions in the translation process is the difference between forms Russian and English verbs have. In the English language there exists a verb form called gerund which has no analogues in Russian.

The way this difference influences the domain of LF constructions is the following. Among lexical functions a lot (more than a nundred) have nouns as their arguments and verbs (sometimes accompanied by prepositions or adverbs) as their values. For example, all LFs mentioned above belong to this group.

Many of these verbs form verbal nouns which in the majority of cases inherit the meaning of the verbs. Therefore, if a verb V is a value of lexical function lf1 for the argument X (lf1(X) = V), and V forms a verbal noun N_V, in the majority of cases there will be the same semantic correlation between X and V, on one hand, and X and N_V, on the other hand.

But syntactic links between a verb and a noun (X+V) and a noun and a noun (X+N_V) are of course different, so word-combinations formed by two nouns cannot be described as representing the same lexical functions as word-combinations formed by a noun and a verb.

This problem was solved by making a special group of lexical functions for the description of word-combinations formed by a verbal noun and another noun. If lf1(X) = V, and semantic correlation between the meanings of V and X and of N_V and X is the same, then N_V = S0_lf1 (X). For example, see (7).

(7) IncepOper1 (compromise) = arrive at
S0_IncepOper1 (compromise) = arrival at

Lexical-functional constructions of S0_lf type exist both in the Russian and the English languages, but their number is bigger in Russian. The reason for this seems to be the fact that the English language has gerund which can be used in a function of a noun. Russian has no verb form of this kind. So, it has to form verbal nouns which are actually formed quite freely. English, being able to use gerunds in many constructions, does not need such a big number of verbal nouns.

The result of this difference is that in many cases Russian LF word-combination representing a lexical function of S0_lf type corresponds to the English construction with gerund. In order not to lose the information about lexical-functional correspondences it is possible to establish the following correlation:

S0_lf (in Russian) → lf (in English) (gerund)

Such a correspondence can be established for all (or nearly all) the lexical functions of S0_lf type. The article is too small to list them all[7], so only several examples are given:

№	Russian	→	English
1.	S0_Oper1 (X)	→	Oper1 (X) (gerund)
2.	S0_IncepOper1 (X)	→	IncepOper1 (X) (gerund)
3.	S0_FinOper1 (X)	→	FinOper1 (X) (gerund)

Table 4. Some correspondences of S0_lf type and nonS0_lf type lexical functions

Here are some illustrations for LF correspondences from table 4:

(8) S0_Oper1 (sport) = zanatija
Oper1 (sport) = go in for
(*sport* – sport, *zanatija* – work)

(9) S0_IncepOper1 (soglashenije) = dostizenije
IncepOper1 (agreement) = arrive at
(*soglashenije* – agreement)
(*dostizenije* - reaching)

(10) S0_FinOper1 (biznes) = uhod iz
FincOper1 (business) = go out of
(*biznes* – business)
(*uhod iz* = going away from)

Translation equivalents for examples (8)-(10) are:

(8a) zanatija sportom → going in for sports
(9a) dostijenije soglashenija → arriving at an agreement
(10a) uhod iz biznesa → going out of business

[7] There are more then 100 LF of S0_lf type in the ETAP-3 system.

3.3 Other differences

Lexical-functional correspondences described in the previous section are the result of system differences between the Russian and the English languages. This section is devoted to LF correspondences which are not caused by difference in strategies these languages use, but which are still applicable to large amount of word-combinations.

These LF correspondences consist of functions from Func group.

Lexical functions of Func type include, among others, Func0 and Func1. Func0 describes situations when X takes place, Func1 describes situations when X takes place for something/somebody or characterizes something/somebody (X is an argument of LF and a grammatical subject, something/somebody is a principal complement)[8].

There are a lot of cases when a word in one of the languages (X1) is an argument of Func1, and its translation equivalent in the other language (X2) is an argument of Func0 and is not an argument of Func1.

In such cases the result of the translation would be much better if we replace the value of Func1 for X1 with the value of Func0 for X2 than if we replace it with the trivial translation equivalent of X1. So, there can be established the following lexical-functional correspondence: Func1→Func0.

It is important to note that, unlike all the other "real" LF correspondences described above, this correspondence works in both Russian-English and English-Russian translation.

Here is the example for the described LF correspondence:

(11) Func0 (anger) = reign
Func1 (gnev) = vladet'
(*gnev* – anger, *vladet'* – possess)
Im vladeet gnev. – Anger reigns.

There is one more reason for the establishment of Func1→Func0 LF correspondence. A lot of words are arguments of both Func0 and Func1. But in many cases information about one of these links is not yet included in the system by mistake, or by chance, or because of the lack of time. In this case Func1→Func0 correspondence works as a technical one, not being able to provide the best translation result, but making it as good as possible.

Func1→Func0 is not the only lexical-functional correspondence of Func type. Other LF correspondences for lexical functions of Func group can be established. This is the list of them:

№	L1	→	L2
1.	IncepFunc1 (X)	→	IncepFunc0 (X)
2.	FinFunc1 (X)	→	FinFunc0 (X)
3.	CausFunc1 (X)	→	CausFunc0 (X)
4.	LiquFunc1 (X)	→	LiquFunc0 (X)

Table 5. Correspondences of Func type lexical functions

Here are some illustrations for LF correspondences from table 5:

(12) IncepFunc0 (doubt) = arise
IncepFunc1 (somnenije) = voznikat' u
(*somnenije* – doubt, *voznikat' u* – appear at)

(13) FinFunc0 (doubt) = disappear
FinFunc1 (somnenije) = pokidat'
(*pokidat'* – leave)

(14) CausFunc0 (indignation) = arouse
CausFunc1 (vozmuschenije) = vyzyvat' u
(*vozmuschenije* – indignation)
(*vyzyvat' u* – cause at)

(15) LiquFunc0 (confidence) = shatter
LiquFunc1 (doverije) = podryvat'
(*doverije* – confidence)
(*podryvat'* – undermine)

Translation equivalents for examples (12)-(15) are:

(12a) U nego voznikajet somnenije. →
Doubt arises.

(13a) Somnenija pokidajut ego. →
Doubts leave him.

(14a) Eto vyzyvajet vozmuschenije u vseh. →
This arouses everybody's indignation.

(15a) Eto podryvajet doverije ludej. →
It shatters people's confidence.

3.4 Rare correspondences

Despite the differences, all lexical-functional correspondences described above have one common feature: they take a word-combination

[8] These definitions of lexical functions were created by Ju.D. Apresjan (Apresjan, Tsinman, 2002).

X1+lf1(X1) and transform it into a word-combination X2+lf2(X2).

But the situation is not always that simple. Let us look at the following examples:

(16) carry convinction
(17) privodit' v izumlenije[9]
(18) prihodit' v izumlenije

All the examples (16)-(18) represent different lexical functions.

(16a) CausFunc0 (conviction) = carry
(17a) CausOper1 (izumlenije) = privodit'
(18a) IncepOper1 (izumlenije) = prihodit'
(*izumlenije* – astonishment)
(*prihodit'* – come, *privodit'* – lead)

The fact is that none of the examples (16)-(18) can be translated with the help of any of LF correspondences described above in this article. These word-combinations are transformed in the process of translation into one verb. Fortunately this verb is a value of CausV0 for the translation equivalent of X1, so it is possible to establish the following LF correspondences:

№	L1 →	L2
1.	CausFunc0 (X) + X →	CausV0 (X)
2.	CausOper1 (X) + X →	CausV0 (X)
3.	IncepOper1 (X) + X →	CausV0 (X) (passive voice)

Table 6. Some rare LF correspondences

Here are translation equivalents for table 6.

(16b) to carry conviction → ubejdat'
(*ubejdat'* – convince)
(17b) privodit' v izumlenije → to astonish
(18b) prihodit' v izumlenije → to be astonished

4 Mechanism of translation

In sections 2 and 3 different types of lexical-functional correspondences were described. This chapter shows how the use of these LF correspondences is realised in the system of machine translation ETAP-3.

"Linguistically, ETAP-3 consists of various sets of rules and dictionaries… All the rules … are subdivided into three main types: (i) general rules that apply to all the sentences in the course of their processing; (ii) class-specific rules that hold for compact groups of words and are referred to by their names in the dictionary entries of the respective items; (iii) word-specific rules that are characteristic of individual lexical items and are stored directly in their dictionary entries. The second and third types of rules are activated only on condition that the processed sentence contains the relevant lexical items." (Apresjan et al, 2003). As for general rules, it is important to note that they work one after another, in the fixed order, so the order they are listed in the system is very important.

To implement the above-described lexical-functional correlations we have to include them into the system in a form of translation rules. It means two main problems to be solved:
1) what type these rules must belong to,
2) if they are general, what their order must be.

As for the type, the decision seems to be the following: all the correspondences except ones from section 3.4 (rare correspondences) must become general rules, and those from 3.4 must become class-specific ones. The latter are very rare and can be implemented, perhaps, only for several words each. There is no use in making them general, and they must not be word-specific, too, because they describe groups of constructions, not singular cases. So, class-specific type is ideal for them.

All the other types of LF correspondences described above are worth being implemented with the help of general rules. First, they describe big groups of constructions. Second and the most important is the reason for they usage: we need them to work automatically in case the main rule of translation with the help of a lexical function (described in the introduction) does not work, and the only way to provide this is to make them general.

As these correspondences are implemented with the help of general rules, it is very important to put them into the proper order. I would like to propose the following one.

1) The first rule in the list of rules responsible for the translation with the help of LFs is of course the main rule described in introduction. All the other rules can work only in case the first one did not work. So, the first rule is:

lf1 (X) → lf1 (X)

[9] Russian examples (17) and (18) will be translated below.

2) The second block of rules is the block responsible for Func-Oper correspondences. In case the general rule does not work this block provides the most correct translation equivalents. So, the second block is:

№	Russian	↔	English
1.	Func1 (X)	↔	Oper1 (X)
2.	IncepFunc1 (X)	↔	IncepOper1 (X)
3.	FinFunc1 (X)	↔	FinOper1 (X)
4.	Func2 (X)	↔	Oper2 (X)

The order of rules inside this block (as well as inside all the other blocks) is not of great importance. It can be the same as in the table.

3) The third block is the one responsible for correspondences of Func type.

№	L1	→	L2
1.	Func1 (X)	→	Func0 (X)
2.	IncepFunc1 (X)	→	IncepFunc0 (X)
3.	FinFunc1 (X)	→	FinFunc0 (X)
4.	CausFunc1 (X)	→	CausFunc0 (X)
5.	LiquFunc1 (X)	→	LiquFunc0 (X)

It is very important for the third block to be implemented only after the second one, because there are lexical functions both blocks work with (Func1, IncepFunc1, and FincFunc1). If it is impossible to replace the value of one of these functions with the value of the same one, we must first try to replace it with its Oper equivalent and only in case it is impossible pass to Func0 correspondence. Oper equivalent is better than Func0 one because the former allows to preserve the information about all the actants of the verb – value of an LF, while the latter loses the information about one of the actants.

4) The fourth block is the one transforming nouns into gerunds.

№	Russian	→	English
1.	S0_Oper1 (X)	→	Oper1 (X) (gerund)
2.	S0_IncepOper1 (X)	→	IncepOper1 (X) (gerund)
3.	S0_FinOper1 (X)	→	FinOper1 (X) (gerund)

As for the order of the third and the fourth blocks, they do not interfere with each other so it is of no importance which one is the first. The only problem is that in reality the fourth block is very big (it includes the majority of LFs of S0_lf type), so it is just more convenient to have it after the third block.

5) And the last one is the block responsible for false homonymy correspondences.

This block was created "just in case", so it is worth being placed at the end of the list. Besides, it works with Oper functions of different types, so in any case it must be placed after the second block.

This block causes one additional problem. If we list all its correspondences in one column, we will see that CausFunc0 can become both Oper1 and IncepOper1. Establishment of both of these rules in the system will not improve the translation but will produce plenty of wrong translation variants. As the rules of this block are not of great importance, it seems better not to use these two problem rules at all. So, this block will be the following:

№	L1	→	L2
1.	IncepOper1 (X)	→	CausFunc0 (X)
2.	Oper1 (X)	→	CausFunc0 (X)
3.	FinOper1 (X)	→	LiquFunc0 (X)
4.	LiquFunc0 (X)	→	FinOper1 (X)

In the majority of cases the transformation of one LF construction into the other one entails changes in syntactic roles of actants. Information about these changes is also included into the rules.

5 Conclusion

The paper described different types of lexical-functional correspondences between the Russian and the English languages. It showed how the information about LF links included in the dictionaries and translation rules of machine translation system ETAP-3 allowed to consider these correspondences in the process of translation and thus to improve its results.

References

Andreyeva, A.S. 2007. Lexical Functions and Homonymy. *MTT-2007, Proceedings of the 3rd International Conference on Meaning-Text Theory.* Wiener Slawistischer Almanach, Sonderband 69. München – Wien

Apresjan, Jury D., Igor M. Boguslavsky, Leonid L. Iomdin, Alexander V. Lazursky, Vladimir Z. Sannikov, Victor G. Sizov, and Leonid L. Tsinman. 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. *MTT 2003, First International Conference on Meaning – Text Theory.* Paris, Ecole Normale Superieure, Paris, 279-288

Apresjan, Jury D., and Leonid L. Tsinman. 2002. Formal'naja model' perifrazirovanija predlozhenij dlja sistem pererabotki tekstov na estestvennyx jazykax [A Formal Model of Sentence Paraphrasing for NLP Systems]. *Russkij jazyk v nauchnom osveshchenii*, No. 4, pp. 102-146

Mel'čuk, Igor 1974. *Opyt teorii lingvisticheskix modelej "Smysl ⇔ Tekst"* [A Theory of Meaning ⇔ Text Linguistic Models"]. Moscow, Nauka.

Mel'čuk, Igor, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, and Adèle Lessard. 1984. *Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques I.* Les Presses de l'Université de Montréal.

Mel'čuk, Igor, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, and Suzanne Mantha. 1988. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II.* Les Presses de l'Université de Montréal.

Mel'čuk, Igor, Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, and Suzanne Mantha. 1992. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III.* Les Presses de l'Université de Montréal.

Mel'čuk, Igor, and Alexander Zholkovskij. 1984. *Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka.* [An Explanatory Combinatorial Dictionary of the Contemporary Russian Language] Wiener Slawistischer Almanach, Sonderband 14.

The "Close-Distant" Relation of Adjectival Concepts Based on Self-Organizing Map

Kyoko Kanzaki, Hitoshi Isahara
National Institute of Information and Communications Technology
3-5, Hikaridai, Seikacho,
Sorakugun, Kyoto, 619-0289,
Japan
{kanzaki,isahara}@nict.go.jp

Noriko Tomuro
School of Computer Science, Telecommunications and Information Systems
DePaul University
Chicago, IL 60604
U.S.A
tomuro@cs.depaul.edu

Abstract

In this paper we aim to detect some aspects of adjectival meanings. Concepts of adjectives are distributed by SOM (Self-Organizing map) whose feature vectors are calculated by MI (Mutual Information). For the SOM obtained, we make tight clusters from map nodes, calculated by cosine. In addition, the number of tight clusters obtained by cosine was increased using map nodes and Japanese thesaurus. As a result, the number of extended clusters of concepts was 149 clusters. From the map, we found 8 adjectival clusters in super-ordinate level and some tendencies of similar and dissimilar clusters.

1 Introduction

This paper aims to find a diversity range of adjectival meanings from a coordinate map in which "close-distant" relationships between adjectival classes is reflected. In related research over adjectives, Alonge et.al (2000), Solar (2003), Marrafa and Mendes (2006) suggested that WordNet and EuroWordNet lack sufficient adjectival classes and semantic relations, and extended the resources over such relations.

For the sake of identifying the diversity of adjectival meanings, it is necessary to analyze adjectival semantics via "close-distant" relationships extracted from texts. In our work on extracting adjective semantics, we consider abstract nouns as semantic proxies of adjectives. For the clustering method, we utilized a self-organizing map (SOM) based on a neural network model (Kohonen, 1997). One of the features of SOM is that it assigns words coordinates, allowing for the possibility of visualizing word similarity. SOM has two advantages for our task. One is that we can utilize the map nodes of words to locate members of clusters that clustering methods have failed to classify. The other is that the map shows the relative relations of whole clusters of adjectival concepts. By observing such a map in which the relations of clusters are reflected, we can analyze the diversity of adjectival meaning.

2 Abstract Nouns that Categorize Adjectives

Collocations between adjectives and nouns in "concrete value and its concept" relations can be used to represent adjectival semantics. Nemoto (1969) indicated that expressions such as "*iro ga akai* (the color is red)" and "*hayasa ga hayai* (literally, the speed is fast)" are a kind of tautology. Some studies have suggested that some abstract nouns collocating with adjectives are hypernymic concepts (or concepts) of those adjectives, and that some semantic relations between abstract nouns and adjectives represent a kind of repetition of meaning.

This paper defines such abstract nouns as the semantic categorization of an adjective (or an adjectival concept).

The data for this study was obtained by extracting adjectives co-occurring with abstract nouns in 100 novels, 100 essays, and 42 years of newspaper articles.

We extracted the abstract nouns according to the procedure described by Kanzaki et.al (2006). Here, they evaluated the category labels of adjectives obtained by the proposed procedure and found that for 63% of the adjectives, the ex-

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 73–76
Manchester, August 2008

tracted categories were found to be appropriate. We constructed a list as follows:

Abstract Nouns:
Adjectives modifying abstract nouns
KIMOCHI (feeling):
ureshii (glad), *kanashii* (sad),
shiawasena (happy) …

In this list, "*KIMOCHI* (feeling)" is defined by "*ureshii* (glad), *kanashii* (sad), and *shiawasena* (happy)", for example. Here, each abstract noun conveys the concept or hypernym of the given adjectives.

Next we classify these abstract nouns based on their co-occurring adjectives using SOM.

