Proceedings of the Workshop

Semantic Relations.

Theory and Applications

18 May 2010

Editors:
Verginica Barbu Mititelu, Viktor Pekar, Eduard Barbu

The Workshop Programme

Tuesday, May 18

09:00 – 09:10 Welcome and introduction
Verginica Barbu Mititelu

09:10 – 09:35 Semantic relations of Adjectives and Adverbs in Estonian WordNet
Kadri Kerner, Heili Orav, Sirli Parm

09:35 – 10:00 PredXtract, a generic platform to extract in texts predicate argument structures (PAS)
Elisabeth Godbert, Jean Royaute

10:00 – 10:25 Discovering semantic relations by means of unsupervised sense clustering
Marianna Apidianaki

10:30 – 11:00 Coffee break

11:00 – 11:25 From the People's Synonym Dictionary to fuzzy synsets – first steps
Lars Borin, Markus Forsberg

11:25 – 11:50 Consistency of Sense relations in a Lexicographic Context
Peter Meyer, Carolin Muller-Spitzer

11:50 – 12:15 Natural and contextual constrains for domain-specific relations
Pilar Leon Arauz, Pamela Faber

12:00 – 13:00 Invited talk:
Wheels for the mind of the language producer: microscopes, macroscopes, semantic maps and a good compass
Michael Zock

Workshop Organisers

Verginica Barbu Mititelu, RACAI
Viktor Pekar, Oxford University Press
Eduard Barbu, Center for Mind/Brain Sciences

Workshop Programme Committee

Dan Cristea, University Al. I. Cuza of Iassy
Brigitte Endres-Niggemeyer, University of Applied Sciences and Arts
Amac Herdagdelen, Center for Mind/Brain Sciences
Diana Inkpen, University of Ottawa
Radu Ion, RACAI
Gerhard Kremer, Center for Mind/Brain Sciences
Claudia Kunze, Qualisys GmbH
Gianluca Lebani, Center for Mind/Brain Sciences
Lothar Lemnitzer, Berlin-Brandenburgische Akademie der Wissenschaften
Alessandro Lenci, Universita di Pisa
M. Lynne Murphy, University of Sussex
Reinhard Rapp, Universitat Rovira e Virgili
Didier Schwab, Laboratoire d'Informatique de Grenoble
Dan Tufiş, RACAI

Table of Contents

Author Index

Forward

Semantic relations have been a subject of interest of various disciplines since ancient times. The 20th century structural semantics has fostered new perspectives on semantic relations as the basis for lexicon organization. More recently, semantic relations have become a major theme of interest of Computational Linguistics, as they present a convenient and natural way to organize huge amounts of lexical data in ontologies, wordnets and other machine-readable lexical resources. Semantic relations are, thus, a key to various important practical NLP tasks, involving semantic analysis and generation of text, such as Information Extraction, Question Answering, Automatic Summarisation, Knowledge Acquisition and many others. On the other hand, practical interests in these tasks have stimulated further linguistic research into the nature of semantic relations from both the paradigmatic and syntagmatic perspective, as can be seen from a growing attention to corpus-based studies of the subject, as well as to complex analysis of large manually compiled lexical resources. So, at present we witness that these two areas increasingly tend to mutually support and foster each other's advancements.

Over the past decades, theoretical linguists have made considerable progress collecting, defining and providing detailed characterization to semantic relations (synonymy, hyponymy, homonymy, polysemy, etc.) from the point of view of their concrete organisation within large-scale models of the lexical systems of languages, such as WordNet. They have also extensively studied the relationships between lexical meaning and its surface realization, i.e. the lexical and syntactic patterns between words or phrases that express a certain semantic relation. These insights proved to be extremely useful for language engineers. The described properties of semantic relations have allowed an economical design of language resources - for example, the transitivity of the hyponymy relation have been found to be conveniently reflect the hierarchical organization of nouns, which mirrors the inheritance of properties in natural languages.

In order to create lexical resources in a fast and cheap way, NLP research has been aiming to automatically extract the knowledge from corpora and proposed a broad variety of methods, including, but not limited to, lexico-syntactic patterns, distributional similarity, knowledge-based methods exploiting complex lexical resources, machine learning techniques operating on a broad variety of orthographical, morphological, syntactic and lexical features of text, etc. This work continues to draw upon knowledge gained from linguistic exploration of corpora and large lexical resources, which provides insight on the concrete nature of semantic relations and specifics of their organisation in the lexical systems of natural languages.

At this workshop, we aimed at bringing together researchers in computational linguistics and lexical semantics, discussing theoretical and practical aspects of semantic relations and answering the question of how computational linguists could benefit from the work done by theoretical linguists and vice versa. Specifically, in the call for papers we solicited papers on the following topics:

- Knowledge representation and semantic relations
- Extraction of semantic relations from various sources (lexical ontologies, corpora, Wikipedia, WWW)
- Exploitation of semantic relations in NLP applications
- Distributional methods for recognition of semantic relations
- Lexico-syntactic patterns and semantic relations
- Machine Learning approaches to recognition of semantic relations
- Semantic relations in language generation
- Semantic relations and terminology

In fact, most of these topics lie at the heart of the papers that were accepted to the workshop.

We would like to thank all the authors who submitted papers, as well as the members of the Program Committee for the time and effort they contributed in reviewing the papers. We are grateful to Michael Zock for accepting to give an invited talk. The talk is available as a paper in the proceedings.

The Editors

Discovering semantic relations by means of unsupervised sense clustering

Marianna Apidianaki

Language and Translation Technology Team (LT^3)
Faculty of Translation Studies, University College Ghent, Belgium
marianna.apidianaki@hogent.be

Abstract

Electronic sense inventories are needed for Word Sense Disambiguation (WSD) in Natural Language Processing (NLP) applications. However, existing sense inventories are criticized for being inadequate for this task. The fine granularity of the senses described in these resources is not always needed in NLP applications. Furthermore, the senses are listed without any information on their distinguishability and their relations. However, this type of knowledge can be automatically extracted from textual data using machine learning techniques. In this paper, we show how an unsupervised sense induction method permits to capture two different types of semantic relations: first, semantic similarity relations between the translation equivalents of ambiguous words; second, relations between their senses, which are automatically identified by means of semantic clustering. We analyze the results of this method and compare them with the semantic descriptions provided in a typical multilingual semantic resource.

1. Introduction

Pre-defined sense inventories are often criticized as being inadequate for Word Sense Disambiguation (WSD) in Natural Language Processing (NLP) applications. The main reason is that these resources contain a high number of too fine-grained senses which are listed without any information on their relations. This lack is partly due to the incapacity of the sense enumeration techniques used in lexicography to justify a distinction between different types of ambiguity (Dolan, 1994; Pustejovsky, 1995). Additionally, it is explained by the fact that the majority of the resources were initially developed for use by humans, who can identify word sense relations even if the relevant information is not explicitly mentioned.

Nevertheless, the fine granularity of the semantic descriptions found in the existing resources does not seem to be necessary for efficient WSD. According to Ide and Wilks (2007), NLP applications, when they need WSD, seem to need homograph-level disambiguation. Finer-grained distinctions are rarely needed, and when they are, more robust and different kinds of processing are required.