3. A Map of Adjective Semantics

3.1 Input Data

In our SOM, we use adjectives which occur more than four times in our corpus. The number of such adjectives was 2374. Then we identified 361 abstract nouns that co-occurred with four or more of the adjectives. The maximum number of co-occurring adjectives for a given abstract noun in the corpus was 1,594.

In the data, each abstract noun was defined by a feature vector, in the form of noun co-occurrences represented by *pointwise mutual information* (Manning and Schutze, 1999). Mutual information (MI) is an information theoric measure and has been used in many NLP tasks, including clustering words (e.g. Lin and Pantel, 2002).

3.2 SOM

Kohonen's self-organizing map (SOM) is an unsupervised learning method, where input instances are projected onto a grid/map of nodes arranged in an *n*-dimensional space. Input instances are usually high-dimensional data, while the map is usually two-dimensional (i.e., $n = 2$). Thus, SOM essentially reduces the dimensionality of the data, and can be used as an effective tool for data visualization – projecting complex, high-dimensional data onto a low-dimensional map. SOM can also be utilized for clustering. Each node in a map represents a cluster and is associated with a reference vector of *m*-dimensions, where *m* is the dimension of the input instances. During learning, input instances are mapped to a map node whose (current) reference vector is the closest to the instance vector (where SOM uses Euclidean distance as the measure of similarity by default), and the reference vectors are gradually smoothed so that the differences between the reference vector and the instance vectors mapped to the node are minimized. This way, instances mapped to the same node form a cluster, and the reference vector essentially corresponds to the centroid of the cluster.

SOM maps are self-organizing in the sense that input instances that are similar are gradually pulled closer during learning and assigned to nodes that are topographically close to one another on the map. The mapping from input instances to map nodes is one-to-one (i.e., one instance is assigned to exactly one node), but from map nodes to instances, the mapping is one-to-many (i.e., one map node is assigned to zero, one, or more instances).

The input data was the set of 361 abstract nouns defined by the 2,374 co-occurring adjectives, as described in the previous section. These abstract nouns were distributed visually on the 2-dimensional map based on co-occurring adjectives. This map is a "map of adjective semantics" because the abstract nouns are identified as proxies for adjective semantics.

As mentioned before, similar words are located in neighboring nodes on the 2-dimensional map. The next step is to identify similar clusters on the map.

4. Clusters of Adjective Semantics

4.1 Tight Clusters from the Map Nodes

In SOMs, each node represents a cluster, i.e. a set of nouns assigned to the same node. These nouns are very similar and can be considered to be synonyms. However, nouns that are similar might map to different nodes because the algorithm's self-organization is sensitive to the parameter settings. To account for this, and also to obtain a more (coarse-grained) qualitative description of the map, tight clusters—clusters of map nodes whose reference vectors are significantly close—were extracted. All groupings of map nodes whose average cosine coefficient between the reference vectors in the group was greater than 0.96 were extracted (Salton and McGill, 1983).

4.2 Result

The total number of clusters was 213. Excluding singleton clusters, the number of clustes was 81. 229 concepts were classified into 81 clusters, with 132 concepts not classified into any cluster.

In order to evaluate the quality of the conceptual classification, we utilized the "*Bunruigoihyou"* Japanese thesaurus (National Institute of Japanese Language, 1964). In "*Bunruigoihyou,*" each category is assigned a 5-digit category number, with close numbers indicating similar categories.

Among the 81 with two or more concepts, the number of clusters containing words with the same class was 36. That is, for 44% of the clusters, the constituent nouns had the same "*Bunruigoihyou*" class label. The ratio of concept agreement between "*Bunruigiohyou*" and our obtained clusters was found to be 20.87/81=0.25. We also compared tight clusters by performing hierarchical clustering with the *k*-means algorithm.

The results of the hierarchical clustering were as follows:

1) The rate of clusters agreeing with "*Bunruigoihyou*": 30/96 = 0.31
2) The average rate of agreement for each tight cluster: 21.07/96 = 0.21

In the case of *k*-means:

3)The rate of clusters agreeing with "*Bunruigoihyou*": 33/143 = 0.23
4) The average rate of agreement for each tight cluster: 28.37/143 = 0.198

From these results, we can observe that clusters obtained with cosine similarity agree more with the Japanese thesaurus than the other two methods. Therefore, in terms of quality, clusters obtained by cosine similarity seem to be superior to the others.

4.3 Using the Position of Map Nodes

However, even for the result obtained with cosine similarity, 132 concepts were not classified into any clusters. Additionally, the clusters appear to be overly fine grained: most tight clusters include 1, 2 or 3 concepts. In order to find similar concepts that cosine similarity failed to cluster together, we used the position information of the map nodes.

After we plotted clusters obtained by cosine similarity on the map, we checked for singleton concepts located near a cluster which are members of the same "*Bunruigoihyou*" class. Also, we checked to see if concepts in clusters located at neighboring nodes could be clustered together using the category numbers of "*Bunruigoihyou. "*

By extending the clusters, we generated a total of 149 clusters, including 68 with two or more elements and 81 singleton clusters.

5. Interpreting the Adjectival Clusters

In our final map, 361 concepts were distributed based on 2374 adjectives into 149 clusters. Among the 149 clusters, 68 contained two or more concepts.

5.1 "Close-Distant" Relations of Clusters and Adjectives

In the final map, clusters at the superordinate level are located around the center of the map. Upper level concepts tend to agree with clusters in "Bunruigoihyou." For examples, "image and impression," "situation and state", "feeling and mood" are located around the center of the map.

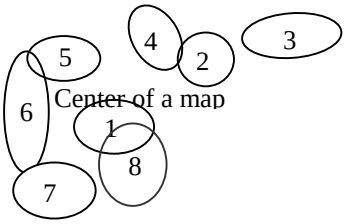


Fig7. Cluster 7 on the map

Cluster1 (Center of the map): *koto* (matter), *in'shou* (impression), *men* (side of something or someone), and *kankaku* (sense/feeling)
Cluster2: *seishitsu* (characteristics of someone/something), *yousou* (aspect)
Cluster3: *kanten* (viewpoint), tachiba (standpoint), *bun'ya* (domain)
Cluster4: *taido* (attitude), *yarikata* (way of doing)
Cluster5: *gaikan*, *gaiken*, *sugata* (outlook and appearance of someone/something)
Cluster6: *fun'iki*, *kuuki*, *kehai* (atmosphere)
Cluster7: *kimochi*, *kanji* (feeling)
Cluster8: *joutai* (state), *joukyou* (situation)

In our experiment, at the top level, adjectival concepts seem to be divided into 8 basic clusters. From the distribution of the map, we find "close-distant" relationships between clusters, that is clusters located far from each other tend to be semantically disparate. In terms of adjective semantics, the semantic relationship between "*kimochi, kanji* (feeling)" (Cluster7) and "*seishitsu* (characteristics of someone/something), *yousou* (aspect)" are distant.

However, "*kimochi, kanji* (feeling)" (Cluster7) has a close relation to "*fun'iki, kuuki, kehai* (atmosphere) " (Cluster6) and also "*joutai* (state), *joukyou* (situation)" (Cluster8).

1. In our experiment, 77 adjectives belonged to one or two clusters. Though there is the possibility of data sparseness, there is also the possibility that the meanings of these adjectives are specific. Examples of adjectives belonging to specific clusters are as follows:

Adjectives in distant relationships;
- Clusters 2: *keisandakai* (seeing everything in terms of money), *ken'meina* (wise), …
- Cluster 7: *akkenai* (disappointing/easily), *kiyasui* (feel at home),…

Adjectives in close relationships;
- Cluster 6: *ayashigena* (fishy)
- Cluster7: *akkenai* (disappointing /easily), *kiyasui* (feel at home)
- Cluster8: *meihakuna* (obvious), *omoshiroi* (interesting), *makkurana* (dark)

Japanese adjectives are often said to represent "kanjou (mental state)", "joutai (state)," "seisitsu (characteristics)" and "teido (degree)", in addition to "positive/negative image." In our experiment, the SOM unearthed not only these adjectival meanings, but also "inshou (impression)", "taido (attitude)", "kanten (viewpoint)" and "sugata (outlook)", which seem to be discriminative meanings of adjectives.

6. Future work

We classified 361 concepts based on 2374 adjectives using a self-organizing map. Since the SOM shows the distribution visually, it provides not only clusters of adjectives but also "close-distant" relationships between clusters. As a result, adjectival concepts at the superordinate level are divided into 8 main clusters. The results not only verify previous work but also suggest new discriminative adjective classes. One of the advantages of SOM is that it presents its outputs visually. As a result, we can explore "close- distant" relationships between clusters, and analyze the meaning of each. In addition to increasing the range of adjectival classes and improving our method, our method provides the means to analyze concepts which did not agree with those in existing thesauri such as "Bunruigoihyou", the EDR dictionary or Japanese Word Net.

References

Alonge, Antonietta., Francesca Bertagna, Nicoletta Calzolari, Andriana Roventini and Antonio Zampolli. 2000. Encoding Information on Adjectives in a Lexical-semantic Net for Computational Applications, *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-00)* :42-49

Kyoko Kanzaki,Qing Ma, Eiko Yamamoto and Hitoshi Isahara, 2006, Semantic Analysis of Abstract Nouns to Compile a Thesaurus of Adjectives, *In Proceedings of The International Conference on Language Resources and Evaluation (LREC-06)*

Kohonen, Teuvo. 1997. *Self-Organizing Maps, Second Edition*, Springer.

Lin, Dekang., and Patrick Pantel. 2002. Concept Discovery from Text, *Proceedings of the 19th International Conference on Computational Linguistics(COLING-02)*: 768-774

Manning, Christopher D., and Hinrich Shütze. 1999. *Foundations of Statistical Natural language Processing*, The MIT Press.

Marrafa, Palmira., and Sara Mendes. 2006. Modeling Adjectives in Computational Relational Lexica, *Proceedings of the COLING/ACL2006*:555-562

National Institute for Japanese Language. 1964. *Bunruigoihyou* (Word List by Semantic Principles).

Nemoto, Kesao. 1969. The combination of the noun with "ga-Case" and the adjective, *Language research 2 for the computer*, National Language Research Institute: 63-73 (in Japanese)

Salton, Gerard., and Michael J. McGill. 1983. Introduction to Modern Information Retrieval. McGraw Hill.

Solar, Clara. 2003. Extension of Spanish WordNet, Proceedings of the third International WordNet Conference(GWC-06):213-219

Looking up phrase rephrasings via a pivot language

Aurélien Max
LIMSI-CNRS & Université Paris-Sud 11
Orsay, France
aurelien.max@limsi.fr

Michael Zock
LIF-CNRS
Marseilles, France
michael.zock@lif.univ-mrs.fr

Abstract

Rephrasing text spans is a common task when revising a text. However, traditional dictionaries often cannot provide direct assistance to writers in performing this task. In this article, we describe an approach to obtain a monolingual phrase lexicon using techniques used in Statistical Machine Translation. A part to be rephrased is first translated into a pivot language, and then translated back into the original language. Models for assessing fluency, meaning preservation and lexical divergence are used to rank possible rephrasings, and their relative weight can be tuned by the user so as to better address her needs. An evaluation shows that these models can be used successfully to select rephrasings that are likely to be useful to a writer.

1 Introduction

Once an initial draft of a text is ready, writers face the difficult phase of *text revision*. Changes may be made for various reasons: correcting spelling or grammatical errors, making the text locally more fluent (for example, in case it contains wordings that are literal translations from another language), avoiding close repetitions or enforcing terminological consistency, or better conveying the writer's ideas. All these changes can affect text spans of various sizes, and can globally be seen as cases of *rephrasing*. Paraphrasing involves rephrasings that are semantically equivalent, but targets terminology and style that are more suited to the context of use of a text. In a broad sense, rephrasing may involve wordings that convey different meanings in an attempt to correct or make the writer's thoughts more precise. Research concerned with the study of changes between writers' drafts (*textual genetic criticism*) can help in understanding writers' rewriting processes, and can be supported by automatic tools (e.g. (Bourdaillet et al., 2007)).

In this work, we address the issue of how writers can be assisted in finding wordings that correspond to multi-word phrases of any nature. Given an original text span, the writer is presented with a list of rephrasings that are organized by taking into account the context of the rephrasing and user-specified preferences. Our proposal can therefore be used as a lexicon operating at the phrasal level, which can be used either when writers are faced with a tip-of-the-tongue lexical access problem, or when they are not completely satisfied with some initial wording. In the former case, they may be able to come up with some words or phrases that would be different in meaning from what they are looking for, and in the latter they may be looking for a near-synonymous wording that is more appropriate to a given context, for example to avoid close repetitions. To define such a phrase lexicon and its possible mode of use, the following questions should be considered: (a) how the lexicon entries are obtained, (b) what can be the entry points and how can one navigate in the results, and (c) how the results are displayed.

Rephrasing can be more or less complex and problematic depending on the consequences at the various levels:

- In the simplest case, replacing one element

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 77–85
Manchester, August 2008

by another does not have any consequences overall. This is often the case when a word is replaced by its synonym or a similar word.

- An entire expression or sentence is replaced by its equivalent. In this case the problem is generally to obtain a good fit with regard to the surrounding text, the replacing unit being well-formed by definition.

- The replacing element may require syntactic changes of the matrix, i.e. the text in which it is embedded. This occurs if the source word and the target word have different syntactic requirements, and this can be seen as a good reason to replace entire sentences, or at least sentence fragments. This assumes a pattern dictionary, where patterns achieving the same conceptual goal are grouped together.

In the next section, we discuss limitations of traditional dictionaries with respect to the targeted task, and describe an approach to obtain phrase rephrasings through a pivot translation into another language. In section 3, we discuss the issue of the organization of the results along various axis: fluency of rephrasings, preservation of meaning, and lexical divergence between original text spans and rephrasings. We then present an initial evaluation of our approach on French rephrasing in section 4. Related work is presented in section 5, and we finally discuss our approach and our future work in section 6.

2 Lexicon of phrase rephrasings

Dictionaries and semantic resources such as thesauri can be used to find words by following links of different kinds from a given entry point. WordNet (Fellbaum, 1998) is one such resource. For a proposal of other kinds of links and navigational aids see also (Zock and Bilac, 2004; Zock, 2006; Zock, 2007).

Words are the traditional units that people expect to find in dictionaries. Whereas some types of dictionaries can contain multiword expressions, such as compound nouns and terms, those correspond to linguistically-motivated units. In order to rephrase phrases of any type with a dictionary, a writer may have to look up several words, combine various information and validate the result using her experience of the language or throught the use of a concordancer. Moreover, dictionary lookups are in most cases insensitive to the actual context of words in an existing text. It is therefore the responsibility of its users to ensure that a choice is appropriate for a given context, which can be quite difficult, for example when writing in a second language.

One way of obtaining phrase rephrasings is by looking at phrases that occur in similar contexts in a monolingual corpus (e.g. (Munteanu and Marcu, 2006)). In order to extract a comprehensive phrase lexicon, a very large number of sentences should be compared to extract potential rephrasings, which furthermore may often correspond to phrases that are too remotely connected. Parallel corpora provide the interesting advantage that it is reasonable to assume that elements from one side of the corpus should be aligned to elements on the other side, and that associations of elements can be reinforced by the number of times they occur in the corpus. Various approaches for word alignment from parallel corpora have been proposed (see e.g. (Och and Ney, 2003)), and the phrase-based approach to Statistical Machine Translation (Koehn et al., 2003) has led to the development of heuristics for obtaining alignments between phrases of any number of words.

Unfortunately, monolingual parallel corpora aligned at the sentence level, such as various translations of a novel in a foreign language, are resources that are extremely scarce. Using bilingual parallel corpora, a much more common resource, one can obtain various possible phrase translations for a given source phrase, as well as some estimate of the distribution of probabilities for the various translations of that phrase. Such $N \rightarrow M$ alignements can capture lexical translations (e.g. *exigeons → ask for, call for, demand, expect, request,* etc.) and phrasal literal or idiomatic translations (e.g. *un bon début → a good approach, a good first move, a good starting point, a positive initiative, an encouraging start, the right road*, etc.), but can also capture noise depending on the alignment heuristics used (e.g. *les états candidats (candidate countries) → Member States, the candidate countries were to, the accession countries have called for, candidate, the,* etc.) Different target phrases associated with a given source phrase can either represent paraphrases or phrases with different meanings. Among the limitations of this type of phrasal alignments are their inability to model non-consecutive words and to generalize the con-

tents of phrases, and the fact that their translations are not conditioned on their context.

If phrase extraction is performed in two opposite directions, then it is possible to find the possible translations of a given phrase (and their conditional probabilities), and then to translate back those phrases into the original language. In this approach proposed by (Bannard and Callison-Burch, 2005), the second language acts as a pivot, as illustrated on figure 1. Because of the nature of the possible alignments, this pivot can represent various senses, which in context can be equivalent or comparable to that of the original phrase. In turn, the same phenomena can take place when translating back from the pivot phrases to the original language, and the resulting rephrasings can be equivalent or comparable in meaning to that of the original phrase in some context, may also be incomplete and/or require other changes in the rephrased sentence.

Bannard and Callison-Burch have defined a *paraphrase probability* between two phrases p_1 and p_2 (with $p_1 \neq p_2$) that uses conditional probabilities between phrases and sums over all possible pivot phrases:

$$P(p_2|p_1) = \arg\max_{p_2 \neq p_1} \sum_{pivot} P(pivot|p_1)P(p_2|pivot) \quad (1)$$

(Callison-Burch, 2007) measured the importance of various factors impacting the quality of the paraphrases obtained. Using manually built alignments yields a significant improvement in paraphrase quality, showing that if better alignments are available the proposed approach can produce better paraphrases. Alignments between several languages can be used for finding pivot phrases, and using several simulateously tend to improve alignment quality and therefore paraphrases themselves. Using a language model to find paraphrases that maximize its score in the original sentencial context leads to improved fluency, but has a negative impact on meaning preservation. Lastly, restricting pivot phrases to those actually aligned in a test aligned bilingual corpus improves paraphrase quality, which illustrates the importance of disambiguating source phrases relatively to the pivot language.

The rephrasings obtained can be classified into several categories when used in context:

- A rephrasing can be a paraphrase that is valid in all contexts (e.g. ***je vous donne raison*** → ***je suis d'accord avec vous***), in specific grammatical contexts (e.g. *pouvoir accueillir* ***dans de bonnes conditions*** *les pays* → ***comme il se doit***) and/or pragmatic contexts (e.g. ***c'est un bon début*** → ***nous partons du bon pied***).

- A rephrasing can contain shifts in meaning with the original phrase which might be acceptable or not (e.g. *nous voulons* ***apporter notre contribution*** *à ce débat* → ***donner de la valeur***). Some such rephrasings reveal a natural bias towards the bilingual corpus used (e.g. *le prochain élargissement constitue* ***la principale tâche*** → ***l' objectif principal***).

- A rephrasing can be ill-formed but still contain elements of interest to a writer (e.g. *ceux qui disent que ...* ***se trompent*** → ***devrions à nouveau réfléchir***; here a rephrasing such as ***devraient à nouveau réfléchir*** could be deemed acceptable in some contexts).

- A rephrasing may introduce a contradiction in a specific context (e.g. ***ce n'est pas le moment de*** *se montrer hésitant* → ***il est trop tôt pour***)

- A rephrasing may be inexploitable because it is syntactically ill-formed in context and does not contain any element of interest, or is too close to the original phrase.

The most natural entry point to such a resource is by entering a phrase or selecting it in a text under revision. Approximate search can also be of use, as done in some concordancer software, for example by allowing the user to enter word-based regular expressions mixing literal words, word lemmas, word part-of-speech or even word classes (e.g. types of named entities). Boolean queries on indexes of word lemmas can also be used to offer yet more flexibility to search the lexicon, but at the cost of more candidate results. Once results are returned, they can recursively be reused as source phrases, so as to offer a means to navigate by iterative refining.

3 Evaluation of rephrasings in context for ranking results

Each candidate phrase rephrasing for a given phrase must be evaluated in order to define a ranking order for presentation to the user, and possibly

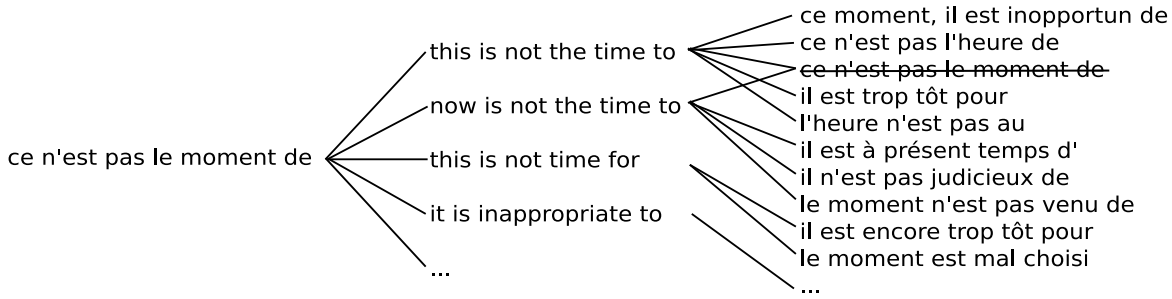


Figure 1: Example of rephrasing for the French phrase *ce n'est pas le moment de* using English as pivot.

to discard some of them. The proposed ranking should reflect as best as possible the preferences of the user for the task at hand in order to minimize reading time and maintain the user's interest in using the phrase lexicon. It is essential to give the user some control over how the results are returned depending on what is more important to her. For example, (Ferret and Zock, 2006) have proposed to present results from a dictionary enriched with topical associations in chunks to allow for categorial search. There will be cases where the user may find acceptable only grammatical results, while in other cases the user might accept agrammatical results provided they contain interesting suggestions. Moreover, it seems extremely important that result ranking can take into account the phrase substitution into the original context.

Considering how the proposed phrase lexicon is built, the pivot paraphrasing probability of equation 1 (PIV) can be used as a baseline ordering. Such a model reflects some strength of association between a rephrased phrase and the original phrase using the extracted phrases and conditional probabilities derived from a bilingual training corpus. It is therefore expected that results will be biased towards that corpus if the latter belongs to a particular genre or theme. Nonetheless, one can expect that some associations will be general enough to be of general interest.

In addition, several models that users can interpret as ranking criterion can be used simulateneously using the log-linear framework traditionally used in SMT systems. However, contrary to what is done in SMT, the weight of the models cannot be automatically optimized if we do not use an automatic evaluation of rephrasing quality, the definition of which depending heavily on the subjective appreciation of a user. Equation 2 shows how the score of a rephrasing p_2 of p_1 can be computed, where M is the set of models used, h_m is the logarithm of the normalized score of a model and λ_m its weight (with $\sum_{m \in M} \lambda_m = 1$), and C is the original sentence and the placeholder for the rephrased phrase.

$$s(p_2, p_1, C) = \sum_{m \in M} \lambda_m h_m(p_1, p_2, C) \quad (2)$$

3.1 Control over fluency

As noted by (Mutton et al., 2007), the notion of sentence-level fluency is not uniformely agreed upon, and its evaluation by human judges is sometimes found subjective, but in practice judges can obtain high levels of agreement about what can be considered fluent or not. Like (Callison-Burch, 2007), we can use a language model (LM) to assess the local fluency of a sentence after a phrase has been substituted with a rephrasing. A degradation in score (with a fluent original sentence) can indicate that the rephrasing segment should be adapted to the sentence, and/or that the sentence itself should be modified in order to integrate the new phrase as is.

Syntax parsers can produce various information that can be relevant for assessing the fluency of sentences, which can be used as features from different parsers for classification that can correlate well with human judgment (Mutton et al., 2007). When substituting a part of a sentence with another phrase and if this substitution does not require other changes in the sentence, then at least the dependency relationships between words outside that phrase should be preserved. This seems coherent with our objective of focussing on the task of phrase rephrasing when it is possible to modify only a given phrase and obtain an acceptable result.

3.2 Control over meaning preservation

The preservation of dependency relationships outside of the rephrased phrase can also play a role in terms of meaning preservation. Dependency relationships connecting words in the phrase and words outside the phrase (i.e., whose governor is outside the phrase and dependant inside it, or the opposite) should still exist after such a substitution, but possibly with a modified dependency target in the phrase. Indeed, those relationships denote the grammatical role of the words of the phrase relative to their context, and if those are preserved then it is more likely that meaning is preserved.

We use a model based on dependency preservation (DEP) which involves relationships outside the rephrased phrase and relationships crossing a boundary of that phrase. The score is based on some proportion of the number of such dependencies found after substitution over the number of original dependencies (see (Max, 2008) for details). Another way of controlling for meaning preservation is to ensure that only the pivot phrases with the same meaning as the original phrase are kept (and then their back translations). (Callison-Burch, 2007) has shown the positive impact on paraphrase quality of using a controlled pivot present in an aligned sentence in a test bilingual corpora. Phrase disambiguation techniques have been proposed for SMT and could be applied to the problem at hand (e.g. (Stroppa et al., 2007)). In an interactive context, it makes sense to let the user the opportunity to control for phrase sense by rejecting bad pivot phrases if she wants to, which is then similar to Callison-Burch's experiment settings. This manual selection must of course be optional, but can be used when a user prefers a stricter control on meaning. Another possibly interesting use is to disambiguate in a pivot language corresponding to one's native language when writing in a foreign language.

3.3 Control over lexical divergence

There will be cases when possible rephrasings will be very close to their original phrase, differing for example by only punctuation marks or verbal forms[1]. Writers may sometimes prefer rephrasings that differ by just one word, or on the contrary rephrasings that use a set of completely different words. To account for differents words between an original phrase and its rephrasing, we use a model (LEM) that returns a proportion of lemmas for full words that only belong to a rephrasing over all such lemmas for an initial phrase and its rephrasing (see (Max, 2008)).

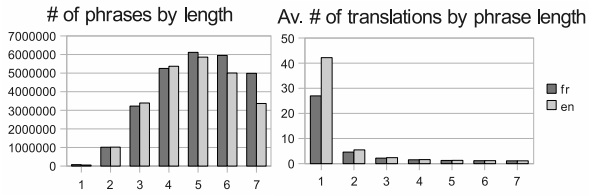


Figure 2: Bilingual phrase lexicon statistics

4 Experiments and evaluation

We carried out an evaluation on the local rephrasing of French sentences, using English as the pivot language.[2] We extracted phrase alignments of up to 7 word forms using the Giza++ alignment tool (Och and Ney, 2003) and the `grow-diag-final-and` heuristics described in (Koehn et al., 2003) on 948,507 sentences of the French-English part of the Europarl corpus (Koehn, 2005) and obtained some 42 million phrase pairs for which probabilities were estimated using maximum likelihood estimation. Statistics for the extracted lexicons are reported on figure 2. Entries of the monolingual phrase lexicon are built dynamically from the entries of the monolingual lexicons.

For the LM model, we used a 5-gram language model trained on the French part of the corpus using Kneser-Ney smoothing. The robust parser for French SYNTEX (Bourigault et al., 2005) was used to obtain lemmas for word and labeled dependency relationships between words, used respectively for the LEM and DEP models. Robust parsers provide the advantage that they can provide partial analysis for correct chunks in agrammatical sentences, but they can also recover information from agrammatical chunks which can be undesirable in this case.[3]

A test corpus of 82 sentences that were not used for extracting phrase alignments and learning the

[1] This is particularly the case when aligning between low and highly inflected languages.

[2] The main motivation for this choice was that we could easily have access to French native speakers for manual evaluation. We plan however to start new experiments using English, as well as experiments using another highly inflected language as pivot such as Spanish.

[3] We intend to use several parsers for English implementing different approaches as in (Mutton et al., 2007), but we had access to only one parser for French.

language model was built. A human judge selected one phrase of length 3 words or more per sentence that would be a good candidate for rephrasing, and which was accepted if it belonged to the French-English lexicon[4]. We kept at most the 20 first rephrasings obtained using the baseline PIV model, and asked two French native speakers to evaluate on a 5-level scale each the 1648 reformulated sentences obtained on *fluency*, *meaning preservation*, and *authoring value*, where the latter was described in the following way: (5) the rephrasing can be directly reused for revising a text, (4) the rephrasing can be used with a minor change, (3) the rephrasing contains elements that could be used for a good rephrasing, (2) the rephrasing contains elements that could suggest a rephrasing, and (1) the rephrasing is useless.

After the judges had completed manual annotation, smoothing of the scores was done by keeping mean scores for each sentence. We measured a value of 0.59 standard deviation for score differences between judges for grammaticality, 0.7 for meaning preservation and 0.8 for authoring value. Those values can indicate a growing difficulty in judging those characteristics, and in particular that judging authoring value on the proposed scale is more dependant on personal judgment. Results of mean scores for the first rank solutions with various model combinations with uniform weights are reported on figure 3, and results for mean authoring value scores depending on the number of top results presented to the user are reported on figure 4.

Authoring value scores are lower, which can be explained by the fact that rephrasings with bad fluency and/or meaning preservation scores will penalize authoring value scores according to our scale. The best results are obtained when combining all models, which remains true when considering mean results up to at least 8 rephrasings.

The baseline PIV model seems to have the most impact, but all other models also contribute in different ways. This suggests that which model should be used (or its weight in our framework) could be chosen by a user. In the following example, the LEM model helped select a rephrasing which obtained good scores:

Original sentence: *ce que je vous propose donc, c'est de travailler dans cette direction ...* (*what I therefore propose is to work towards this* . . .)

Rephrased sentence: *ce que je vous propose donc, c'est de coopérer dans ce sens ...* (*work towards this goal* . . .)

[4]This is a limitation of our evaluation, as our annotator was not strictly speaking revising a text that she wrote. We hope to be able to conduct task-based experiments in the future.

	fluency	meaning	authoring
PIV (baseline)	4.46	4.18	3.62
LM	4.28	3.62	3.45
DEP	4.35	3.68	3.43
LEM	4.05	3.21	3.28
PIV+LM	**4.65**	4.06	3.82
PIV+DEP	4.58	**4.27**	3.66
PIV+LEM	4.37	4.00	3.76
LM+DEP	4.49	3.81	3.68
LM+LEM	4.28	3.59	3.56
PIV+LM+DEP	**4.65**	4.05	3.92
PIV+LM+LEM	4.61	4.02	3.97
PIV+DEP+LEM	4.57	4.17	**4.02**
LM+DEP+LEM	4.37	3.69	3.64
PIV+LM+DEP+LEM	**4.68**	4.09	**4.05**

Figure 3: Mean results at first rank for various model combinations (uniform weighting)

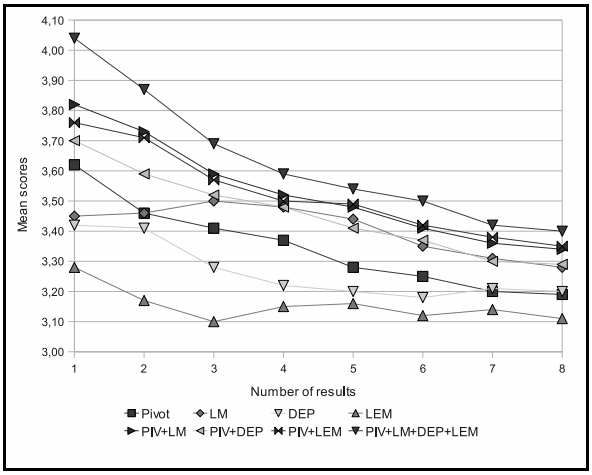


Figure 4: Mean authoring value scores depending on the number of results presented to the user

Figures 5 and 6 show two examples of rephrasings in French, whereby for each rephrasing the ranks given by PIV, LM and the combination of all mentioned models are shown.

5 Related work

While the traditional view of lexicons is word-based, we may as well consider larger units, including sentences. Corpus Pattern Analysis (CPA) (Hanks and Pustejovsky, 2005) is concerned with the prototypical syntagmatic patterns with which words in use are associated. For example, the meaning of *take place* is different from the mean-

Rephrasings	Ranks given by model(s)		
	PIV	LM	PIV+LM+DEP+LEM
quelques points essentiels	1	3	1
les points essentiels	19	1	2
plusieurs questions importantes	17	4	3
des points essentiels	8	6	4
deux ou trois questions importantes	5	9	5
plusieurs points importants	11	2	5
un certain nombre de questions importantes	17	7	7
certains points importants	2	5	8
un certain nombre de points importants	3	8	9
certains éléments très importants	13	11	10
une série de points importants	4	12	11
quelques accents importants	5	15	11
des choses extrêmement importantes	13	14	11
quelques remarques importantes ,	8	16	14
des points importants	12	10	15
quelques choses très importantes	13	17	16
certains points importants ,	8	13	17
quelques points essentiels sur	20	18	17
de certains éléments très importants	13	19	19
placer quelques accents importants	5	20	20

Figure 5: Examples of rephrasings for the phrase *quelques points importants* in *je voudrais mentionner* ***quelques points importants*** *de la directive*

Rephrasings	Ranks given by model(s)		
	PIV	LM	PIV+LM+DEP+LEM
vous avez raison	1	1	1
je suis d' accord avec vous	2	2	2
je suis d' accord	3	6	3
je conviens avec vous	6	5	4
je partage votre avis	7	4	5
vous avez raison de dire	10	3	5
je pense comme vous	7	8	7
je suis parfaitement d' accord avec vous	12	7	8
je partage votre point de vue	12	9	9
je vous rejoins	7	10	10
, je vous donne raison	3	12	11
là , je vous donne raison	3	13	12
tu as raison	16	11	12
vous avez raison de	10	14	14
je partage votre point	12	15	15
je partage votre point de	12	16	16

Figure 6: Examples of rephrasings for the phrase *je vous donne raison* in *à cet égard bien précis ,* ***je vous donne raison*** *, monsieur le commissaire*

ing of *take his place*, due to the possessive determiner. The actual meaning of words depends on the context in which they are used. The work done by the team of Gross on lexicon-grammar (e.g. (Gross, 1984)) showed that a relatively small set of clause patterns and syntactic constraints suffices to cover most of common French.

Comparable monolingual corpora have been used for automatic paraphrasing. Barzilay and Lee (Barzilay and Lee, 2003) learned paraphrasing patterns as pairs of word lattices, which are then used to produce sentence level paraphrases. Their corpus contained news agency articles on the same events, which allows precise sentence paraphrasing, but on a small sets of phenomena and for a limited domain. As sentential paraphrasing is more likely to alter meaning, Quirk *et al.* (Quirk et al., 2004) approached paraphrasing as a monotonous decoding by a phrase-based SMT system. Their corpus consisted of monolingual sentences extracted from a comparable corpus that were automatically aligned so as to allow aligned phrase extraction. Pang *et al.* (Pang et al., 2003) used parallel monolingual corpora built from news stories that had been independantly translated several times to learn lattices from a syntax-based alignment process.

Bannard and Callison-Burch (Bannard and Callison-Burch, 2005) proposed to use pivot translation for paraphrasing phrases. Fujita (Fujita, 2005) proposed a transfer-and-revision framework using linguistic knowledge for generating paraphrases in Japanese and a model for error detection. At the lexical level, a recent evaluation on English lexical substitution was held (McCarthy and Navigli, 2007) in which systems had to find lexical synonyms and disambiguate the context.

6 Discussion and future work

In this article, we have presented an approach for obtaining rephrasings for short text spans from parallel bilingual corpora. These rephrasings can be ranked according to user-defined preferences, and the weights of the models used can be dynamically adjusted by a user depending on what features are more important to her, for instance after an initial list of candidates has been proposed by the system. Indeed, good candidates include paraphrases, but also more generally phrases that could help a writer revise a text with some shifts in meaning, even if at the cost of some corrections to make the resulting text grammatical. Furthermore, search for rephrasings can be iteratively performed using candidate rephrasings as source phrases, and the user can have some fine-grained control if selecting or rejecting possible pivot phrases manually. Possible user interfaces to this proposed bilingual phrase lexicon could include rephrasing memory features to learn from interaction with the user, and concordancing features to display the context of use in the bilingual corpus of the segments used to build the relevant lexicon entries. In the latter case, the similarity used to select examples could take the context of the phrases into account in terms of dependency relationships.