The fine granularity of the semantic descriptions found in existing resources, combined to the great divergences in their structure and content, put a hindrance to their compatibility. Their exploitation in multilingual applications is rather problematic as well, given the difficulty to establish correspondences between fine-grained senses of words and their translation equivalents in other languages (Specia et al., 2006).

However, the semantic information needed for WSD can be acquired directly from texts without recourse to pre-defined inventories. In this paper, we first analyze the problems raised during WSD by the fine granularity of word senses and we explain how it is possible to obtain coarser senses from pre-defined inventories. Then, we argue in favor of data-driven semantic analysis methods. We show how an unsupervised sense induction method can reveal two types of semantic relations: a) the relations between the translation equivalents (TEs) of ambiguous words of one language in another language, and b) the relations between the senses of the ambiguous words, which are correspond to sense-clusters of their TEs. The results of this method are analyzed and compared to the descriptions provided in a predefined multilingual semantic resource.

2. Using pre-defined resources for WSD

2.1. Sense granularity

Electronic sense inventories provide the list of the candidate senses of polysemous words that are needed for WSD. The task of a WSD algorithm is to select, from this list, the most appropriate sense for each new occurrence of a polysemous word in texts. However, this selection is complicated when the WSD algorithm is confronted to a high number of fine-grained senses. This criticism has been mainly formulated by reference to WordNet, widely exploited in WSD tasks (Edmonds and Kilgarriff, 2002; Ide and Wilks, 2007).

The pertinence of the selection of one among close senses can be doubted as well, as it can lead to arbitrary decisions in cases where the occurrences of the polysemous words could correspond to more than one specific fine-grained sense. Given that inter-sense relations are not described, the selection would provoke a loss of useful information on the semantics of the word (Dolan, 1994).

The difficulty of selecting one among fine-grained senses for a new occurrence of a polysemous word is also observed in human annotation tasks. The inter-annotator agreement for word-sense tagging is lower when language users are asked to assign refined sense tags – such as those found in WordNet – especially when the definition entries are short and only a few or no example sentences are provided for the usage of each word sense (Veronis, 1998; Ng et al., 1999).

2.2. Sense clustering

The criticisms addressed to existing semantic resources, combined with some scepticism concerning the need for fine-grained semantic distinctions in NLP applications, enhanced the development of methods for deriving coarser sense inventories from existing ones (Peters et al., 1998; Mihalcea and Moldovan, 2001; Navigli, 2006; Navigli et al., 2007). These *sense clustering* methods, developed by

reference to WordNet and EuroWordNet, discover the relations between the sets of synonyms (synsets) that describe the senses of the words and collapse them into clusters. These methods perform clustering by exploiting different types of information: information concerning the similarity of the words found in the synsets describing different senses; information on the similarity of the relations between them and other synsets of the network; probabilistic information extracted from corpora; syntactic criteria concerning alternations with similar subcategorization frames; semantic criteria concerning the semantic class of arguments, the subject domain and the underlying predicate-argument structures (Resnik, 1995; Jiang and Conrath, 1997; Mihalcea and Moldovan, 2001; Palmer et al., 2006). The sense granularity reduction performed has been shown to improve the performance of WSD. It also facilitates the establishment of correspondences between the senses described in different resources, increasing their compatibility.[1] Nevertheless, these knowledge-based clustering methods have been developed by reference to specific resources and cannot be generalized. This limitation and the increasing availability of text corpora have fostered the development of unsupervised methods capable of acquiring information on sense relations directly from texts.

3. Data-based semantic analysis

3.1. Discovering word sense relations from texts

Methods for discovering word sense relations from textual data have been developed in a monolingual as well as in a bi- and multi-lingual context. Monolingual methods are based on the distributional hypotheses of meaning and of semantic similarity, according to which semantically similar words present similar distributional behavior (Miller and Charles, 1991). The cooccurrences of the words in texts, or the syntactic frames in which they occur, constitute their sets of context features. The similarity of these sets reveals the similarity of the corresponding words (Church and Hanks, 1990; Dagan et al., 1993; Pereira et al., 1993; Pantel and Lin, 2002).

In a bi- (or multi-) lingual setting, word sense relations can be discovered by using translational information. In this case, the context of the SL words, that serves to calculate their similarity, corresponds to their translation equivalents (TEs) in other languages. The TEs are used to build vectors for the SL words whose similarity shows their semantic relatedness (van der Plas and Tiedemann, 2006). However, apart from revealing the relations of SL words, their TEs can also serve to analyze their semantics.

3.2. Cross-lingual sense induction

In the cross-lingual approach to sense induction, the TEs of the instances of polysemous words in parallel corpora can be used for identifying their senses. Ide et al. (2002) and Tufis et al. (2004) build vectors for the instances of SL words in a multilingual corpus by using as features their TEs in different languages. The vectors and the corresponding SL instances are clustered and the generated clusters describe the senses of the words. However, this method needs parallel corpora in many different languages, which are hard to be found. Additionally, the clustering is performed by an agglomerative algorithm which creates disjoint clusters. Consequently, the acquired senses are disjoint and their relations are not taken into account.

Another common approach to cross-lingual sense induction consists in using each TE of a polysemous word as describing one distinct sense. The most important merits of this approach are that it permits to bypass the subjectivity issue inherent in sense identification tasks and to derive senses relevant for translation (Resnik and Yarowsky, 1998).[2] Nevertheless, the TEs of polysemous words may not always constitute valid sense indicators.

In cases of *parallel* (or *translation*) *ambiguity*, for instance, the TEs present the same ambiguity as the SL word.[3] Ide and Wilks (2007) describe cases where the same historical processes of sense "chaining" occurs in different languages and the words extend their original sense in the same way.[4] Relying on cross-language lexicalization for sense distinction in such cases would disregard the sense deviations characterizing the words and would lead to the conclusion that both have a single sense.

Another danger in using TEs as straightforward sense indicators is that they may carry senses valid only in the TL which should not indicate a sense split in the SL. This happens, for instance, when a generic word in one language describes senses expressed by distinct lexical elements in another.[5]

Additionally, given that translators often use synonyms and near-synonyms in order to avoid repetitions in the translated texts, it is common that the TEs of a polysemous word be semantically similar in the target language (TL).[6] Consequently, these TEs translate the same sense of the SL word and should not be used to induce senses in the SL.

Finally, the senses induced by using the TEs as straightforward sense indicators are uniform : clear-cut and finer sense distinctions are listed without any description of their relations. The theoretical and practical problems posed by this cross-lingual approach to sense identification are discussed more thoroughly in Apidianaki (2008) and Apidianaki (2009). In the following section we show how the re-

[1] In EuroWordNet (Vossen, 1999), where the fine-grained English WordNet served as an Interlingual Index (ILI), the creation of coarse-grained inter-lingual entries facilitated the correspondences between equivalent senses in different languages.

[2] It has been adopted in the multilingual tasks of the Senseval (Chklovski et al., 2004) and SemEval (Jin et al., 2007) exercises and in works on WSD in Machine Translation (Cabezas and Resnik, 2005; Carpuat et al., 2006).

[3] An example of this type of ambiguity is the TE of the English noun *interest* in French, *intérêt*, which also carries the "financial" and "personal" senses of the English word (Resnik and Yarowsky, 1998).

[4] The English *wing* and its equivalent *ala* in Italian which both have extended their original sense from birds to airplanes, to buildings and to soccer positions.

[5] For example, Japanese has different words for "wear", depending on what part of the body is involved, but this distinction is not performed in English (Gale et al., 1993).

[6] According to Baker (1996), this particularity can be considered as a universal feature of translated texts.

ambiguous word	sense-clusters
movement	a. {μετακίνηση(metakinisi), κίνηση(kinisi), διακίνηση(diakinisi)}
	b. {κίνηση(kinisi), διακίνηση(diakinisi), κυκλοφορία(kikloforia)}
	c. {μετακίνηση(metakinisi), διακίνηση(diakinisi), κινητικότητα(kinitikotita)}
	d. {κίνημα(kinima)}
plant	a. {μονάδα(monada), εγκατάσταση(egkatastasi)}
	b. {σταθμός(stathmos), εργοστάσιο(ergostasio)}
	c. {σταθμός(monada), μονάδα(monada)}
	d. {φυτό(fyto)}

Table 1: Sense-clusters of "movement" and "plant"

lations between senses can be automatically identified during cross-lingual sense induction.

4. Semantic clustering

4.1. The method

An unsupervised sense induction method capable of revealing the relations between word sense from parallel corpora has been proposed in (Apidianaki, 2008). The method identifies the senses of polysemous words by clustering their TEs in another language, on the basis of their semantic similarity. The sense-clustering results show : a) the semantic relations between the TEs of polysemous SL words, and b) the relations between their senses.

The clustering is performed by exploiting, on the one hand, translational information relative to the TEs that translate the polysemous words in a parallel training corpus and, on the other hand, distributional information relative to the SL instances that are translated by each TE in the corpus. The semantic proximity of the TEs is estimated by combining these sources of information.

The method is trained on a bi-lingual parallel corpus, which has been sentence- and word- aligned, lemmatized and tagged by part-of-speech. The aligned sentences where a polysemous SL word occurs are extracted and grouped by reference to its TEs. Then, the SL sentences are analyzed: a frequency list is built for each TE, which contains the lemmas of the content words that occur in the context of the SL word whenever it is translated by this TE. These SL content words form a feature set that is assigned to the TE. The feature sets of the TEs are compared pairwise by using a modified version of the Weighted Jaccard Coefficient (Grefenstette, 1994), presented in Apidianaki (2008). The results of this similarity calculation reveal the semantic similarity of the corresponding pairs of TEs. The assumptions underlying this procedure is that the instances of the SL word that occur in similar contexts are semantically similar (Miller and Charles, 1991), and that the TEs that translate similar instances are semantically close as well.

The results of the pairwise similarity calculations are exploited by a clustering algorithm (SEMCLU) that groups the most similar TEs into clusters. A similarity threshold is defined dynamically for each polysemous word which corresponds to the mean of the scores assigned to the pairs of TEs. A pair of TEs having a score above the threshold is considered as having a pertinent semantic relation.

The TEs grouped in clusters are most often near-synonyms but it is possible to find TEs with other relations as well (hyponyms or hyperonyms) that translate the same SL sense in the corpus. The clusters obtained for a polysemous SL word are similar to WordNet synsets and describe its senses. This sense clustering procedure permits to analyze the semantic relations of the TEs and to avoid considering semantically similar TEs as indicators of distinct senses.

4.2. Fuzzy clustering

Apart from the semantic relations between the TEs of ambiguous words, the sense induction method described above also captures the relations between the senses of the ambiguous words. An interesting property of the SEMCLU algorithm, which constitutes the core of this method, is that it permits to perform a *fuzzy clustering*. This means that the resulting sense clusters are not disjoint but may present overlaps. The overlaps of the clusters are formally described by the non-empty intersection of their elements. This property of the algorithm is very important for sense induction: given that the clusters describe word senses, their overlaps can be perceived as describing the relations between the corresponding senses.

Capturing word sense relations gives the possibility to perform a differentiation between close and distant (or antagonistic) senses. Furthermore, the proximity of the senses can serve to modify the *granularity* of the proposed sense descriptions. Overlapping clusters often describe nuances or sub-senses of coarse-grained senses. Consequently, their merging makes it possible to obtain a description of the main sense distinctions characterizing a word.

A side benefit of this type of representation is that it allows to take into account the cases of translation ambiguity. Such cases are observed when a TE is found in the intersection of clusters. This means that the TE is ambiguous between the senses described by the clusters, which may be distant or close, and translates both in the TL.

4.3. Clustering examples

The sense induction method presented above builds bilingual sense-cluster resources for different language pairs automatically. What is needed is a parallel training corpus and tools for sentence and word alignment, part-of-speech tagging and lemmatization, which are available in many languages.

In this paper, we focus on the clustering performed for two polysemous English nouns (*movement* and *plant*). Our intention is to analyze the semantic representations obtained

by this method from a qualitative point of view. The results presented here are obtained by training the sense induction method on the English–Greek part of the multilingual INTERA parallel corpus (Gavrilidou et al., 2004).

– movement

The English noun *movement* has six TEs in Greek in the training corpus: κυκλοφορία (kikloforia / 251), διακίνηση (diakinisi / 38), κίνηση (kinisi / 28), μετακίνηση (metakinisi / 19), κίνημα (kinima / 11), κινητικότητα (kinitikotita / 6).[7] In the "traditional" approach to cross-lingual sense induction, each TE would describe one distinct sense of the English word and, consequently, *movement* would be considered as having six distinct senses. However, this analysis would not be correct because some of the TEs of the word are semantically related in the TL.

In the clustering solution generated by SEMCLU, the semantically similar TEs of *movement* are grouped into clusters which describe its senses. The clustered TEs translate occurrences of *movement* that are semantically related, i.e. found in similar contexts. The clusters obtained for *movement* are described in Table 1 and schematically illustrated in Figure 1.

We observe that the clusters (a), (b) and (c) share some elements and that they overlap. The TEs found in the overlapping clusters translate the "physical movement" sense of the word. So, each cluster describes a nuance of this coarse sense. However, the cluster (d) is clearly distinguished from the others. Its unique element (κίνημα)translates the sense of "social movement".

The "physical" and "social" senses of *movement* are more distant than those described by the overlapping clusters. This is clearly illustrated in Figure 1. The relations and the distinctions identified between the clusters reflect the different status of the corresponding senses. In the following section, we will show how this information can be exploited in order to modify the granularity of the obtained senses.

Figure 1: Sense-clusters of "movement"

– plant

The TEs of *plant* in the training corpus are the following : φυτό(fito94), μονάδα(monada15), εγκατάσταση (egkatastasi14), εργοστάσιο(ergostasio14), σταθμός (stathmos7). The clusters obtained for this word are described in Table 1 and illustrated in Figure 2.

[7]The TEs that translate less than five instances of the SL word in the corpus are not considered, as the available contextual information is not sufficient for efficient processing.

Figure 2: Sense-clusters of "plant"

As in the case of *movement*, some of the obtained sense-clusters overlap while others are disjoint. The relations and distinctions of the clusters indicate the relations between the corresponding senses of the SL word. *Plant* constitutes a case of homonymy where the implicated lexical units carry two distinct senses, the "botanical" and the "industrial" senses. The "industrial" sense is translated by the TEs found in the overlapping clusters ((a), (b) and (c)), which describe nuances of this coarse sense, while the "botanical" sense of the word is described by the disjoint cluster (d). The TE found in this cluster (φυτό), is the most frequent TE of *plant* in the corpus and the only that translates its "botanical" sense. So, the distinction between the disjoint and the overlapping clusters reflects the coarse distinction between the two non-related senses of *plant*.