There are several open issues to the presented work. Important issues are where the phrases can come from and the bias introduced by the resource used. Using a bilingual corpora such as the Europarl corpus with this pivot approach yields both generic and domain/genre-specific rephrasings, and it is important to be able to determine their appropriate context of use. It would also be interesting to investigate enriching this framework with phrases learnt from monolingual corpora from a given domain or genre, and to use features from the current text under revision. More generally, we would need to get some idea of the degree of possible reuse of a given rephrasing.

Another important group of issues concerns limitations due to the nature of phrases for the task at hand. As we have said, phrases as units of rephrasing are limited because they cannot model non-consecutive words and because of the rigidity of their content. Various types of entry points to the rephrasing lexicon such as using word-based regular expressions can in some way alleviate this problem, but work could be done on the lexicon itself. As shown by Callison-Burch (Callison-Burch, 2007), much can be gained by using better alignments. Alignments techniques using syntactic information could eliminate weak rephrasing candidates (i.e. increase in overall precision), but interesting phrasal alignments could be lost as well (decrease in overall recall). Furthermore, information from the context of alignments could also be used to disambiguate the source phrase and get only pivot phrases that are compatible with the context of a given rephrasing, in similar ways as recently done for SMT (Stroppa et al., 2007).

References

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*, Ann Arbor, USA.

Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL/HLT*, Edmonton, Canada.

Bourdaillet, Julien, Jean-Gabriel Ganascia, and Irène Fenoglio. 2007. Machine assisted study of writers' rewriting processes. In *Proceedings of NLPCS, poster session*, Madeire, Portugal.

Bourigault, Didier, Cécile Fabre, Cécile Frrot, Marie-Paule Jacques, and Sylvia Ozdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Proceedings of TALN*, Dourdan, France.

Callison-Burch, Chris. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh.

Fellbaum, Christiane, editor, 1998. *WordNet: An Electronic Lexical Database and some of its Applications*. MIT Press.

Ferret, Olivier and Michael Zock. 2006. Enhancing electronic dictionaries with an index based on associations. In *Proceedings of COLING/ACL*, Sydney, Australia.

Fujita, Atsushi. 2005. *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Ph.D. thesis, Nara Institute of Science and Technology.

Gross, Maurice. 1984. Lexicon-grammar and the analysis of french. In *Proc. of the 11th COLING*, pages 275–282, Stanford, CA.

Hanks, Patrick and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Koehn, Philipp, Franz Josef Och, , and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL/HLT*, Edmonton, Canada.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, Phuket, Thailand.

Max, Aurélien. 2008. Local rephrasing suggestions for supporting the work of writers. In *Proceedings of GoTAL*, Gothenburg, Sweden.

McCarthy, Diana and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Semeval-2007 Workshop at ACL*, Prague, Czech Republic.

Munteanu, Dragos S. and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of COLING/ACL 2006*, Sydney, Australia.

Mutton, Andrew, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU : Automatic evaluation of sentence-level fluency. In *Proceedings of ACL*, Prague, Czech Republic.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pang, Bo, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of NAACL/HLT*, Edmonton, Canada.

Quirk, Chris, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, Barcelona, Spain.

Stroppa, Nicolas, Antal van den Bosch, and Andy Way. 2007. Exploiting source similarity for smt using context-informed features. In *Proceedings of TMI*, Skvde, Sweden.

Zock, Michael and Slaven Bilac. 2004. Word lookup on the basis of associations : from an idea to a roadmap. In *Workshop on 'Enhancing and using electronic dictionaries'*, pages 29–35, Geneva. COLING.

Zock, Michael. 2006. Navigational aids, a critical factor for the success of electronic dictionaries. In Rapp, Reinhard, P. Sedlmeier, and G. Zunker-Rapp, editors, *Perspectives on Cognition: A Festschrift for Manfred Wettler*, pages 397–414. Pabst Science Publishers, Lengerich.

Zock, Michael. 2007. If you care to find what you are looking for, make an index: the case of lexical access. *ECTI, Transaction on Computer and Information Technology*, 2(2):71–80.

Toward a cognitive organization for electronic dictionaries, the case for semantic proxemy

Bruno Gaume, Karine Duvignau, Laurent Prévot, Yann Desalle
Université de Toulouse, CNRS
{gaume,duvignau,prevot,desalle}@univ-tlse2.fr

Abstract

We compare a psycholinguistic approach of mental lexicon organization with a computational approach of implicit lexical organization as found in dictionaries. In this work, we associate dictionaries with 'small world' graphs. This multidisciplinary approach aims at showing that implicit structure of dictionaries, mathematically identified, fits the way young children categorize. These dictionary graphs might therefore be considered as 'cognitive artifacts'. This shows the importance of semantic proximity both in cognitive and computational organization of verbs lexicon.

1 Introduction

According to (Dik, 1991) a linguistic theory should be compatible with psycholinguistic research on language acquisition, treatment, production, interpretation and memorization of linguistic expressions. We agree with this view and postulate that elaborating electronic dictionaries on the ground of a linguistic theory, satisfying Dik's principle, will confer them good ergonomics that will increase their usability. Our approach is to some extent comparable to WordNet initiative (Fellbaum, 1998), in the sense that we are trying to characterize speakers' mental lexicon.

In this paper, we focus on verb lexical organization through the examination of verbal pivot metaphorical utterances (VPMU). Such utterances involve an understudied structural aspect of the lexicon: *interdomain co-hyponymy* (Duvignau, 2002; Duvignau and Gaume, 2008). In this context, we take semantic proximity as a central principle for cognitive ergonomics influencing dynamic lexical acquisition and adult lexical organization. VPMU generally consists in substituting elements from different semantic domains. They are usually considered as deviants while they might constitute a linguistic illustration of the categorial flexibility advocated in (Piaget, 1945; Ny, 1979; Hofstadter, 1995). They might therefore reveal an early lexical structuring mode that may form a ground for improving electronic dictionaries.

This paper presents a mathematical method able to discover the areas in which this structuring mode appears in dictionaries. Our approach is to take advantage of the mathematical structure of the network generated by verb definitions. This structure has been mentioned in (Watts and Strogatz, 1998), studied for WordNet by (Sigman and Cecchi, 2002), refined in (Gaume et al., 2002) and exploited in the current proposal.

The paper is organized as follows. The next section brings evidence of categorization by semantic proximity from early lexicon acquisition experiments. Section 3 presents the computational model, hereafter *'proxemy'*. Section 4 details our work on lexical graphs while section 5 compares the results of experimental studies with those of the computational model.

2 Toward a categorization by semantic proximity: evidences from early lexicon acquisition

In order to show the importance of semantic approximation, we have chosen to support our claim with productions observed at the crucial period of lexical construction (between 2 and 4 years-of-

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 86–93
Manchester, August 2008

age) and to compare these with adult speakers that have a stabilized lexicon.

2.1 Inter-domains vs. intra-domain semantic approximations

Studies in this field are almost exclusively limited to nominal utterances. (Duvignau et al., 2005) established the existence of the production of metaphor-like utterances with a verbal pivot in 2-4 years-old children and proposed to consider them, at this stage of language development, as semantic approximations and not as mistakes or true metaphors. Duvignau distinguished two kinds of semantic approximations: Inter-domains proximity and intra domain proximity between verbs (Duvignau, 2002).

- Inter-domains proximity / co-hyponymy between verbs : a 'linguistic approximation'

(1) Elle déshabille l'orange (She undresses the orange) [Age: 3 years] [movie: a lady peels an orange]

In this category of approximation, the verb used by the speaker constitutes a reference to a semantic domain different from the one of element it is combined to *('undress' / 'orange')*. For this reason, the approximate character of the verb is understandable independently of the context of the utterance: detecting the approximation occurs at the linguistic level. We call this type of production 'semantic approximation'. They might constitute a metaphor or an 'analogic surextention'.

When someone has a conventional verb in the mental lexicon *('to peel')* and use a non conventional but relevant verb like *'to undress the orange'* for the action *[to peel the orange* his verbal semantic approximation constitutes a metaphor. On the contrary when someone does not have a conventional verb in the mental lexicon but manages to use a non conventional but relevant verb in saying *'to undress'* for this action, his verbal semantic approximation constitutes a 'surextension' but not an error because of the lexical relation that links these verbs. In fact, according to (Duvignau and Gaume, 2008) *'to undress'* and *'to peel'* are related by an inter-domains synonymic relation.

- Intra-domain proximity / co-hyponymy between verbs: a 'pragmatic approximation' In this category, illustrated by (2) the approximate character of the verb comes only from a non-correspondence between the verb used and the reality it designates. This happens with utterances in which the use of the verbal form does not create any semantic tension within the utterance but designates a way of carrying out an activity that does not correspond precisely to the action undertaken.

(2) Elle coupe l'orange (She cuts the orange)[age: 3 years][movie: a lady peels an orange]

We propose an experimental study of the production of verbal semantic approximations like (2) or (1) by way of a naming task of 17 action-movies with young children (from 2 to 4 years old). We compare their performances with adult's ones.

2.2 Experimental Design

In order to elicit the production of semantic approximations we proposed to all our participants an action-video naming task. The population sample consisted of:

- 54 non-disturbed children (2-4 years old), monolingual in French
- 77 non-disturbed young adults (18-40 years old), monolingual in French

The action movies sequences are coming from the *Approx* protocol (Duvignau et al., 2005). The material consists in 17 action-movies sequences described in table 1.

The 17 action movies are presented in random order to each participant. Instructions were given at the time the action in the movie was completed and its results were visible (e.g when the glass is broken). At that moment a question was asked to the participant:*'What did the woman do? (just now)'*

2.3 Results

Each of the children produced between 2 and 5 approximations: *'Elle casse la tomate' -She breaks a tomato '* [action = to squash], *'Elle épluche le bois' - She peels the wood'* [action = to strip the bark off a log]. Globally, children produced semantic approximations for 34 % of the naming tasks, which were distributed as follows: 24 % intra-domain semantic approximations, 10 % inter-domains semantic approximations. They produced them significantly more frequent than adults : 5 % with 4% intra-domain semantic approximations and 1 % inter-domains semantic approximations.

/DAMAGE/	/TAKE OF /	/SEPARATE/
1- burst a balloon 2- screw up a piece of paper 3- break a glass with a hammer 4- squash a tomato with the hand 5- tear up a newspaper	6- peel a carrot with a peeler. 7- peel an orange with one's hands. 8- strip the bark off a log 9- undress a doll 10- take apart a lego structure. 11- peel a banana	12- make bread-crumbs by hand. 13- slice bread with a knife. 14- break up bread with one's hands 15- shred parsley with a knife. 16- saw a wooden plank 17- tear up a shirt

Table 1: *Approx* 17 action movies

The *student Test* shows the difference between children and adults in terms of production of semantic approximation is very significant: here $p < 0,01$ while $p < 0,05$ is enough.

These results signal the importance of semantic approximations and of semantic proximity between verbs in the cognitive organization of verbs lexicon.

In the rest of the paper we present a computational model of semantic proximity and then compare this model with the experimental data obtained from the children.

3 *Proxemy*: a computational approach

A theory of language useful for computational work must account for language statistical regularities. Zipf law (Zipf, 1949) satisfy this observation but provides little insight on lexical structural organization. More recent graph theory studies (Ferrer-i-Cancho and Sole, 2001; Sigman and Cecchi, 2002), capitalizing on results in other scientific domains, provided interesting contributions to the establishment of such a theory of language. All structures discovered in this field research satisfy the "hierarchical small word" (HSW) definition (see section 3.1). Our approach takes place in this general framework. Our specificities are:

- a new linguistic and psycholinguistic insight that guides us and help us on our results validation;
- the kind of objects studied (dictionaries);
- our analysis of graph structure resulting in a computational model of *semantic proximity* among vertices (here vertices are French verbs).

The study by (Resnik and Diab, 2000) signaled that although existing models for verb similarity performed reasonably well against human judgments, none managed to handle certain types of metaphorical pairs such as *to undress / to peel off* that are nonetheless declared to be rather similar by speakers. We aim to develop a model addressing this issue.

3.1 Small World Networks

Networks corresponding to structures found in real world (henceforth *real world networks*) are sparse: in a graph with n nodes, the maximum number of possible edges is $O(n^2)$ while the number of edges in real networks is generally inferior to $O(nlog(n))$. Watts and Strogatz (Watts and Strogatz, 1998) proposed two indicators to characterize a large sparse graph G:

- **L** : *the characteristic path length*, i.e the mean of the shortest path between two nodes of G
- **C** : *the clustering coefficient*, $C \in [0, 1]$, it measures the graph tendency to host zones very dense in edges. (The more clustered the graph is, the more the graph's **C** approaches 1, whereas in random graphs **C** is very close to 0).

In applying these criteria to different types of graphs, Watts and Strogatz found that:

- real world networks have a tendency to have a small **L**: generally there is at least one short path between any two nodes ;
- real world networks have a tendency to have a large **C**: this reflects a relative tendency for two neighbors on the same node to be interconnected;
- random graphs have a small **L**: If someone builds a graph randomly with a density of edges comparable to real world networks, it will obtain graphs with a small *L*;
- random graphs have a small **C**: They are not composed of aggregates. In a random graph

there is no reason why neighbors of a same node are more likely to be connected than any two other nodes, hence their poor tendency to form aggregates.

Watts and Strogatz proposed to call the graphs having these two characteristics (a small **L** and a large **C**) *small worlds* (SW). They recognized these SW in all the real world networks they observed, and therefore postulated for being a SW was an universal property of real world networks. A complete presentation of *Small Words* can be found, for example, in (Newman, 2003).

More recent research has shown that most SW also have a hierarchical structure (*hereafter hierarchical small worlds, HSW*). The distribution of the vertices incidence degrees follows a power law. The probability $P(k)$ that a given node has k neighbors decreases as a power law, $P(k) \approx k^{-\lambda}$, where λ is a constant characteristic of the graph (Barabási and Albert, 1999), while random graphs conforms to a Poisson Law.

In the next section, we present *'proxemy'*, a semantic proximity measure based on a distance we define. A interesting particularity of this distance is to calculate the distance between two vertices on the ground of the complete graph, and not only on their direct neighbors.

3.2 The mathematical model

PROX (PROXemy) is a stochastic method designed for studying "Hierarchical Small Worlds".[1] This method takes graph as input and transform them in a Markov chain whose states are graph vertices. Metaphorically, energy particles wander randomly from vertex to vertex through the edges of the graph. It is their trajectory dynamics that give us the structural properties of the graph.

PROX takes a graph in input and output a similarity measure between the vertices of the graph. Our problem is therefore the opposite than the one of Pathfinder networks (PFNETs see (Schvaneveldt et al., 1988)). PFNETs take a full proximity matrix in input and output a sparse graph. Their goal is to minimize the number of edges required in the sparse graph to be able to approximate the full distance matrix corresponding to the initial full proximity matrix.

PROX build a similarity measure between the vertices. The hypothesis is that areas having a high density in edges (hereafter, these areas will be called *aggregates*) correspond to closely related verb meanings (in a graph of verbs).

Given a graph with n vertices, $G = (V, E)$, we will note $[G]$ the matrix $n \times n$ such that $\forall r, s \in V$, $[G]_{r,s} = 0$ if $\{r, s\} \notin E$ and 1 otherwise. $[G]$ is called the adjacency matrix of G.

Given $G = (V, E)$ a reflexive graph with n vertices. $[\hat{G}]$ is a $n \times n$ matrix defined by $\forall r, s \in V$, $[\hat{G}]_{r,s} = \frac{[G]_{r,s}}{\sum_{x \in V}\{[G]_{r,x}\}}$. $[\hat{G}]$ is the Markovian matrix of G.

$[\hat{G}]$ is the $n \times n$ matrix is a transition matrix of the homogeneous Markov chain whose states are the vertices of the graph such that the probability of going from one vertex $r \in V$ at an instant t onto another $s \in V$ at the instant $t + 1$ is equal to:

- 0 if $\{r, s\} \notin E$ (s is not neighbor of r)
- $1/D$ if $\{r, s\} \in E$ and r has D neighbors (s is a neighbor of r)[2]

Given $G = (V, E)$ a reflexive graph with n vertices and $[\hat{G}]$ its Markovian matrix, $\forall r, s \in V, \forall t \in \mathbb{N}^*$, $PROX(G, t, r, s) = [\hat{G}^t]_{r,s}$

$PROX(G, t, r, s)$ is therefore the probability for a particle departing from r at the instant zero to be on s at the instant t.

Therefore when, $PROX(G, t, r, s) > PROX(G, t, r, u)$, the particle has more probability to be, at instant t on s than on u and it is graph structure that determine these probabilities.

For the rest of this paper we will set the value of t to 4 since **L** is less than 4 in the kind of graph we are concerned with. Therefore, we take into account the global graph simply by calculating $PROX(G; 4; r; s)$.

Now we have defined our model we will present lexical graphs on which we apply it.

4 Lexical graphs

Several types of lexical graphs can be built according to the type of the semantic relation used for defining the graph's edges. The two principal types of relations used are:

[1] In this paper we will use the term *'proxemy'* to refer to the obtained by PROX algorithm. It corresponds to some kind of semantic proximity.

[2] In the context of this presentation of the model we do not consider weighted graphs. However when building the graphs we do consider information, such as the position of the word in the definition, for giving weight to the edges.s

- Syntagmatic relationships, like co-occurrence relationships: they define edges between nodes corresponding to words found near to each other in a corpus.
- Paradigmatic relationships, like synonymy: they define, on the ground of lexical databases such as WordNet (Fellbaum, 1998), edges between nodes of words being in a synonymy relationship in such resource.

Moreover, we are interested into less specific relations, called *semantic proximity relations* or *semantic relatedness*, and which covers both paradigmatic and syntagmatic dimensions.