It is important to note that the isolation of a TE in a disjoint cluster does not always indicate a clear-cut semantic distinction. This isolation may also be provoked by the low-frequency of a TE, because of data sparseness.[8] Nevertheless, the high frequency of a TE (like φυτό) constitutes a reliable clue for spotting a semantic distinction, given that the relative contextual information is sufficient for identifying the semantic relation of the TE to the other TEs of the SL word, if such a relation exists.

The creation of overlapping and disjoint clusters reflects the differences in the nature and status of the obtained senses. This information on the distinguishability of the senses and on their relations is important from a lexicographic point of view, as it offers the possibility to identify different types of ambiguity that are not taken into account in other semantic resources. It also permits to obtain a dynamic representation of the semantics of words that can be adapted to the uses of the inventory. In the next section, we show how the granularity of the senses can be modified by exploiting the information on their relations found in the sense-clustering results.

4.4. Sense granularity modification

The semantic representations generated by the sense induction method presented in this paper give the possibility to automatically modify the granularity of the obtained senses. The overlapping clusters describe fine-grained senses of the SL words or nuances of coarse senses. The clusters may also overlap because of the lack of pertinent links between some of their elements, which may be due to data-sparseness and not to the absence of a semantic

[8]As the contextual information used concerns first-order cooccurrences, the method is rather sensible to data sparseness. The use of more abstract information concerning higher-order occurrences constitutes an avenue for future work.

relation. These clusters can be merged on the basis of their overlaps into bigger ones, describing coarser-grained senses.
In the case of *movement*, the merging of the overlapping clusters ((a), (b) and (c)) would generate a bigger cluster describing the "physical movement" sense. The distinction between this cluster and the cluster (d), which describes the "social movement" sense, reflects the main semantic distinction characterizing the SL word.

1. *movement* - {{μετακίνηση, κίνηση, διακίνηση}, {κίνηση, διακίνηση, κυκλοφορία}, {μετακίνηση, διακίνηση, κινητικότητα}}
2. *movement* - {κίνημα}

In the case of *plant*, the fusion of the overlapping clusters ((a), (b) and (c)) generates a bigger one describing the "industrial" sense of the word, which is contrasted to the "botanical" sense, described by (d). These two clusters reflect the homonymic distinction characterizing *plant*.

1. *plant* - {{μονάδα, σταθμός}, {μονάδα, εγκατάσταση}, {σταθμός, εργοστάσιο}}
2. *plant* - {φυτό}

4.5. Discussion

We have shown how the inter-sense relations captured by the clusters overlaps can serve to modify the granularity of the obtained senses. Instead of going "bottom-up" (i.e. grouping fine-grained senses to form coarser ones), it would also be possible to adopt a "top-down" approach (i.e. to move from coarse-grained senses, described by disjoint clusters, to fine-grained ones). As there is no unique answer to the question of the granularity of the word sense descriptions needed in different NLP tasks, the possibility to modify it automatically permits to adapt the descriptions to the needs of specific applications. Taking into account the proximity of the senses facilitates the task of WSD algorithms as well, which do not have to make a selection among a large number of close but distinct, and hardly distinguishable, senses.
Additional benefits become evident during WSD evaluation. When the relations between senses are not taken into account, errors concerning close or distant senses are considered as equally important. Considering the inter-sense relations makes possible a differing penalization of WSD errors (Resnik and Yarowsky, 1998) and renders the evaluation more flexible and sophisticated. The advantages of exploiting this type of semantic representation in multilingual WSD, and in WSD and Machine Translation evaluation are presented in Apidianaki (2009) and Apidianaki et al. (2009).

5. Evaluation

5.1. Difficulties

The evaluation of automatic sense induction methods is difficult due to the lack of a gold standard in lexical semantics and the great divergences in the content and the coverage of existing hand-crafted semantic resources. Another important issue is that WSD constitutes an intermediate task in NLP applications, which aims to improve their performance (Wilks and Stevenson, 1996). Consequently, different applications have varying WSD needs which have an impact on the semantic descriptions that should be used. So, the results of an evaluation of the contents of a semantic inventory would not always be meaningful for the usefulness of the inventory in different settings.
An extrinsic evaluation of the results of the sense induction method presented in this paper has been performed in Apidianaki (2009). In this work, it is shown how the exploitation of the automatically built sense-cluster inventory can improve the performance of a WSD method in a bilingual context. Furthermore, the work of Apidianaki et al. (2009) shows how the use of this type of inventory can be beneficial in MT evaluation. Here, we present a more focused qualitative evaluation of the clustering results, by comparing them to the contents of an existing multilingual semantic resource.

5.2. Qualitative evaluation

We compare the senses obtained by the clustering method to the ones found for the same ambiguous English words in BalkaNet (Stamou et al., 2002; Tufis et al., 2004). BalkaNet is a multilingual semantic network that comprises wordnets in six different languages (Greek, Bulgarian, Romanian, Turkish, Serbian and Czech). Each wordnet contains concepts organized in semantic taxonomies, which are put into correspondence with their semantic equivalents in the other languages via an Interlingual Index (ILI) composed by concepts of Princeton WordNet. The ILI connects the languages between them and makes possible the transition from the concepts of one language to semantically similar concepts in another.
The senses found in the BalkaNet ILI for *movement* and *plant* are described in Table 2. We present the ILI synsets describing each sense of the words, the corresponding Greek synsets and the provided sense descriptions (glosses), which are common for the synsets of the two languages.

– movement

The first three senses given for *movement* in BalkaNet are too fine-grained and hardly distinguishable. The words found in the ILI synsets are very similar and the corresponding glosses as well. The Greek synsets corresponding to the ILI synsets do not clarify the proposed semantic distinctions either.
The Greek synsets corresponding to the first two senses contain the same equivalent, meaning "change of position".[9] So, no distinction can be performed at this level. TheTE found in synset (3), κίνησηι, is also contained in sense-clusters (a) and (b) that describe the "physical movement" sense. Nevertheless, the information provided in the sense-clusters of *movement* is richer than the information found BalkaNet. While in BalkaNet κίνησηι is isolated in a synset, in the sense-clusters it is linked to other TEs

[9] The proposed equivalent is not found in the clustering results because only one-word TEs are retained.

ambiguous word	ILI synsets	Sense description	Greek senses
movement	1. {motion, move, movement}	the act of changing location from one place to another	αλλαγή θέσης (alagi thesis)
	2. {mobility, motion, move, movement}	a change of position that does not entail a change of location	αλλαγή θέσης (alagi thesis)
	3. {motion, movement}	a natural event that involves a change in the position or location of something	κίνηση (kinisi)
	4. {front, movement, social movement}	a group of people with a common ideology who try together to achieve certain general goals	κίνημα (kinima)
plant	1. {flora, plant, plant life}	a living organism lacking the power of locomotion	φυτό (fyto)
	2. {industrial plant, plant, works}	buildings for carrying on industrial labor	βιομηχανικές εγκαταστάσεις (viomihanikes egkatastaseis)
	3. {plant}	something planted secretly for discovery by another	παγίδα (pagida)
	4. {plant}	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience	ηθοποιός (ithopoios)

Table 2: Balkanet senses for "movement" and "plant"