4.1 Dictionary graphs

Meaning in dictionary definition is at least partially brought by the relations they create between the words constituting the entries. Our approach consists in exploiting the small word properties of the graphs corresponding to dictionaries. More precisely, we are taking advantage of our hypothesis that *aggregates* correspond to areas of closely related senses. We illustrate our approach on two kinds of dictionary, two traditional dictionaries, *Le Grand Robert*[3] and *TLFi*[4], and an synonym dictionary *(Dicosyn)* made of compilation of synonym relations extracted from seven other dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert).[5]

We create a graph from a dictionary in the following way. The entries constituted the vertices. Edges between two vertices A and B were added if and only if B appears in A's lemmatized definition[6] as illustrated in Figure 4.1

We proceed in this way for each entry and obtained a graph of the dictionary. By extracting the subgraph composed only of verbs, the 'neighborhood' we get for the verb 'écorcer' is illustrated by Figure 4.1. Then we render the graph symmetric and reflexive. These modifications on the graphs are allowed thanks to its paradigmatic nature. Graphs created in this way are typical *small world network*. For example, DicoSyn-Verb has 9043 vertices and 50948 edges, its **L** is 4,1694 and its *C* 0,3186.

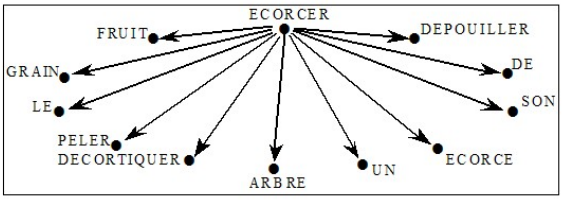


Figure 1: Sub-graph near 'écorcer (to bark – a tree–)' from *Le Grand Robert*

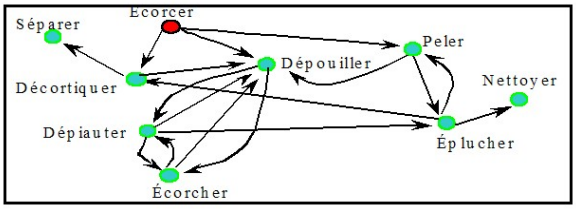


Figure 2: Sub-graph of the verbs near 'écorcer' from Robert

(Duvignau, 2002) has shown that co-hyponymy verb lexical organization according fits with a power law distribution of incidence degrees. In our opinion, (i) the hierarchical organization of dictionaries is a consequence of the special role of the hypernymy relation together with the polysemy of some specific vertices; (ii) the strong **C** reflects the role of interdomain co-hyponyny (Duvignau, 2002; Duvignau and Gaume, 2003). For example, in French language, *'casser (to break)'* appears in many definitions: *'émietter (to crumble)', 'fragmenter (to fragment)', 'détériorer (to damage)', 'révoquer (to dismiss)', 'abroger (to abrogate)'*. This results in a very high incidence for the vertex *'casser (to break)'*. Moreover, many triangles exist *({casser, émietter, fragmenter}, {casser, révoquer, abroger}...,)* and they help to create aggregates. These areas that are bringing co-hyponyms closer in the resulting graph.

4.2 Disambiguization for creation dictionary graphs

Word Sense Disambiguation is a general issue for natural language processing that we need to address when we build our graphs. We need to disambiguate the verbs we found in the definition facing a similar problem as (Harabagiu et

[3] A significant amount of work has been done to encode *'Le Grand Robert* in a graph.

[4] We would like to thank ATILF for making the TLFi resource available to us.

[5] *Dicosyn* has been first realized at ATILF (Analyse et Traitement Informatique de la Langue Française), before being corrected at CRISCO laboratory *(http://elsap1.unicaen.fr/dicosyn.html)*.

[6] Lemmatization has been realized with TreeTagger *(http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)*.

al., 1999). For example, in French dictionary *Le Grand Robert*, there are two distinct entries for the verb 'causer': *to cause* (3) and *to chat* (4).

(3) CAUSER-1: être la cause de. *(to be the cause of)*

(4) CAUSER-2: S'entretenir familièrement avec qqn. *to chat with*

Of course, the word 'causer' may appear in other definitions like 'bavarder' *(to chat)* . Although a French speaker knows that the 'causer' in (5) refers to the definition (4) our system for building the graph cannot disambiguate. The solution we propose is to (i) first create a fictive vertex which is not a dictionary entry and then (ii) adds two edges {CAUSER, CAUSER-1} and {CAUSER, CAUSER-2 }. When 'causer' is found in another definition like (5), we add the edge { BAVARDER, CAUSER } as illustrated in Figure (5).

(5) BAVARDER "Parler beaucoup, longtemps ou parler ensemble de choses superficielles. - Parler; babiller, bavasser (fam.), cailleter, caqueter, **causer**, discourir, discuter, jaboter, jacasser, jaser, jaspiner (argot), lantiponner (vx), papoter, potiner. Bavarder avec qqn ... "

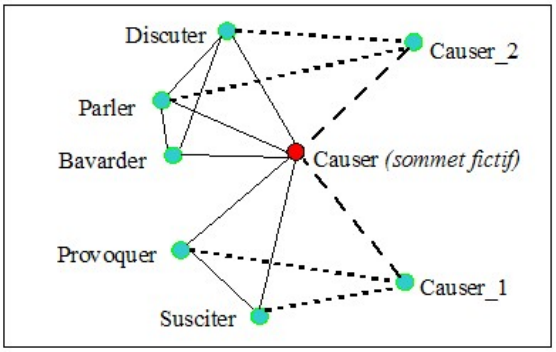


Figure 3: Disambiguation: 'Causer', fictive vertice

In Figure (5), many edges are hidden for clarity reasons. Dashed edges ({Discuter, Causer2}) result from the fact 'Discuter' and 'Parler' are in the definition of 'Causer-2'.

At this stage, we apply PROX to such graph as the one Figure (5) in order to get a matrix $[\hat{G}^4]$ as defined in section 3.2. $[\hat{G}^4]_{bavarder,causer-1} < [\hat{G}^4]_{bavarder,causer-2}$. This comparison allows us to disambiguate.

More generally, let suppose we found a word with k entries in a definition, we will then have $S_1, \ldots, S_k$ vertices corresponding to the entries a fictive vertex S. In case there is an edge $\{A, S\}$ it is replaced by $\{A, S_i\}$ where S_i is such that $[\hat{G}^4]_{A,S_i} = MAX_{0<i\leq k}\{\hat{G}^4]_{A,S_i}\}$. Then we remove all fictive vertices from the graph to get a disambiguated graph.

We can then apply PROX a last time on the disambiguated graph in order to get the closest word of a word according to our proxemy measure. For example, the PROX-closest words of *écorcer (to bark –a tree–)*, calculated with $t = 6$ are: *1 ECORCER (to bark), 2 DÉPOUILLER (strip), 3 PELER (peel), 4 TONDRE (mow, shear), 5 ÔTER (remove), 6 ÉPLUCHER (peel, pare), 7 RASER (shave), 8 DÉMUNIR (divest), 9 DÉCORTIQUER (decorticate), 10 ÉGORGER (slit the throat of), 11 ÉCORCHER (skin), 12 ÉCALER (husk), 13 VOLER (steal), 14 TAILLER (prune), 15 RÂPER (grate), 16 PLUMER (pluck), 17 GRATTER (scrape), 18 ENLEVER (remove), 19 DÉSOSSER (bone), 20 DÉPOSSÉDER (dispossess), 21 COUPER (cut), 22 BRETAUDER (shear sloppily), 23 INCISER (incise), 24 GEMMER (tap), 25 DÉMASCLER (remove first layer of cork)*[7]

5 Proxemy and Experimental studies

Prox is a robust method: changing randomly a few edges does not change significantly the results. The repartition of aggregates is not strongly affected by a random redistribution of some edges. However the relevance of our proxemy approach of lexical networks is tied to the linguistic representativity of the networks we use. Therefore, we tested the PROX model of four different dictionary graphs and we compared them to the psycholinguistic experimental results presented in section 2. The graph we compared were:

1. Graph.TLFI.Verb, a graph built as explained in 4.1 from TLFi[8] dictionary,
2. Graph.Robert.Verb, a graph built as explained in 4.1 from *Le Grand Robert* dictionary,
3. Graph.DicoSyn.Verb, in which there is a edge between two verbs if there are given as syn-

[7] Proposing a translation for such fine grained and sometimes polysemous words is impossible since proposing the translation include a certain form of disambiguisation as it is suggested by the work of (Gale et al., 1992).

[8] http://atilf.atilf.fr/tlf.htm

onyms by one of the synonym dictionary composing DicoSyn

4. Graph.DicoSyn_20 built from Graph.DicoSyn but in which 20% of the edges are randomly removed and re-added.

For each of these graphs we looked at two variables to be related with the psycho-linguistics experiments: the answers incidence and the proximity of answers to a *'reference verb'*

Answers incidence We compare in the graph the average incidence degree between adult (ID_{adult}) and children answers ($ID_{children}$).

Average incidence	Children	Adults
Graph.TLFi.Verb	61	29
Graph.Robert.Verb	236	126
Graph.DicoSyn.Verb	102	58
Graph.DicoSyn.Verb _20	66	40

Table 2: Results for 'Answers incidence'

The proximity of answers to a *'reference verb'* Three linguist judges determined together for each movie which was the most appropriate verb to describe the action performed in the movie (hereafter R_i is the reference verb for the movie M_i). For a given movie M_i, an answer may therefore be ranked according to its proxemy according to R_i.

For a lexical graph $G = (V, E)$ composed of n words, and for a reference verb $R_i \in V$, one can define $rank_{Ri}$ for ranking all the vertices of V in decreasing order resulting from a PROX iteration $PROX(G, t, Ri, \bullet)$ on V (see section 3.2).

Average rank relatively to reference verbs	Children	Adults
Graph.TLFi.Verb	270	173
Graph.Robert.Verb	121	76
Graph.DicoSyn.Verb	105	44
Graph.DicoSyn.Verb _20	185	94

Table 3: Proximity between answers and reference verb

Our first hypothesis was that $ID_{adult} < ID_{children}$. According to the hypothesis children would learn first words corresponding to high incidence vertices. Then they would use them for talking about an large lexical area (e.g 'casser' (to break) is used by children while adults use a more precise verb like 'déchirer' (to tear) which has a lower incidence in dictionary graphs).

Our second hypothesis was that the mean of the rank of the children answers according to the reference verb is higher that the adult ones. When a child is attempting to communicate an event *(e.g déchirer un livre, to tear a book)* for which he does not have an already constituted verbal category, he would do an analogy with a past event *(e.g to break a glass)* and use this verb for describing the current event *(e.g casser un livre, to break a book)*. The adult could use a number of more accurate verbs but their proxemic rank, with regard to the reference verb, is generally lower than the children ones.

The table 2 shows the results concerning answers incidence. Although some variability is observed across the graphs, our first hypothesis is validated for the 4 graphs. On the three first graphs the average incidence of answers is roughly twice as the adults one.

The table 3 illustrates the results concerning proxemic rank of answers according to the reference verb. Again, in spite of some variability across the graphs our second hypothesis is validated as well. Moreover, having in mind that the graph has about 10 000 vertices, we observe that although less close that adults answers, the children answers remain relatively close to the reference verb according to our proxemic measure.

6 Conclusion

Our psycholinguistic approach allows us to establish that semantic proximity between verbs play a fundamental role during the period of early lexical acquisition. We signaled the existence in the organization of the lexicon of a relation of co-hyponymy between verbs. Based on these first observations we consider that productions based on semantic proximity are particularly interesting: they manifest the existence, at the surface level of discourse, of a lexical relation of inter-domain 'semantic proximity' between verbs not yet considered in linguistics.

Moreover we have seen that semantic approximations for verbs appear to fit the proximity values calculated by PROX. On the ground of these first results, we postulate that constructing electronic dictionaries on the ground of linguistic theory of lexical semantic organization that fits with early lexicon acquisition as well with adult lexical organization will provide them interesting ergonomics properties. This should increase their usability and

might be taken into account for normalizing electronic dictionaries.

For example, we are developing a 'proxemic electronic dictionary' from TLFi. Such dictionaries enable to find an uncommon but precise verb like *'to bark'* by using (i) a common verb like *'to undress'* which is related to *'to bark'* by semantic proximity and (ii) a word *(e.g 'tree')* bringing a relevant semantic domain. Moreover, in the definition of *'to bark'* one can find: *'tree', 'grain' 'fruit'* which are close from each other according to PROX ran on nouns. Finally, when we look for verbs that are close from both *'to undress'* and *'tree'*, PROX provides the verbs: *'to cut, to ring, to peel, to notch, to bark, to incise,...'* which constitute relevant verbs. Such a dictionary can be useful for didactic studies where it can complements approaches like and NLP for word sense desambiguization (Gaume et al., 2004) or de-metaphorization.

References

Barabási, Albert-László and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509–512, October.

Dik, S. 1991. Functional grammar. In Droste, F. and J. Joseph., editors, *Linguistic theory and grammatical description*. Amsterdam : Benjamins.

Duvignau, K. and B. Gaume. 2003. Linguistic, psycholinguistic and computational approaches to the lexicon: Contributions to early verb-learning. *Journal of the European Society for the Study of Cognitive Systems*, 6(1).

Duvignau, Karine and Bruno Gaume. 2008. Between words and world: Verbal 'metaphor' as semantic or pragmatic approximation? In *Proceedings of International Conference 'Language, Communication and Cognition'*.

Duvignau, K., B. Gaume, and S. Kern. 2005. Semantic approximations intraconceptăvs. interconcepts in early verbal lexicon: flexibility against error. In *Proceedings of ELA 2005, Emergence of language abilities: ontogeny and phylogeny*.

Duvignau, K. 2002. *La métaphore berceau et enfant de la langue*. Ph.D. thesis, Université Toulouse Ű Le mirail.

Fellbaum, C., editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ferrer-i-Cancho, Ramon and Ricard V. Sole. 2001. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, November.

Gale, W., K. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the humanities*, 26(2):415–439.

Gaume, B., K. Duvignau K., O. Gasquet O., and M-D. Gineste. 2002. Forms of meaning, meaning of forms. *Journal of Experimental and Theoretical Artificial Intelligence*, 14:61–74.

Gaume, B., N. Hathout, and P. Muller. 2004. Désambiguisation par proximité structurelle. In *Proceedings of TALN 2004*.

Harabagiu, Sanda M., George A. Miller, and Dan I. Moldovan. 1999. Wordnet 2 - a morphologically and semantically enhanced resource. In *SIGLEX 1999*.

Hofstadter, D. 1995. *Fluid concepts and creative analogies*. New York : Basic Books.

Newman, M. 2003. The structure and function of complex networks.

Ny, J-F. Le. 1979. *La sémantique psychologique*. PUF.

Piaget, J. 1945. *La formation du symbole chez lŠenfant,*. Delachaux et Niestlé.

Resnik, P. and M. Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*.

Schvaneveldt, R. W., D. W D.W Dearholt, and F.T Durso. 1988. Graph theoric foundations of pathfinder networks. *Computers and Mathematics with Applications*, 15:337–445.

Sigman, M. and G.A. Cecchi. 2002. Global organization of the wordnet lexicon. *Proc. Natl. Acad. Sci.*, 99(3):1741–1747.

Watts, D.J. and S.H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature*, 393:440–442.

Zipf, G. K. 1949. *Human behavior and the principle of least effort*. Addison-Wesley.

Cognitively Salient Relations for Multilingual Lexicography

Gerhard Kremer
CIMeC
University of Trento
gerhard.kremer@unitn.it

Andrea Abel
EURAC
Bolzano
aabel@eurac.edu

Marco Baroni
CIMeC
University of Trento
marco.baroni@unitn.it

Abstract

Providing sets of semantically related words in the lexical entries of an electronic dictionary should help language learners quickly understand the meaning of the target words. Relational information might also improve memorisation, by allowing the generation of structured vocabulary study lists. However, an open issue is *which* semantic relations are cognitively most salient, and should therefore be used for dictionary construction. In this paper, we present a concept description elicitation experiment conducted with German and Italian speakers. The analysis of the experimental data suggests that there is a small set of concept-class–dependent relation types that are stable across languages and robust enough to allow discrimination across broad concept domains. Our further research will focus on harvesting instantiations of these classes from corpora.

1 Introduction

In electronic dictionaries, lexical entries can be enriched with hyperlinks to semantically related words. In particular, we focus here on those related words that can be seen as systematic *properties* of the target entry, i. e., the basic concepts that would be used to define the entry in relation to its superordinate category and coordinate concepts. So, for example, for animals the most salient relations would be notions such as "parts" and "typical behaviour". For a horse, salient properties will include the mane and hooves as parts, and neighing as behaviour.

Sets of relevant and salient properties allow the user to collocate a word within its so-called "word field" and to distinguish it more clearly from neighbour concepts, since the meaning of a word is not defined in isolation, but in contrast to related words in its word field (Geckeler, 2002). Moreover, knowing the typical relations of concepts in different domains might help pedagogical lexicography to produce structured networks where, from each word, the learner can naturally access entries for other words that represent properties which are salient and distinctive for the target concept class (parts of animals, functions of tools, etc.). We envisage a natural application of this in the automated creation of structured vocabulary study lists. Finally, this knowledge might be used as a basis to populate lexical networks by building models of concepts in terms of "relation sketches" based on salient typed properties (when an animal is added to our lexicon, we know that we will have to search a corpus to extract its parts, behaviour, etc., whereas for a tool the function would be the most important property to mine).

This paper provides a first step in the direction of dictionaries enriched with cognitively salient property descriptions by eliciting concept descriptions from subjects speaking different languages, and analysing the general patterns emerging from these data.