(μετακίνηση, διακίνηση, κυκλοφορία) that are semantically related to it. If the clusters are merged, κίνηση is also linked to κινητικότητα, another semantically close TE.
So, we observe that the
The senses described by the BalkaNet synsets (1), (2) and (3) could be grouped in one, describing the coarser sense of "physical movement". The clustering of the synsets could be done on the basis of their relations, which could be identified by the methods described in section 2.2. The semantic relations of these synsets could be discovered on the basis of the similarity of the lexical information found in the synsets (*motion, movement, move*) and the glosses (*change, location, position*). The relation between the synsets (1) and (2) could also be captured by exploiting the relations that they share in the taxonomy (the same hyperonym: *change*, and the same derivational relation: *move*). However, the taxonomic information would not be sufficient to identify the relations of these synsets to synset (3), as this has a different hyperonym (*happening, natural event, occurrence*). Actually, the relations between the senses are not always reflected in their hyperonyms in WordNet, which often emphasize the differences in what is being highlighted by each sense, rather than their similarities (Palmer et al., 2006).
The synset (4) describes the abstract sense of "social movement". We observe that this sense, which is more distant from the others, is situated at the same level as the three related senses of *movement*. Like WordNet, BalkaNet is based on the enumeration approach, which does not operate distinctions between close and distant senses and different types of ambiguity. So, the distinct senses of an entry are sequentially listed and not hierarchically organized. The distinction of the senses by reference to their status could however be performed by exploiting taxonomic information (like the hyperonym of synset (4), which is the synset "social group"). However, even if it is possible to discover the relations and distinctions of the senses on the basis of the information available in the network, these are not explicitly described in the inventory. The senses are organized by frequency of use and there is no indication of the degree of distinguishability between them. These lacks result in the uniform processing of the proposed senses during WSD and have a negative impact during WSD evaluation, as errors concerning close and distant senses are equally penalized (Resnik and Yarowsky, 1997; Resnik and Yarowsky, 1998).
The relations and distinctions between the senses are explicitly described in the clustering solution obtained for *movement*. The coarse-grained senses acquired after merging the overlapping clusters correspond to the senses that would be obtained if the three fine-grained synsets of the words were grouped. The merging of the clusters could thus be considered as performing the same function as the knowledge-based methods that have been proposed for reducing the granularity of WordNet senses.

– plant

The senses described by the first two synsets of *plant* in BalkaNet correspond to its "botanical" and the "industrial" sense. This distinction between the homonymic senses of the word is also performed in the sense clustering solution. The other two senses provided in BalkaNet are very rare. Instances of the *plant* carrying these senses do not appear in our training corpus, so information about these two senses is not found in the sense-clusters. The suggestion of too rare senses or of senses not relevant to the domains of the processed texts constitutes a drawback of exploiting predefined resources for WSD (Pantel and Lin, 2002). The consideration of such senses increases the number of possible choices during WSD which makes processing more complex without any benefit.
The synsets describing the rare senses of *plant* (3 and 4) contain only this word (no synonyms) and have no rela-

tion (hyponymic, meronymic, etc.) with other synsets of the network. On the contrary, the first two synsets that describe its homonymic senses are linked to other synsets. The hyperonym of the first synset is the synset {being, organism}; its hyponyms are {tracheophyte, vascular plant} and {fungus}, and it also retains a meronymy relation with {plant part, plant structure} and a rev_category_domain relation with {microorganism}. The hyperonym of the second synset is {building complex, complex} and its hyponym {factory, manufactory, manufacturing plant, mill}.
The information contained in the Greek synsets corresponding to the ILI synsets of *plant* is limited, as each synset contains only one TE: the synset describing the "botanical" sense of *plant* contains the TE φυτό and the one describing its "industrial" sense contains the multiword term βιομηχανικές εγκαταστάσεις (industrial installation). The sense-clusters describing these senses contain richer information than the synsets. The cluster of the TE εγκατάσταση , also contains the semantically similar word μονάδα.If the overlapping clusters are merged, more similar words are included: the TEs σταθμός and εργοστάσιο.The TE εργοστάσιο is found in BalkaNet in the hyponym of the synset βιομηχανικές εγκαταστάσεις (its hyperonym is the synset 'κτηριακό συγκρότημα' (building block).
Concerning the two rare senses, it would be difficult to consider the TEs found in the corresponding Greek synsets (παγίδα (trap) and ηθοποιός (actor)) as translations of *plant* in Greek. Additionally, the information provided by these synsets is poor as each synset contains only one word while no relations to other synsets are described.
The divergences concerning the quantity of available information are justified by the different status of the concerned senses. However, the difference between the two main homonymic senses and the very rare senses of the word is not described in the resource. This is due to the difficulty of sense enumeration lexicons to make a difference between the senses regarding their status. The senses described in these resources are all situated at the same level and treated as uniform.
The differences in the status of the senses are described in the sense-clusters obtained for *plant*: the mutually exclusive senses are described by distinct clusters while the related senses and sub-senses are described by overlapping clusters.

6. Conclusion

In this paper, we have shown how word sense relations can be identified from textual data by means of an unsupervised sense induction method. The method identifies the semantic relations between the translation equivalents of ambiguous words in a parallel corpus and groups them into clusters, which describe the senses of the source language words. Using the clusters as sense indicators, instead of the separate TEs, increases the pertinence of the identified senses in comparison to the cross-lingual approach to sense induction where TEs serve to identify distinct senses. Additionally, by creating overlapping clusters the method discovers the relations between the corresponding senses.
In the generated semantic inventory, the senses of the words in one language are described by the clusters of their equivalents in another. A merit of this approach is that the acquired senses are not simply enumerated but their relations are taken into account as well. This important difference of the obtained semantic representation from that found in traditional resources is illustrated by the comparison of the sense-clustering results to the corresponding descriptions in an existing multilingual resource, BalkaNet. The description of the relations between the word-senses permits to automatically modify their granularity and to adapt them to the WSD needs of specific applications. Further advantages of this unsupervised semantic analysis method are its language-independency and the relevance of the acquired descriptions to the processed data.

7. References

M. Apidianaki, Y. He, and A. Way. 2009. Capturing lexical variation in MT evaluation using automatically built sense-cluster inventories. In *23rd Pacific Asia Conference on Language, Information and Computation (PACLIC'09)*, Hong Kong.

M. Apidianaki. 2008. Translation-oriented sense induction based on parallel corpora. In *Language Resources and Evaluation Conference*, pages 3269–3275, Marrakech, Morocco.

M. Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 77–85, Athens, Greece.

M. Baker. 1996. Corpus-based translation studies: the challenges that lie ahead. In H. Somers, editor, *Terminology, LSP and Translation: studies in language engineering in honour of Juan C. Sager.*, pages 175–186. John Benjamins Publishing Company, Amsterdam/Philadelphia.

C. Cabezas and P. Resnik. 2005. Using WSD Techniques for Lexical Selection in Statistical Machine Translation. Technical report. Technical report CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42.

M. Carpuat, Y. Shen, Y. Xiaofeng, and D. Wu. 2006. Towards Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 37–44, Kyoto, Japan.

T. Chklovski, R. Mihalcea, T. Pedersen, and A. Purandare. 2004. The senseval-3 multilingual English-Hindi lexical sample task. *Proceedings of Senseval-3, Third International Workshop on Evaluating Word Sense Disambiguation Systems.*, pages 5–8.

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

I. Dagan, S. Marcus, and S. Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 164–171, Columbus, Ohio.

W. B. Dolan. 1994. Word Sense Ambiguation: Clustering Related Senses. *Proceedings of the 15th Interna-*

tional Conference on Computational Linguistics (COLING), pages 712–716.

P. Edmonds and A. Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(4):279–291.

W. A. Gale, K. W. Church, and D. Yarowsky. 1993. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, 26:415–439.

M. Gavrilidou, P. Labropoulou, E. Desipri, V. Giouli, V. Antonopoulos, and S. Piperidis. 2004. Building parallel corpora for eContent professionals. In *Proceedings of MLR 2004, COLING Workshop on Multilingual Linguistic Resources*, Geneva, Switzerland.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*.

N. Ide and Y. Wilks. 2007. Making Sense About Sense. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

N. Ide, T. Erjavec, and D. Tufis. 2002. Sense discrimination with parallel corpora. In *ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia.

J. J. Jiang and D. W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING)*, pages 19–33, Taipei, Taiwan.

P. Jin, Y. Wu, and S. Yu. 2007. SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).*, pages 19–23.

R. Mihalcea and D. I. Moldovan. 2001. Automatic generation of a coarse grained wordNet. *Proceedings of the 14th FLAIRS Conference*, pages 454–458.

G. A. Miller and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

R. Navigli, K. Litkowski, and O. Hargraves. 2007. SemEval 2007 Task 07 : Coarse-Grained English All-Words Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic.

R. Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING/ACL'06)*, pages 105–112, Sydney, Australia.

H. T. Ng, Y. L. Chung, and K. F. Shou. 1999. A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99)*, pages 9–13, College Park, Maryland.

M. Palmer, C. Fellbaum, and H. T. Dang. 2006. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 12(3):137–163.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Alberta, Canada.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional Clustering of English Words. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, Ohio.

W. Peters, I. Peters, and P. Vossen. 1998. Automatic sense clustering in euroWordNet. *Proceedings of the 1st International Conference on Language Resources and Evaluation (LRECU98)*, pages 409–416.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

P. Resnik and D. Yarowsky. 1997. A perspective on Word Sense Disambiguation Methods and Their Evaluation. In *Proceedings of SIGLEX Workshop Tagging Text with Lexical Semantics: What, why and how?*, pages 79–86, Washington, D.C.

P. Resnik and D. Yarowsky. 1998. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(3):113–133.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint conference for Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, Canada.

L. Specia, G. C. B. Ribeiro, M. d. G. V. Nunes, and M. Stevenson. 2006. The Need for Application-Dependent WSD Strategies: a Case Study in MT. In *7th Workshop on Computational Processing of Written and Spoken Portuguese (Propor'06), LNAI 3960*, pages 233–237, Itatiaia.

S. Stamou, K. Oflazer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, and M. Grigoriadou. 2002. BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *International Wordnet Conference*, pages 12–14, Mysore, India.

D. Tufis, R. Ion, and N. Ide. 2004. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *20th International Conference on Computational Linguistics (COLING'04)*, pages 1312–1318, Geneva, Switzerland.

L. van der Plas and J. Tiedemann. 2006. Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL/COLING)*, pages 866–873, Sydney, Australia.

J. Veronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advances papers of the Senseval workshop*, Herstmonceux Castle, UK.

P. Vossen. 1999. EuroWordNet General Document. EuroWordNet: (le2-4003, le4-8328). Technical Report Part A, Final Document.

Y. Wilks and M. Stevenson. 1996. The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report Research Memoranda, CS-

96-05, University of Sheffield, Department of Computer Science.

Natural and contextual constraints for domain-specific relations

Pilar León Araúz and Pamela Faber
University of Granada
Buensuceso, 11 18002 Granada, Spain
E-mail: pleon@ugr.es, pfaber@ugr.es

Abstract

EcoLexicon is a specialized knowledge base on the environment. It is linked to an ontology in order to apply reasoning techniques and enhance user queries. Until recently, semantic relations in specialized term bases, if they existed at all, were mainly restricted to generic-specific and part-whole relations. The extensive use of ontologies for knowledge representation has contributed to the development of certain formal criteria for conceptual description. However, the problem now is the lack of consensus on what kind and how many semantic relations should be covered. It is our claim that domain-specific relations show certain natural constraints that could guide conceptual description. On the other hand, concepts are context-sensitive and they do not always show the same relational behaviour. Accordingly, conceptual networks in EcoLexicon account for contextual reconceptualization.

1. Introduction

EcoLexicon [1] is a specialized knowledge base on the environment. So far, it has 3,042 concepts and 10,597 terms in English, Spanish and German. It is primarily hosted in a relational database which is now linked to an ontology in order to apply reasoning techniques and enhance user queries. As is well-known, ontologies are a powerful mechanism that help to prevent inconsistencies, and make terminology management a more empirical and coherent process. For this purpose, we have developed a systematic way of describing environmental concepts, and the semantic relations that link them. This is a way of mapping possible conceptual configurations in terms of multidimensionality and dynamism. In the following sections we offer a brief overview of the needs for an in-depth analysis of semantic relations in Terminology as well as the application of natural and contextual constraints in EcoLexicon.

2. Terminology and semantic relations

Until recently, semantic relations in specialized term bases, if they existed at all, were mainly restricted to generic-specific and part-whole relations. This was conducive to static configurations, which were at odds with the need to represent dynamic action in domain models (Barrière, 2001: 137). Hyponymy has been widely studied not only because it underlies categorization, but also because it guides property inheritance (Barrière, 2004: 244). However, according to Dancette and Halimi (2005: 202), any terminological resource which is not also enriched by other types of relations will fail to achieve its goals.

In general terms, the relationships between concepts have received relatively little attention compared to the attention devoted to concepts and concept classes (Green et al., 2002: vii). Nevertheless, over the past decades, semantic relations have been at the core of much work in different disciplines, such as philosophy, linguistics and artificial intelligence. Moreover, the development of corpus linguistics and the extraction of lexico-syntactic patterns (Condamines, 2002; Barrière, 2004; Marshman et al., 2002), have improved the identification of all possible ways in which specialized concepts may relate to others.

The extensive use of ontologies for knowledge representation has contributed to the development of certain formal criteria for conceptual description. However, the problem now is the lack of consensus on what kind and how many non-hierarchical relations should be covered, since they are always created with very different purposes in mind, depending on the domain being represented (Hovy, 2002: 92).

3. Domain-specific semantic relations in EcoLexicon

In our experience, semantic relations largely depend on the type of concept being described, its nature, and relational power. In EcoLexicon, some relations are domain-specific, and reflect dynamism and change as the result of a process-oriented management (Faber et al., 2006). We have developed an inventory of semantic relations based on three general concept types: entities, events and properties. For instance, Figure 1 shows a network with processes (INFILTRATION), entities (ALLUVIAL FAN) and properties (ALLUVIAL).