It is worth distinguishing our approach to enriching connections in a lexical resource from the one based on free association, such as has been recently pursued, e. g., within the WordNet project (Boyd-Graber et al., 2006). While we do not dispute the usefulness of free associates, they are irrelevant to

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 94–101
Manchester, August 2008

our purposes, since we want to generate systematic, structured descriptions of concepts, in terms of the relation types that are most salient for their semantic fields. Knowing that the word *Holland* is "evoked" by the word *tulip* might be useful for other reasons, but it does not allow us to harvest systematic properties of flowers in order to populate their relation sketch: we rather want to find out that tulips, being flowers, will have *colour* as a salient property type. As a *location* property of tulips, we would prefer something like *garden* instead of the name of a country or individual associations. To minimise free association, we asked participants in our experiments to produce concept *descriptions* in terms of characteristic properties of the target concepts (although we are not aware of systematic studies comparing free associates to concept description tasks, the latter methodology is fairly standard in cognitive science: see section 2.2 below).

To our knowledge, this sort of approach has not been proposed in lexicography, yet. Cognitive scientists focus on "concepts", glossing over the fact that what subjects will produce are (strings of) words, and as such they will be, at least to a certain extent, language-dependent. For lexicographic applications, this aspect cannot, of course, be ignored, in particular if the goal is to produce lexical entries for language learners (so that both their first and their second languages should be taken into account).

We face this issue directly in the elicitation experiment we present here, in which salient relations for a set of 50 concepts from 10 different categories are collected from comparable groups of German and Italian speakers. In particular, we collected data from high school students in South Tyrol, a region situated in Northern Italy, inhabited by both German and Italian speakers. Both German and Italian schools exist, where the respective non-native language is taught. It is important to stress that the two communities are relatively separated, and most speakers are *not* from bilingual families or bilingual social environments: They study the other language as an intensively taught L2 in school. Thus, we move in an ideal scenario to test possible language-driven differences in property descriptions, among speakers that have a very similar cultural background.

South Tyrol also provides the concrete applicative goal of our project. In public administration and service, employees need to master both languages up to a certain standardised level (they have to pass a "bilingual" proficiency exam). Therefore, there is a big need for language learning materials. The practical outcome of our research will be an extension of ELDIT[1], an electronic learner's dictionary for German and Italian (Abel and Weber, 2000).

2 Related Work

Lexicographic projects providing semantic relations and experimental research on property generation are the basis for our research.

2.1 Dictionaries

In most paper-based general and learners' dictionaries only some information about synonyms and sometimes antonyms is presented. Newer dictionaries, such as the "Longman Language Activator" (Summers, 1999), are providing lists of related words. While these will be useful to learners, information about the *kind* of semantic relation is usually missing.

Semantic relations are often available in electronic resources, most famously in WordNet (Fellbaum, 1998) and related projects like Kirrkirr (Jansz et al., 1999), ALEXIA (Chanier and Selva, 1998), or as described in Fontenelle (1997). However, these resources tend to include few relation types (hypernymy, meronymy, antonymy, etc.). The salience of the relations chosen is not verified experimentally, and the same set of relation types is used for all words that share the same part-of-speech. Our results below, as well as work by Vinson et al. (2008), indicate that different concept classes should, instead, be characterised by different relation types (e. g., function is very salient for tools, but not at all for animals).

2.2 Work in Cognitive Sciences

Several projects addressed the collection of property generation data to provide the community with feature norms to be used in different psycholinguistic experiments and other analyses: Garrard et al. (2001) instructed subjects to complete phrases ("concept is/has/can. . . "), thus restricting the set of producible feature types. McRae et al. (2005) instructed their subjects to list concept properties without such restrictions, but providing them with some examples. Vinson et al. (2008)

[1] URL `http://www.eurac.edu/eldit`

gave similar instructions, but explicitly asked subjects not to freely associate.

However, these norms have been collected for the English language. It remains to be explored if concept representations in general and semantic relations for our specific investigations have the same properties across languages.

3 Data Collection

After choosing the concept classes and appropriate concepts for the production experiment, concept descriptions were collected from participants. These were transcribed, normalised, and annotated with semantic relation types.

3.1 Stimuli

The stimuli for the experiment consisted of 50 concrete concepts from 10 different classes (i. e., 5 concepts for each of the classes): *mammal* (dog, horse, rabbit, bear, monkey), *bird* (seagull, sparrow, woodpecker, owl, goose), *fruit* (apple, orange, pear, pineapple, cherry), *vegetable* (corn, onion, spinach, peas, potato), *body part* (eye, finger, head, leg, hand), *clothing* (chemise, jacket, sweater, shoes, socks), *manipulable tool* (comb, broom, sword, paintbrush, tongs), *vehicle* (bus, ship, airplane, train, truck), *furniture* (table, bed, chair, closet, armchair), and *building* (garage, bridge, skyscraper, church, tower). They were mainly taken from Garrard et al. (2001) and McRae et al. (2005). The concepts were chosen so that they had unambiguous, reasonably monosemic lexical realizations in both target languages.

The words representing these concepts were translated into the two target languages, German and Italian. A statistical analysis (using Tukey's honestly significant difference test as implemented in the R toolkit[2]) of word length distributions (within and across categories) showed no significant differences in either language. There were instead significant differences in the frequency of target words, as collected from the German, Italian and English WaCky corpora[3]. In particular, words of the class *body part* had significantly larger frequencies across languages than the words of the other classes (not surprisingly, the words *eye*, *head* and *hand* appear much more often in corpora than the other words in the stimuli list).

[2] URL `http://www.r-project.org/`
[3] URL `http://wacky.sslmit.unibo.it/`

3.2 Experimental Procedure

The participants in the concept description experiment were students attending the last 3 years of a German or Italian high school and reported to be native speakers of the respective languages. 73 German and 69 Italian students participated in the experiment, with ages ranging between 15 and 19. The average age was 16.7 (standard deviation 0.92) for Germans and 16.8 (s.d. 0.70) for Italians.

The experiment was conducted group-wise in schools. Each participant was provided with a random set of 25 concepts, each presented on a separate sheet of paper. To have an equal number of participants describing each concept, for each randomly matched subject pair the whole set of concepts was randomised and divided into 2 subsets. Each subject saw the target stimuli in his/her subset in a different random order (due to technical problems, the split was not always different across subject pairs).

Short instructions were provided orally before the experiment, and repeated in written format on the front cover of the questionnaire booklet distributed to each subject. To make the concept description task more natural, we suggested that participants should imagine a group of alien visitors, to each of which a particular word for a concrete object was unknown and thus had to be described. Participants should assume that each alien visitor knew all other words of the language apart from the unknown (target) word.

Participants were asked to enter a descriptive phrase per line (not necessarily a whole sentence) and to try and write at least 4 phrases per word. They were given a maximum of one minute per concept, and they were not allowed to go back to the previous pages.

Before the real experiment, subjects were presented an example concept (not in the target list) and were encouraged to describe it while asking clarifications about the task.

All subjects returned the questionnaire so that for a concept we obtained, on average, descriptions by 36.48 German subjects (s.d. 1.24) and 34.34 Italian subjects (s.d. 1.72).

3.3 Transcription and Normalisation

The collected data were digitally transcribed and responses were manually checked to make sure that phrases denoting different properties had been properly split. We tried to systematically apply the

criterion that, if at least one participant produced 2 properties on separate lines, then the properties would always be split in the rest of the data set.

However, this approach was not always equally applicable in both languages. For example, *Transportmittel* (German) and *mezzo di trasporto* (Italian) both are compounds used as hypernyms for what English speakers would probably rather classify as *vehicles*. In contrast to *Transportmittel*, *mezzo di trasporto* is splittable as *mezzo*, that can also be used on its own to refer to a kind of vehicle (and is defined more specifically by adding the fact that it is used for transportation). The German compound word also refers to the function of transportation, but *-mittel* has a rather general meaning, and would not be used alone to refer to a vehicle. Hence, *Transportmittel* was kept as a whole and the Italian quasi-equivalent was split, possibly creating a bias between the two data sets (if the Italian string is split into *mezzo* and *trasporto*, these will be later classified as hypernym and functional features, respectively; if the German word is not split, it will only receive one of these type labels). More in general, note that in German compounds are written as single orthographic words, whereas in Italian the equivalent concepts are often expressed by several words. This could also create further bias in the data annotation and hence in the analysis.

Data were then normalised and transcribed into English, before annotating the type of semantic relation. Normalisation was done in accordance with McRae et al. (2005), using their feature norms as guidelines, and it included leaving habitual words like "normally,", "often", "most" etc. out, as they just express the typicality of the concept description, which is the implicit task.

3.4 Mapping to Relation Types

Normalised and translated phrases were subsequently labelled for relation types following McRae et al.'s criteria and using a subset of the semantic relation types described in Wu and Barsalou (2004): see section 4.1 below for the list of relations used in the current analysis.

Trying to adapt the annotation style to that of McRae et al., we encountered some dubious cases. For example, in the McRae et al.'s norms, *carnivore* is classified as a hypernym, but *eats_meat* as a behaviour, whereas they seem to us to convey essentially the same information. In this case, we decided to map both to *eats_meat* (behaviour).

Among other surprising choices, the normalised phrase *used_for_cargo* is seen by McRae et al. as a function, but *used_by_passengers* is classified as denoting the participants in a situation. In this case, we followed their policy.

While we tried to be consistent in relation labelling within and across languages, it is likely that our own normalisation and type mapping also include a number of inconsistencies, and our results must be interpreted by keeping this important *caveat* in mind.

The average number of normalised phrases obtained for a concept presented is 5.24 (s.d. 1.82) for the German participants and 4.96 (s.d. 1.86) for the Italian participants; in total, for a concept in our set, the following number of phrases was obtained on average: 191.28 (German, s.d. 25.96) and 170.42 (Italian, s.d. 25.49).

4 Results

The distribution of property types is analysed both class-independently and within each class (separately for German and Italian), and an unsupervised clustering analysis based on property types is conducted.

4.1 Distributional Analysis

We first look at the issue of how comparable the German and Italian data are, starting with a check of the overlap at the level of specific properties. There are 226 concept–property pairs that were produced by at least 10 German subjects; 260 pairs were produced by at least 10 Italians. Among these common pairs, 156 (i. e., 69% of the total German pairs, and 60% of the Italian pairs) are shared across the 2 languages. This suggests that the two sets are quite similar, since the overlap of specific pairs is strongly affected by small differences in normalisation (e. g., *has a fur*, *has fur* and *is hairy* count as completely different properties).

Of greater interest to us is to check to what extent property *types* vary across languages and across concept classes. In order to focus on the main patterns emerging from the data, we limit our analysis to the 6 most common property types in the whole data set (that are also the top 6 types in the two languages separately), accounting for 69% of the overall responses. These types are:

- category (Wu/Barsalou code: *ch*; "*pear is a fruit*")

- (external) part (WB code: *ece*; "*dog has 4 legs*")
- (external) quality (WB code: *ese*; "*apple is green*")
- behaviour (WB code: *eb*; "*dog barks*")
- function (WB code: *sf*; "*broom is for sweeping*")
- location (WB code: *sl*; "*skyscraper is found in cities*")

Figure 1 compares the distribution of property types in the two languages via a *mosaic plot* (Meyer et al., 2006), where rectangles have areas proportional to observed frequencies in the corresponding cells. The overall distribution is very similar. The only significant differences pertain to category and location types: Both differences are significant at the level $p < 0.0001$, according to a Pearson residual test (Zeileis et al., 2005).

For the difference in location, no clear pattern emerges from a qualitative analysis of German and Italian location properties. Regarding the difference in (superordinate) categories, we find, interestingly, a small set of more or less abstract hypernyms that are frequently produced by Italians, but never by Germans: *construction* (72), *object* (36), *structure* (16). In the these cases, the Italian translations have subtle shades of meaning that make them more likely to be used than their German counterparts. For example, the Italian word *oggetto* ("object") is used somewhat more concretely than the extremely abstract German word *Objekt* (or English "object", for that matter) – in Italian, the word might carry more of an "artifact, man-made item" meaning. At the same time, *oggetto* is less colloquial than German *Sache*, and thus more amenable to be entered in a written definition. In addition, among others, the category *vehicle* was more frequent in the Italian than in the German data set (for which one reason could be the difference between the German and Italian equivalents, which was discussed in section 3.3). Differences of this sort remind us that property elicitation is first and foremost a verbal task, and as such it is constrained by language-specific usages. It is left to future research to test to what extent linguistic constraints also affect deeper conceptual representations (would Italians be faster than Germans at recognising superordinate properties of concepts when they are expressed non-verbally?).

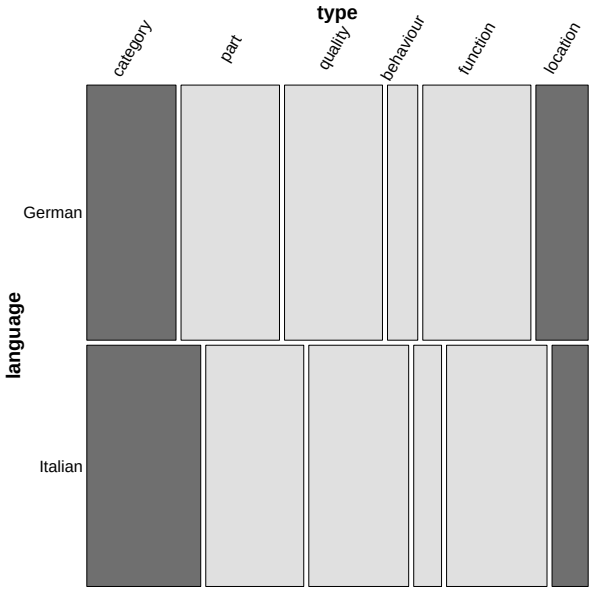


Figure 1: Cross-language distribution of property types

Despite the differences we just discussed, the main trend emerging from figure 1 is one of essential agreement between the two languages, and indicates that, with some caveats, salient property types may be cross-linguistically robust. We, thus, turn to the issue of how such types are distributed across concepts of different classes. This question is visually answered by the association plots in figure 2 on the following page.

Each plot illustrates, through rectangle heights, how much each cell deviates from the value expected given the overall contingency tables (in our case, the reference contingency tables are the language-specific distributions of figure 1). The sign of the deviation is coded by direction with respect to the baseline. For example, the first row of the left plot tells us, among other things, that in German behaviour properties are strongly over-represented in mammals, whereas function properties are under-represented within this class. Like in figure 1, shades of grey cue degrees of significance of the deviation (Meyer et al., 2003).

The first observation we can make about figure 2 is how, for both languages, a large proportion of cells show a significant departure from the overall distribution. This confirms what has already been observed and reported in the literature on English norms – see, in particular, Vinson et. al. (2008):

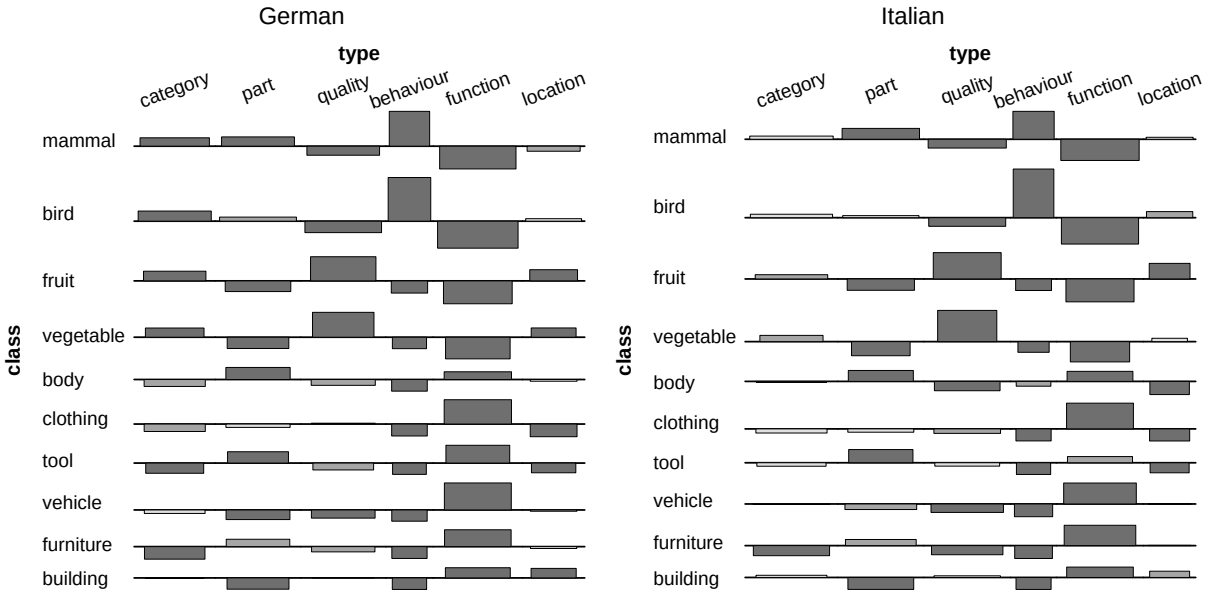


Figure 2: Distribution of property types across classes

property types are highly distinctive characteristics of concept classes.

The class-specific distributions are extremely similar in German and Italian. There is no single case in which the same cell is deviating significantly but in opposite directions in the two languages; and the most common pattern by far is the one in which the two languages show the same deviation profile across cells, often with very similar effect sizes (compare, e. g., the *behaviour* and *function* columns). These results suggest that property types are not much affected by linguistic factors, an intrinsically interesting finding that also supports our idea of structuring relation-based navigation in a multi-lingual dictionary using concept-class–specific property types.

The type patterns associated with specific concept classes are not particularly surprising, and they have been already observed in previous studies (Vinson and Vigliocco, 2008; Baroni and Lenci, 2008). In particular, living things (animals and plants) are characterised by paucity of functional features, that instead characterise all man-made concepts. Within the living things, animals are characterised by typical behaviours (they bark, fly, etc.) and, to a lesser extent, parts (they have legs, wings, etc.), whereas plants are characterised by a wealth of qualities (they are sweet, yellow, etc.) Differences are less pronounced within man-made objects, but we can observe parts as typical of tool and furniture descriptions. Finally, location is a more typical definitional characteristic of buildings (for clothing, nothing stands out, if not, perhaps, the pronounced *lack* of association with typical locations). Body parts, interestingly, have a type profile that is very similar to the one of (manipulable) tools – manipulable objects are, after all, extensions of our bodies.

4.2 Clustering by Property Types

The distributional analysis presented in the previous section confirmed our main hypotheses – that property types are salient properties of concepts that differ from a concept class to the other, but are robust across languages. However, we did not take skewing effects associated to specific concepts into account (e. g., it could be that, say, the property profile we observe for body parts in figure 2 is really a deceiving average of completely opposite patterns associated to, say, heads and hands). Moreover, our analysis already assumed a division into classes – but the type patterns, e. g., of mammals and birds are very similar, suggesting that a higher-level "animal" class would be more appropriate when structuring concepts in terms of type profiles. We tackled both issues in an unsupervised clustering analysis of our 50 target concepts based on their property types. If the postulated classes are not internally coherent, they will not form coherent clusters. If some classes should be merged, they will cluster together.

Concepts were represented as 6-dimensional vectors, with each dimension corresponding to one

of the 6 common types discussed above, and the value on a dimension given by the number of times that concept triggered a response of the relevant type. We used the CLUTO toolkit[4], selecting the `rbr` method and setting all other clustering parameters to their default values. We explored partitions into 2 to 10 clusters, manually evaluating the output of each solution.

Both in Italian and in German, the best results were obtained with a 3-way partition, neatly corresponding to the division into animals (mammals and birds), plants (vegetables and fruits) and objects plus body parts (that, as we observed above, have a distribution of types very similar to the one of tools). The 2-way solution resulted in merging two of the classes animals and plants both in German and in Italian. The 4-way solution led to an arbitrary partition among objects and body parts (and *not*, as one could have expected, in separating objects from body parts). Similarly, the 5- to 10-way solutions involve increasingly granular but still arbitrary partitions within the objects/body parts class. However, one notable aspect is that in most cases almost all concepts of mammals and birds, and vegetables and fruits are clustered together (both in German and Italian), expressing their strong similarity in terms of property types as compared to the other classes as defined here.

Looking at the 3-way solution in more detail, in Italian, the concept *horse* is in the same cluster with objects and body parts (as opposed to German, where the solution is perfect). The misclassification results mainly from the fact that for *horse* a lot of functional properties were obtained (which is a feature of objects), but none of them for the other animals in the Italian data. In German, some functional properties were assigned to both *horse* and *dog*, which might explain why it was not misclassified there.

To conclude, the type profiles associated with animals, vegetables and objects/body parts have enough internal coherence that they robustly identify these macro-classes in both languages. Interestingly, a 3-way distinction of this sort – excluding body parts – is seen as fundamental on the basis of neuro-cognitive data by Caramazza and Shelton (1998). On the other hand, we did not find evidence that more granular distinctions could be made based on the few (6) and very general types we used. We plan to explore the distribution across the remaining types in the future (preliminary clustering experiments show that much more nuanced discriminations, even among all 10 categories, can be made if we use all types). However, for our applied purposes, it is sensible to focus on relatively coarse but well-defined classes, and on just a few common relation types (alternatively, we plan to combine types into superordinate ones, e. g. external and internal quality). This should simplify both the automatic harvesting of corpus-based properties of the target types and the structuring of the dictionary relational interface.

Finally, the peculiar object-like behaviour of body parts on the one hand, and the special nature of horse, on the other, should remind us of how concept classification is not a trivial task, once we try to go beyond the most obvious categories typically studied by cognitive scientists – animals, plants, manipulable tools. In a lexicographic perspective, this problem cannot be avoided, and, indeed, the proposed approach should scale in difficultiese to even trickier domains, such as those of actions or emotions.

5 Conclusion

This research is part of a project that aims to investigate the cognitive salience of semantic relations for (pedagogical) lexicographic purposes. The resulting most salient relations are to be used for revising and adding to the word field entries of a multilingual electronic dictionary in a language learning environment.

We presented a multi-lingual concept description experiment. Participants produced different semantic relation type patterns across concept classes. Moreover, these patterns were robust across the two native languages studied in the experiment – even though a closer look at the data suggested that linguistic constraints might affect (verbalisations of) conceptual representations (and thus, to a certain extent, which properties are produced). This is a promising result to be used for automatically harvesting semantically related words for a given lexical entry of a concept class.

However, the granularity of concept classes has to be defined. In addition, to yield a larger number of usable data for the analysis, a re-mapping of the rare semantic relation types occurring in the actual data set should be conducted. Moreover, the stimuli set will have to be expanded to include, e. g., ab-

[4] URL `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

stract concepts – although we hope to mine some abstract concept classes on the basis of the properties of our concept set (colours, for example, could be characterised by the concrete objects of which they are typical).

To complement the production experiment results, we aim to conduct an experiment which investigates the perceptual salience of the produced semantic relations (and possibly additional ones), in order to detect inconsistencies between generation and retrieval of salient properties. If, as we hope, we will find that essentially the same properties are salient for each class across languages and both in production and perception, we will then have a pretty strong argument to suggest that these are the relations one should focus on when populating multi-lingual dictionaries.

Of course, the ultimate test of our approach will come from empirical evidence of the usefulness of our relation links to the language learner. This is, however, beyond the scope of the current project.

References

[Abel and Weber2000] Abel, Andrea and Vanessa Weber. 2000. ELDIT—A Prototype of an Innovative Dictionary. In Heid, Ulrich, Stefan Evert, Egbert Lehmann, and Christian Rohrer, editors, EURALEX *Proceedings*, volume 2, pages 807–818, Stuttgart.

[Baroni and Lenci2008] Baroni, Marco and Alessandro Lenci. 2008. Concepts and Properties in Word Spaces. *Italian Journal of Linguistics*. To appear.

[Boyd-Graber et al.2006] Boyd-Graber, Jordan, Christaine Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding Dense, Weighted Connections to WordNet. In *Proceedings of the Thirds International WordNet Conference*. Masaryk University Brno.

[Caramazza and Shelton1998] Caramazza, Alfonso and Jennifer R. Shelton. 1998. Domain–Specific Knowledge Systems in the Brain: The Animate–Inanimate Distinction. *Journal of Cognitive Neuroscience*, 10:1–34.

[Chanier and Selva1998] Chanier, Thierry and Thierry Selva. 1998. The ALEXIA system: The Use of Visual Representations to Enhance Vocabulary Learning. In *Computer Assisted Language Learning*, volume 11, pages 489–522.

[Fellbaum1998] Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

[Fontenelle1997] Fontenelle, Thierry. 1997. Using a Bilingual Dictionary to Create Semantic Networks. *International Journal of Lexicography*, 10(4):275–303.

[Garrard et al.2001] Garrard, Peter, Matthew A. Lambon Ralph, John R. Hodges, and Karalyn Patterson. 2001. Prototypicality, Distinctiveness, and Intercorrelation: Analyses of the Semantic Attributes of Living and Nonliving Concepts. *Cognitive Neuropsychology*, 18(2):125–174.

[Geckeler2002] Geckeler, Horst. 2002. Anfänge und Ausbau des Wortfeldgedankens. In Cruse, D. Alan, Franz Hundsnurscher, Michael Job, and Peter Rolf Lutzeier, editors, *Lexikologie. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*, volume 21 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 713–728. de Gruyter, Berlin – New York.

[Jansz et al.1999] Jansz, Kevin, Christopher Manning, and Nitin Indurkha. 1999. Kirrkirr: Interactive Visualisation and Multimedia From a Structured Warlpiri Dictionary. In *Proceedings of the 5th Australian World Wide Web Conference (AusWeb'99)*, pages 302–316.

[McRae et al.2005] McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behaviour Research Methods, Instruments & Computers*, 37(4):547–559.

[Meyer et al.2003] Meyer, David, Achim Zeileis, and Kurt Hornik. 2003. Visualizing Independence Using Extended Association Plots. In *Proceedings of DSC 2003*. Online at URL `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/`.

[Meyer et al.2006] Meyer, David, Achim Zeileis, and Kurt Hornik. 2006. The Strucplot Framework: Visualizing Multi–Way Contingency Tables With vcd. *Journal of Statistical Software*, 17(3):1–48.

[Summers1999] Summers, Della, editor. 1999. *Longman Language Activator. The World's First Production Dictionary*. Longman, Harlow.

[Vinson and Vigliocco2008] Vinson, David P. and Gabriella Vigliocco. 2008. Semantic Feature Production Norms for a Large Set of Objects and Events. *Behaviour Research Methods*, 40(1):183–190.

[Wu and Barsalou2004] Wu, Ling–ling and Lawrence W. Barsalou. 2004. Grounding Concepts in Perceptual Simulation: I. Evidence From Property Generation. Unpublished manuscript.

[Zeileis et al.2005] Zeileis, Achim, David Meyer, and Kurt Hornik. 2005. Residual–Based Shadings for Visualizing (Conditional) Independence. Technical Report 20, Department of Statistics and Mathematics, Wirtschaftsuniversität, Vienna.

The Computation of Associative Responses to Multiword Stimuli

Reinhard Rapp
Universitat Rovira i Virgili
Plaza Imperial Tarraco, 1
43005 Tarragona, Spain
reinhardrapp@gmx.de

Abstract

It is shown that the behaviour of test persons as observed in association experiments can be simulated statistically on the basis of the common occurrences of words in large text corpora, thereby applying the law of association by contiguity which is well known from psychological learning theory. In particular, the focus of this work is on the prediction of the word associations as obtained from subjects on presentation of multiword stimuli. Results are presented for applications as diverse as crossword puzzle solving and the identification of word translations based on non-parallel texts.

1 Introduction

The idea that human memory functions associatively goes back to Aristotle who formulated that the sequence of our memories is determined by the concepts of similarity and proximity (Strube, 1984:34). As early as 1879, Francis Galton tried to systematically observe human associative behaviour by introducing an association experiment. In this experiment, given a particular stimulus word, subjects had to respond with the first other word that occurred to them spontaneously. The resulting tables of associative responses are called association norms.

To explain the behavior documented in the association norms, in the literature a multiplicity of different mechanisms underlying human memory are proposed, thereby, for example, assuming phonological, morphological, syntactical, semantical, and contextual relations between words (Wettler, 1980). However, as yet there is no agreement whether these mechanisms should be considered of equal status, or if some may be derived from others.

Already in 1750 the physiologist David Hartley suggested that it may be possible to reduce the multiplicity of proposed association laws to only a single one based on temporal contiguity. This was formulated as one of the earliest psychological laws by William James (1890: 561): "Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before. This statement we may name the law of mental association by contiguity."

Assuming that the "objects" referred to in this law are words, the law of association by contiguity implies the following two phases:

1) *Learning phase:* When perceiving language, strong associative connections are developed between words that frequently occur in close temporal succession.
2) *Retrieval phase:* These associations determine the words that come to mind during generation. Only words that are strongly interconnected or have strong associations to external stimuli can be uttered or written down.

Pre-supposing the validity of the law of association, it should be possible to derive free word associations from the distribution of words in texts. Following Church & Hanks (1990), Rapp (2004), and Wettler et al. (2005) this actually seems to be successful. The recent simulation algorithms generate results which largely agree with the free word associations as found in the association norms. An example is shown in Table 1, where the observed and the simulated responses to the stimulus word *cold* are compared.

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 102–109
Manchester, August 2008

OBSERVED RESPONSE	NUMBER OF SUBJECTS	PREDICTED RESPONSE	NUMBER OF SUBJECTS
hot	34	hot	34
ice	10	winter	2
warm	7	weather	0
water	5	warm	7
freeze	3	water	5
wet	3	heat	1
feet	2	ice	10
freezing	2	wet	3
nose	2	wind	0
room	2	temperature	0
sneeze	2	shiver	0
sore	2	freeze	3
winter	2	rain	0

Table 1: Observed and predicted associative responses to the stimulus word *cold*.

When judging these results it should be kept in mind that among subjects there is some variation of responses. Therefore, the simulation results can be considered satisfactory if the difference between the predicted and the observed answers is on average not larger than the difference between an answer of an average test subject and the answers of the remaining test subjects.

In the current paper we try to build on these results. However, while most previous work considered only associations to individual stimulus words, the question to be dealt with here is whether the associative responses to several stimuli can likewise be predicted from the co-occurrences of words in texts. This is of considerable interest as all utterances and texts can be considered as accumulations of stimulus words, which together lead to a systematic activation of other words and concepts in the mind of the listener or reader.

How uniform the reactions of test subjects can be upon presentation of several stimulus words can be seen from examples like the word pairs *circus – laugh* or *King – girl* where subjects tend to think of *clown* and *princess*, respectively. Starting from the association norms for individual stimuli, the observed results are not always obvious. For example, in a large database of association norms, namely the *Edinburgh Associative Thesaurus* (Kiss et al., 1973), among the responses to *King* the word *princess* is completely missing, and the same is true for *girl*.

This means that the combination of stimulus words can lead to associations which are only weakly linked to the individual words and therefore cannot easily be deduced from conventional association norms. Accordingly it is not obvious whether the method used for the simulation of the associative behavior to single words can be extended in a straightforward way in the case of several stimulus words.

The organization of this paper is as follows: We first look at association norms collected for pairs of stimulus words. We then introduce a corpus-based algorithm that simulates the observed behavior which is applicable in the case of single or multiple stimuli. We then present some results of the algorithm and apply it to some related problems.

2 Association norms for word pairs

For individual English words, several association norms have been published, with the largest being the Edinburgh Associative Thesaurus. However, in the case of several stimulus words hardly any data seems to exist, with Rapp (1996, 1998) being an exception. This is a study that collected the responses of 31 subjects to pairs of German[2] nouns. In compiling these association norms, a list of 10 common German nouns had been selected, namely Mädchen (girl), Krankheit (illness), Junge (boy), Musik (music), Bürger (citizen), Erde (earth), Straße (street), König (King), Freude (joy), Sorge (worry). Then all 90 possible pairs of these words were constructed, and the answers of the subjects upon presentation of these pairs were collected. The subjects were asked to come up with the first word spontaneously coming to mind. In addition, associations to the individual words were also collected.

As for the pairs it turned out that word order did not have a noticeable effect on the responses, the responses to pairs differing only in word order were merged.

In Table 2 the associative responses as given by the test subjects for two sample pairs of stimulus words are listed. In comparison to responses to individual stimulus words, the responses to pairs of words are generally less uniform, i.e. there is considerably more variation in the case of word pairs. For example 25 of 31 test subjects come up with the association *Mädchen* (girl) given the stimulus word *Junge* (boy). In contrast, the most frequently mentioned associative response upon presentation of the stimulus pair *Junge Mädchen* (boy girl), which is *Kinder* (children), is given by only seven test persons.

[2] As we are not aware of such data for English, the current study was conducted for German, with translations given throughout the paper.

STIMULUS PAIR	ASSOCIATIVE RESPONSES
Erde (earth) Sorge (worry)	Umwelt (environment) 8, Umweltverschmutzung (environmental pollution) 5, Weltuntergang (end of the world) 2, ai (AI), Ausbeutung (exploitation), Katastrophen (catastrophe), Klimakatastrophe (climatical catastrophe), Krieg (war), Luft (air), Macht (might), Müll (garbage), Mutter (mother), Ozonloch (ozone hole), Resignation (resignation), Überbevölkerung (overpopulation), Umweltzerstörung (destruction of the environment), unfruchtbar (infertile), Verschmutzung (pollution), Zerstörung (destruction)
König (King) Mädchen (girl)	Prinzessin (princess) 15, Königin (queen) 3, Tochter (daughter) 2, Abhängigkeit (dependency), Dienerin (maid), Hochzeit (wedding), Kinderspiele (children's games), Kitsch (kitsch), Königspaar (royal couple), Märchen (fairy tale), Mißbrauch (abuse), Pferd (horse), Vater (father), Vorbild (model)

Table 2: Associations to the stimulus pairs "*Erde Sorge*" (earth worry) and "K*önig Mädchen*" (King girl). Figures indicate the number of subjects with the respective response, with the default being one.

For a more exact quantitative analysis of this observation a measure is needed for the homogeneity of the answers. For this purpose, it was computed how many subjects gave the same answer to a particular stimulus pair. On average, this was the case for 4% of the subjects. In comparison, the corresponding value for individual stimulus words is 15%. Thus the impression of a substantially larger homogeneity of the associative answers for individual stimuli is confirmed.

3 Simulation program

The simulation is based on the detection of statistical regularities of the common occurrences between the words in a large text corpus. As we did not have a large and at the same time balanced corpus of German at our disposal, we decided to use a corpus of the newspaper *Frankfurter Allgemeine Zeitung* (FAZ) comprising the years 1993 to 1996 (135 million words). As in the association experiment the subjects rarely answer with inflected forms or function words, for computational reasons we lemmatized this corpus (Lezius, Rapp & Wettler, 1998) and – based on a list of stop words – removed closed class words such as articles, pronouns, and particles.

To determine word co-occurrences, for each word in the corpus it was counted how often its close neighbors occurred within a text window of plus and minus six words. Assuming that approximately every second word is a function word, a window size of plus and minus six words after removal of the function words roughly corresponds to a window size of plus and minus 12 words without such pre-processing. This is a window size that corresponds with what had been found appropriate for the computation of associations in other studies (e.g. Rapp, 2004).

As the co-occurrence counts largely depend on overall word frequency, some association measure needs to be applied to eliminate this undesired influence. Many previous studies have shown that the log-likelihood ratio is well suited for this purpose (Dunning, 1993). It successfully eliminates word-frequency effects and emphasizes significant word pairs by comparing their observed co-occurrence counts with their expected co-occurrence counts. It can be expected that the log-likelihood ratio produces an accurate ranking of word pairs that highly correlates with human judgment (Dunning, 1993), although there are other measures which come close in performance (e.g. Rapp, 1998).

To compute the associations to pairs of stimulus words, it would in principle be possible to consider text positions where both stimulus words occur together, and to count the co-occurrence frequencies with their neighboring words. This would result in a three-dimensional association matrix whose first two dimensions correspond to the two stimulus words and whose third dimension corresponds to their associations. However, the problem of data sparseness would be very severe with such an approach, and it would not scale well if more than two stimulus words were considered.