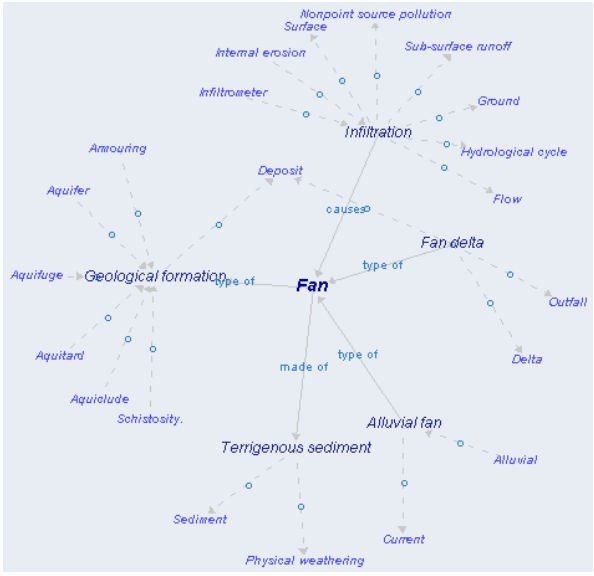


Figure 1. Conceptual network of FAN

[1] http://manila.ugr.es/visual

A great effort has been made to build a systematic method for conceptual description. It is based on the following criteria (Faber et al., 2009):

- *Is_a*: the traditional generic-specific relation reflects hierarchical inheritance in conceptual networks. All entities and events are categorized as subtypes of a particular class. For example, SHEET PILE GROYNE (instance) *is_a* GROYNE (class, subtype) *is_a* COASTAL DEFENSE STRUCTURE (class, subtype) *is_a* COASTAL STRUCTURE (class, subtype).
- *Part_of*: this relation also reflects the hierarchical structure of the domain. In the case of physical entities, this relation directly refers to the parts of each concept (SPILLWAY *part_of* DAM). However, there is another case of abstract meronymy for mental entities, such as scientific disciplines (MICROBIOLOGY *part_of* BIOLOGY).
- *Phase_of*: this is another kind of meronymy applied to processes. In the same way that objects are incomplete and can even lose their identity without one or more of their constituent parts, processes are incomplete without one or more of their phases (PUMPING *phase_of* DREDGING).
- *Made_of*: this relation links both artificial and natural objects to the material they are made of, and thus bears a certain resemblance to the *part_of* relation without being the same. Even though the material of an object is part of it, this relation differs from the latter because materials used to build an artefact can vary. For example, a GROYNE HEAD is *part_of* all GROYNES, but the same cannot be said of the material used to make this type of construction, since GROYNES can be *made_of* STONE, CONCRETE or WOOD.
- *Delimited_by*: this relation pertains to physical objects, and marks the boundaries, dividing one object from another. This is a domain-specific relation, mainly for geographic entities, such as the different layers of the atmosphere or the Earth. For example, the STRATOSPHERE and MESOSPHERE are *delimited_by* STRATOPAUSE.
- *Located_in*: this relation is relevant when the location of a physical object is an essential characteristic for its description. For instance, a GROYNE is not a GROYNE if it is not *located* on the COAST. When the *located_in* relation converges with the *part_of* relation, the *part_of* then overrides *located_in*. For example, a RIVER BED is *part_of* a RIVER instead of *located_in* the RIVER, because a RIVER cannot exist without its BED.
- *Takes_place_in*: this relation describes processes which have spatial and temporal dimensions. The distinction between this relation and *located_in* is based on the fact that processes are not bounded in space as objects, and also have a temporal dimension. For example, LITTORAL DRIFT *takes_place_in* the SEA; and THERMAL LOW *takes_place_in* SUMMER.

The last six semantic relations are all subtypes of meronymy. This distinction is in consonance with some of the subtypes proposed in Winston et al. (1997), since not all parts interact in the same way with their wholes. The reason why we have established six different meronymic relations is based on our domain-specific needs, but especially on ontological reasoning and consistency. For example, if *located_in* were considered as a *part_of* relation, that would cause a fallacious transitivity (Murphy, 2003). If a GABION is *part_of* a GROYNE and a GROYNE *part_of* the SEA, the ontology would infer that GABIONS are *part_of* the SEA, which is not a plausible example. In the same way, if both processes and entities were connected through the same *part_of* relation, there would be no restrictions on category membership or disjunction. However, it is true that if a HARD DEFENCE STRUCTURE is *located_at* the BEACH and the BEACH is *part_of* the COAST, then the DEFENCE STRUCTURE is *located_at* the COAST. In this sense, we are now working on "property chain inclusions" according to the W3C recommendations.

- *Result_of*: this relation is relevant to either events or entities that are derived from other events. Even though events and entities can be the result of another event, an event cannot be the result of an object. For example, ACCRETION is the *result_of* SEDIMENTATION (process), but it cannot be regarded as the *result_of* SEDIMENTS (entity).
- *Causes*: this relation only links entities and events, for example, WATER *causes* EROSION. Even though this relation initially seems to be the inverse of *result_of*, there is a difference stemming from the active role played by certain entities. *Causes* only describes the beginning of a process, whereas *result_of* may link events or entities that are the consequence of another event. When an entity causes a final result in another entity, the following relation applies.
- *Affects*: this relation, along with *causes* and *result_of*, is a crucial relation in dynamic systems such as ours since environmental concepts have a high combinatorial potential. *Affects* relates a wide variety of concepts to their ever-changing environments. It links processes or entities that cause a change in any other object or event without producing a final result (e.g. GROYNE *affects* LITTORAL DRIFT). Moreover, complex conceptual relations such as *affects* can generate a hierarchy of domain-specific relations such as *retards* (BEACH NOURISHMENT *retards* BEACH EROSION), *erodes* (WATER *erodes* ROCKS), etc.
- *Has_function*: this relation not only links entities or processes that are artificially created or carried out with a specific function, but also natural entities which, despite not being goal-directed, can be used for human profit. Natural concepts with a function are AQUIFER (*has_function* WATER SUPPLY), SAND (*has_function* BEACH NOURISHMENT), etc. As in the case of *affects*, *has_function* can also be associated with other domain-specific subordinate relations, such as *measures* for instruments (a PLUVIOMETER *measures* PRECIPITATION); *studies* for sciences (POTAMOLOGY *studies* SURFACE CURRENTS); and *represents* for graphics, maps and charts (a HYDROGRAPH *represents* RATE OF WATER FLOW).
- *Effected by*: this relation is only used for instruments that carry out some process or create an entity. For example, DREDGING is *effected_by* a DREDGER and a MARIGRAM is *effected_by* a TIDE GAUGE. This relation is especially meaningful in those domains where human interaction plays an essential role as it is the case of environmental contexts.