We therefore propose another approach, which to our knowledge is novel in this context: The idea is that a potential associative response to a pair of stimulus words should have a strong and preferably symmetric associative connection to each of the stimulus words, and that a strong association to only one of them does not suffice. Such a behavior can usually be ensured by a multiplication.

However, we do not multiply the association strengths, as the log-likelihood ratio has an inappropriate (exponential) value characteristic. This value characteristic has the effect that a weak association to one of the stimuli can easily be overcompensated by a very strong association to

the other stimulus, which is not desirable. Instead of multiplying the association strengths, we therefore suggest to multiply their ranks. This improves the results considerably.

These considerations lead us to the following procedure: Given an association matrix of vocabulary *V* containing the log-likelihood ratios between all possible pairs of words, to compute the associative response given words *a* and *b*, the following steps are conducted:

1) For each word in *V* (by applying a search-and-compare operation on the association matrix) look up the ranks of words *a* and *b* in its list of associations, and compute the product of these ranks.
2) Sort the words in *V* according to these products, with the sort order such that the lowest value obtains the top rank (i.e. conduct a reverse sort).

Note that this procedure is somewhat time consuming as computations are required for each word in a large vocabulary.[3] On the plus side, the procedure is applicable to any number of stimulus words, and with increasing number of stimuli there is only a moderate increase in computational requirements. (The application presented in section 5.2 successfully processes 30 stimulus words.)

A minor issue is the assignment of ranks to words that have identical log-likelihood scores, especially in the frequent case of zero co-occurrence counts. In such cases, the assignment of possibly almost arbitrary ranks could adversely affect the results. We therefore suggest assigning corrected ranks, which are to be chosen as the average ranks of all words with identical scores.

With large numbers of stimuli, depending on the application it can be helpful to introduce a limit to the maximum rank, thereby reducing the effects of the sparse-data problem. The benefit of this measure is similar to smoothing, but more sophisticated smoothing methods can of course also be considered (as described, e.g. in Church & Gale, 1991). Note that for the current work we only used a rank limit of 10,000, but did not apply any sophisticated smoothing as this usually has little impact if the focus is mainly on the top ranks, as is the case here.

[3] Considerable time savings are possible by using an index of the non-zero co-occurrences.

4 Results

The algorithm as described above was applied to the FAZ corpus. That is, based on a window size of plus and minus six words, an association matrix with log-likelihood scores and (in both rows and columns) comprising all words with a corpus frequency of 200 or higher was computed. For each of the 45 word pairs, the top associations as resulting from the product of ranks were computed. To give some examples, the following tables show the outcome for a few stimulus pairs. Hereby, the columns in the tables have the following meanings:

1) rank
2) corpus frequency of association
3) score (product of stimulus ranks)
4) association

Junge Mädchen (boy girl)

1	247	11.33	fünfzehnjährig (15 year old)
2	2960	9.81	dreizehn (13)
3	398	9.72	gleichaltrig (same age)
4	86559	9.72	alt (old)
5	850	9.66	blond (blond)

Bürger Mädchen (citizen girl)

1	1276	11.51	brav (well behaved)
2	1268	7.26	unschuldig (innocent)
3	223	6.73	verängstigt (scared)
4	979	6.41	anvertrauen (to intrust)
5	362	5.97	belästigen (to molest)

Straße Mädchen (street girl)

1	2509	7.50	tanzen (to dance)
2	242	7.12	pflastern (to pave)
3	272	6.96	Bürgersteig (sidewalk)
4	529	6.87	Prostitution (prostitution)
5	4367	6.76	begegnen (to encounter)

Sorge Mädchen (worry girl)

1	317	7.03	elterlich (parental)
2	210	6.62	Burschen (fellows)
3	222	6.23	Beschneidung (concision)
4	7508	5.81	Eltern (parents)
5	271	5.77	zwölfjährig (12 year old)

Junge Krankheit (boy illness)

1	8891	7.33	leiden (to suffer)
2	3553	7.14	tödlich (lethal)
3	16468	7.04	sterben (to die)
4	423	6.83	Heilung (cure)
5	261	6.62	Schizophrenie (schizophrenia)

Straße Krankheit (street illness)

1	308	6.94	Tuberkulose (tuberculosis)
2	4704	6.74	Unfall (accident)
3	276	6.71	tückisch (malicious)
4	232	6.34	heimtückisch (malignant)
5	620	6.07	anstecken (to infect)

Straße Bürger (street citizen)

1	272	7.21	Bürgersteig (sidewalk)
2	235	7.18	Gibraltar (Gibraltar)
3	207	7.09	flanieren (to stroll)
4	242	7.02	pflastern (to pave)
5	366	6.58	Fußgängerzone (pedestrian zone)

Sorge Freude

1	6331	1.11	bereiten (to cause)
2	8747	9.21	Anlaß (occasion)
3	950	8.54	überwiegen (to outweigh)
4	27136	7.54	Grund (reason)
5	248	7.21	ungetrübt (untroubled)

If we look at all 45 word pairs, we obtain the following evaluation: Whereas an associative answer given by a subject is on average also given by 4% of the other subjects, only about 0.8% of the subjects give the answer produced in the simulation, i.e. the word ending up on the top rank. However, due to the low number of cases, this value may be subject to some sampling error.

A method less sensitive to sampling errors is to look at the overall simulation ranks of the subjects' responses. Hereby it is better to consider the median of the ranks rather than the mean, as the median's treatment of outliers is more appropriate. Note that when computing the median, associative responses given by n subjects obtain an n-fold higher weight. To further reduce the effects of outliers, only responses that are given by at least two subjects are taken into account.

Under these assumptions, the overall median (computed over all stimulus pairs) has a value of 245. With the total vocabulary of corpus frequency 200 and higher comprising about 25000 words, this value is at the 1% level. This compares to 12500 at the 50% level, which could be expected in the case of random behaviour.

5 Applications

5.1 Crossword puzzle solver

As crossword puzzles have definitions which usually consist of several words, the proposed algorithm can be applied as a crossword puzzle solver. In order not to reduce this task to a (for computers) relatively simple combinatorial problem, we hereby only restrict the ranked list of words as produced by the simulation program to those words that have the correct number of characters, but do not utilize as clues the common characters of horizontal and vertical words.

As an example, Figure 1 shows a crossword puzzle which is attributed to be the world's first one. It was designed by Arthur Wynne and published on December 21, 1913 in *The New York World*. Table 3 shows the definitions of this crossword puzzle together with the supposed solutions and the ranks of the respective words as computed by our algorithm based on three different corpora, namely the *British National Corpus* (BNC), the years 1990 to 1994 of the newspaper *The Guardian*, and the English part of the *Wikipedia XML Corpus* (Denoyer & Gallinari, 2006). These three corpora have a size of roughly 100, 150, and 300 million words, respectively. To allow a better judgment of the simulation results, the number of words of the respective length in the underlying vocabulary is specified in column 5.

This vocabulary was chosen to consist of all words that have a corpus frequency of 100 or higher in the BNC but did not occur in our list of about 200 function words. To this vocabulary, all words occurring in the crossword puzzle were added. The purpose of limiting the vocabulary was solely for computational reasons, as our algorithm is rather demanding with regard to both execution time and memory requirements.

Note that the BNC-based vocabulary was also used for the other somewhat larger corpora as not many words were missing there: In the Guardian corpus of the altogether 34,448 words all but 390 occurred at least one time, and in the larger Wikipedia corpus all but 131. We did not lemmatize the English corpora as in several cases inflected forms of words occurred in the definitions or in the solutions of the crossword puzzle.

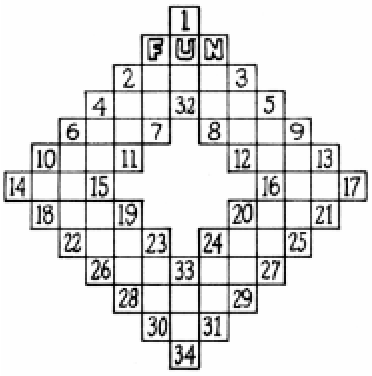


Figure 1: Crossword puzzle by Arthur Wynne.

As described in section 2, for counting the co-occurrences of words a window of plus and minus six words around a given word was considered, and for the computation of the associative strengths the log-likelihood ratio was used. Stop words were also removed from the corpora beforehand, but no lemmatization was conducted.

As many of the words used in the crossword puzzle are rare and several are outdated, solving this problem by a simulation is a non-trivial task. Nevertheless, for the Wikipedia corpus the algorithm got 8 of 31 answers ranked among the top five. When inspecting the examples that the algorithm got wrong, it appears that these are often the ones where humans would also have difficulties. For example, the solution "*side"* for "*to agree with*" got consistently poor ranks with all three corpora. On the other hand, rather surprisingly, the solution for "*such and nothing more*", namely "*mere*", received top rankings despite the fact that there are no salient content words in the description. This may be an indication that the algorithm grasps something that is related to cognitive processes. However, a similar example, namely "*what we all should be*" (→ *moral*) only obtains a reasonable ranking with the Wikipedia corpus. According to the average rankings (bottom line of Table 2), this corpus seems to be better suited for this task than the other two corpora.

5.2 Identifying word translations

The proposed core algorithm also has applications that may come somewhat unexpectedly. What we suggest here is to identify word transla-

POS.	DEFINITION	SOLUTION	LENGTH	WORDS OF THIS LENGTH	RANK BNC	RANK GUARDIAN	RANK WIKIPEDIA
2-3	what bargain hunters enjoy	sales	5	4254	1014	70	338
4-5	a written acknowledgement	receipt	7	5371	2	44	355
6-7	such and nothing more	mere	4	2916	16	17	4
10-11	a bird	dove	4	2916	17	87	4
14-15	opposed to less	more	4	2916	42	34	5
18-19	what this puzzle is	hard	4	2916	1486	115	384
22-23	an animal of prey	lion	4	2916	84	16	324
26-27	the close of a day	evening	7	5371	603	494	185
28-29	to elude	evade	5	4254	80	64	38
30-31	the plural of is	are	3	1424	238	119	412
8-9	to cultivate	farm	4	2916	2316	2783	1070
12-13	a bar of wood or iron	rail	4	2916	1658	1419	925
16-17	what artists learn to do	draw	4	2916	227	1437	86
20-21	fastened	tied	4	2916	15	2335	2078
24-25	found on the seashore	sand	4	2916	124	19	757
10-18	the fibre of the gomuti palm	doh	3	1424	585	279	711
6-22	what we all should be	moral	5	4254	4107	1163	51
4-26	a day dream	reverie	6	5371	489	572	2
2-11	a talon	sere	4	2916	676	803	492
19-28	a pigeon	dove	4	2916	36	8	1
F-7	part of your head	face	4	2916	63	20	143
23-30	a river in Russia	Neva	4	2916	174	413	3
1-32	to govern	rule	4	2916	48	9	13
33-34	an aromatic plant	nard	4	2916	616	2753	393
N-8	a fist	neif	4	2916	---	---	---
24-31	to agree with	side	4	2916	2836	2393	1387
3-12	part of a ship	spar	4	2916	2693	1932	90
20-29	one	tane	4	2916	2814	2773	2680
5-27	exchanging	trading	7	5371	3444	5216	2347
9-25	sunk in mud	mired	5	4254	3	2	1
13-21	a boy	lad	3	1424	3	2	2
AVERAGE RANK					891.6	922.3	520.2

Table 2: Crossword puzzle definitions and the computed ranks of their solutions based on three corpora. ('---' means that a solution does not occur in a corpus (not taken into account when computing average ranks).

tions from monolingual English and German corpora, i.e. from corpora that are not translations of each other (Rapp, 1999). This is a rather difficult task.

As our textual basis, for German we use the FAZ corpus as described above, with exactly the same pre-processing. For English we use a similarly sized corpus of the newspaper "The Guardian", with analogous pre-processing.

We apply a two-stage procedure to compute the translation of a source language word: First, by considering the log-likelihood ratios, its strongest source language associations are determined and translated to the target language using a small pocket dictionary. Hereby, associations that are missing in the dictionary are discarded, and of the remaining associations only the top 30 are selected.

The second step exactly corresponds to the computation of associations when given multiple stimulus words as described above. That is, for each word in the target language vocabulary (comprising all words that in the Guardian corpus occur with a frequency of 100 or higher) the ranks of the 30 translations are determined, and the product of these ranks is computed. The word obtaining the smallest value for the product is considered to be the translation of the source language word. This algorithm turned out to be a significant improvement over the previous algorithm described in Rapp (1999) as it provides a better accuracy and at the same time a considerably higher robustness.

Based on this novel algorithm, a large dictionary for German to English was computed. As for the translation of the source language vectors a base dictionary is required, we adapted for this purpose a small Collins pocket dictionary which comprised in the order of 20 000 entries. In essence, the adaptation procedure involves deriving word equations from the dictionary, each consisting of the source word and its first translation as mentioned in the dictionary.

To give an impression of the results, the following tables show the top ten computed translations for the six words *Historie* (history), *Leibwächter* (bodyguard), *Raumfähre* (space shuttle), *spirituell* (spiritual), *ukrainisch* (Ukranian), and *umdenken* (rethink). Hereby, the columns have the following meanings:

1) Rank of a potential translation
2) Corpus frequency of translation
3) Score assigned to translation
4) Computed translation

Historie (history)

1	29453	13.73	history
2	4997	12.87	literature
3	4758	8.74	historical
4	2670	0.67	essay
5	6969	0.11	contemporary
6	18909	-1.72	art
7	18382	-2.81	modern
8	15728	-4.31	writing
9	1447	-5.52	photography
10	2442	-5.53	narrative

Leibwächter (body guard)

1	949	40.02	bodyguard
2	5619	23.34	policeman
3	2535	8.18	gunman
4	26347	3.69	kill
5	9180	2.92	guard
6	401	-0.56	bystander
7	815	-1.24	POLICE
8	8503	-2.33	injured
9	2973	-3.23	stab
10	1876	-3.58	murderer

Raumfähre (space shuttle)

1	1259	46.20	shuttle
2	666	26.25	Nasa
3	473	25.95	astronaut
4	287	25.76	spacecraft
5	1062	16.92	orbit
6	16086	11.72	space
7	525	9.50	manned
8	125	7.69	cosmonaut
9	254	5.24	mir
10	7080	3.70	plane

spirituell (spritual)

1	2964	56.10	spiritual
2	1380	8.34	Christianity
3	7721	8.08	religious
4	9525	4.10	moral
5	1414	0.63	secular
6	5685	0.06	emotional
7	4678	-1.04	religion
8	6447	-1.49	intellectual
9	8749	-2.25	belief
10	8863	-4.07	cultural

ukrainisch (Ukrainian)

1	1753	50.69	Ukrainian
2	22626	39.88	Russian
3	3205	29.25	Ukraine
4	34572	23.63	Soviet

5	978	21.13	Lithuanian
6	1005	18.88	Kiev
7	10968	15.07	Gorbachev
8	10209	14.51	Yeltsin
9	16616	13.38	republic
10	502	11.71	Latvian

umdenken (rethink)

1	1119	20.76	rethink
2	248	15.46	reassessment
3	84109	13.39	change
4	12497	12.13	reform
5	236	10.00	reappraisal
6	9220	9.97	improvement
7	5212	9.48	implement
8	1139	8.25	overhaul
9	13550	7.89	unless
10	9807	7.88	immediate

6 Summary

It could be shown that word associations to multiple stimuli as collected from test persons can be predicted with reasonable accuracy using a simulation program that analyzes the co-occurrences of words in texts.

This result makes the automatic construction of an associative thesaurus of responses to multiple stimuli feasible. Note that such a thesaurus could not realistically be compiled by collecting the responses of human subjects as there are too many possible combinations of stimuli.

Finally, by looking at two sample applications we showed the pracical utility of the method. In principle, there should be many more applications, as all utterances and texts can be considered as collections of stimulus words. A notable one is search word generation in the context of internet search engines.

Of course, all existing algorithms for speech and text processing, although often not claiming any cognitive plausibility, necessarily also have some implicit mechanisms that deal with multiword stimuli. We nevertheless hope that the specific perspective that we presented here may add to a better understanding of the underlying cognitive mechanisms, and that it offers a systematic way of approaching these challenges.

7 Acknowledgments

This research was in part supported by a Marie Curie Intra European Fellowship within the 6th European Community Framework Programme.

References

Church, Kenneth W.; Gale, William (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5(1), 19–54.

Church, Kenneth W., Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Denoyer, Ludovic; Gallinari, Pattrick (2006). The Wikipedia XML Corpus. *SIGIR Forum*, 40(1), 64–69.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Galton, Francis. (1879). Psychometric experiments. *Brain,* 1, 149–162.

James, William (1890). *The Principles of Psychology*. New York: Dover Publications.

Kiss, George R.; Armstrong, Christine; Milroy, Robert; Piper, James (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley, N. Hamilton-Smith (eds.): *The Computer and Literary Studies*. Edinburgh University Press.

Lezius, Wolfgan; Rapp, Reinhard; Wettler, Mafred (1998). A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. *Proceedings of COLING ACL 1998*, Montreal, 743–747.

Rapp, Reinhard (1996). *Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz.* Hildesheim: Olms.

Rapp, Reinhard (1998). Das Kontiguitätsprinzip und die Simulation des Assoziierens auf mehrere Stimuluswörter. In: B. Schröder, W. Lenders, W. Hess, T. Portele: *Computer, Linguistik und Phonetik zwischen Sprache und Sprechen.* Frankfurt am Main: Peter Lang, 261–272.

Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. In: *Proceedings of the 37th Annual Meeting of the ACL,* College Park, MD, 519–526.

Rapp, Reinhard (2004). Word Sense Induction as Statistical Pattern Recognition. In: Ernst Buchberger (ed.): *Tagungsband der 7. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, Universität Wien, 161–168.

Strube, Gerhard (1984). *Assoziation.* Berlin: Springer.

Wettler, Manfred (1980). *Sprache, Gedächtnis, Verstehen*. Berlin: de Gruyter.

Wettler, Manfred; Rapp, Reinhard; Sedlmeier, Peter (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, 12(2), 111–122.

Author Index