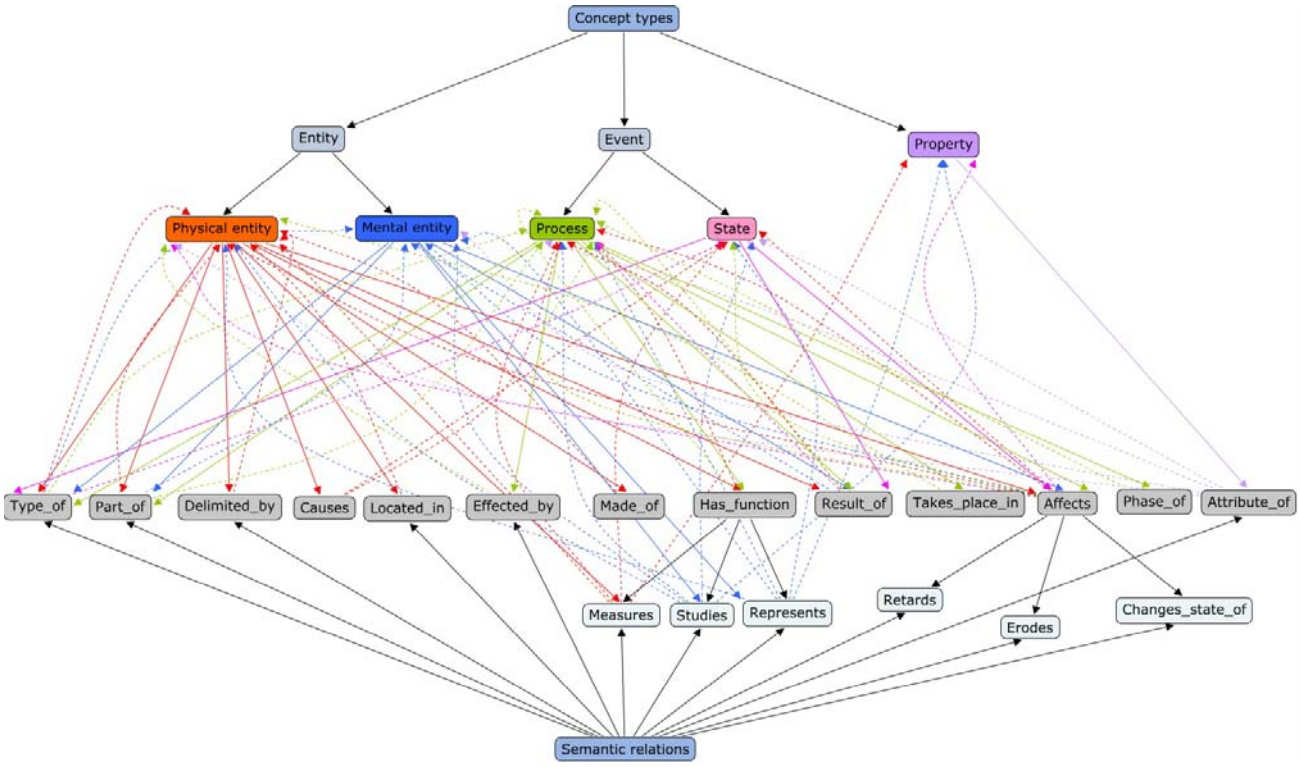


Figure 2. Combinatorial potential and relation types

Obviously, each of the above relations has its inverse relation, except *delimited_by* due to its symmetric nature: *is_a* ↔ *generic_of; result_of* ↔ *has_result; causes* ↔ *caused_by; part_of* ↔ *has part*, etc.

3.1 Natural constraints

According to the above-mentioned criteria, concept nature alone determines the potential activation of certain semantic relations, but at the same time, semantic relations determine which kind of concepts can be part of the same conceptual proposition. This gives rise to all these possible combinations (Figure 2).

This combinatorial potential represents certain constraints associated with the natural aspect of concepts. For instance, a process may activate the relation *effected_by*, but only if it is associated with a physical entity. However, if it activates *affects*, it can be linked to entities, events and properties.

However, the environmental domain has many concepts that can be represented according to very different facets, causing the well-known phenomenon of multidimensionality (Kageura, 1997; Rogers, 2004). This increases the number of possible relations activated by particular concepts, since multidimensionality is intimately linked to the semantic roles concepts may play. In a process-oriented domain the same concept may act as an AGENT or a PATIENT, as a PROCESS or a RESULT[2].

For example, the concept BEACH can be either a PATIENT (e.g. of WAVE ACTION) or an AGENT (e.g. of PROGRADATION) and the prototypical activation of semantic relations generally depends on perspective. Broadly speaking, we could say that BEACH (as a physical entity) may activate nearly all semantic relations, except those exclusive of events and properties: *is_a, generic_of, located_in, location_of, delimited_by, part_of, has_part, made_of, material_of, has_function, effected_by, effects, measures, measured_by, affects, affected_by* and *causes.* Inverse relations are included because they do not have the same behaviour in terms of prototypicality. In Table 1, all these relations are restricted according to semantic roles.

Interestingly enough, hierarchical relations are invariable parameters. Physical entities may *have parts* or be *part of* other wholes whether they are AGENTS or PATIENTS, but that is not the case for non-hierarchical relations. This means that a concept's behaviour depends on its nature and the role that it plays in a particular domain. If a physical entity behaves like a PATIENT it cannot *effect* anything, as it would then become an AGENT. Prototypically, a PATIENT can only activate its inverse relation, *effected_by*. The same applies for relations such as *measures* and *affects*. This does not imply that the same concept cannot play different roles (e.g. BEACH), but rather that the representation of the concept is role-sensitive, and should change accordingly in each case.

Furthermore, there are three relations that are exclusive of AGENTS or PATIENTS (*causes* and *location_of*), and which do not have an inverse. According to our relational criteria, *causes* is only used for entities causing processes, so no PATIENTS (always entities) may be *caused by* other entities. On the other hand, locations are considered to be PATIENTS by default, since they do not imply being active agents.

[2] Our inventory of basic semantic roles has been designed according to the most prototypical roles encountered in the corpus, following the FrameNet approach.

Concept Type	AGENT		PATIENT	
	Relational power			
	Relation	Concept Type	Relation	Concept Type
Physical entity	*Is_a* *Generic_of* *Delimited_by* *Has_part* *Part_of* *Made_of* *Material_of* *Located_in* *Has_function* *Effects* *Measures* *Affects* *Causes*	Physical entity Event Event/Entity Event/Entity Event/Entity Event	*Is_a* *Generic_of* *Delimited_by* *Has_part* *Part_of* *Made_of* *Material_of* *Located_in* *Location_of* *Effected_by* *Measured_by* *Affected_by*	Physical entity Entity/Event

Table 1. Relational constraints according to concept nature and semantic role

3.2 Contextual constrains

Multidimensionality is often regarded as a way of enriching traditional static representations, enhancing knowledge acquisition through different points of view in the same conceptual system. However, it causes a heavy information overload, which ends up jeopardizing knowledge acquisition.

This is the case of certain versatile concepts involved in a myriad of events, such as WATER (Figure 3). Obviously, users would not acquire any meaningful knowledge if all dimensions of WATER were shown in the same network. In Figure 3, WATER is closely linked to many natural and artificial processes, such as EROSION or DESALINATION. However, WATER rarely (if ever) activates both concepts simultaneously since they evoke widely different contexts or situations. In the first context (EROSION), WATER is a natural *agent,* whereas in the second context (DESALINATION), WATER is the *patient* of an artificial process (León Araúz et al., 2009).

The problem of information overload results from the *incompatibility of facets* in such a wide domain. Cruse (1995:4) illustrates this with the example of BOOK and its subtypes. NOVEL, BIOGRAPHY, and TEXTBOOK are hyponyms of BOOK, but so are PAPERBACK and HARDBACK. The hyponyms of BOOK fall into two sets, which display within-set incompatibility and between-set compatibility. For example, if a book is a NOVEL, it cannot be a TEXTBOOK. However, if it is a NOVEL, it can be a PAPERBACK. In that sense, even though WATER subtypes, such as PRECIPITABLE WATER, DRINKING WATER and NAVIGABLE WATER, all represent the same facet *function*, strictly speaking, they are not coordinate concepts.

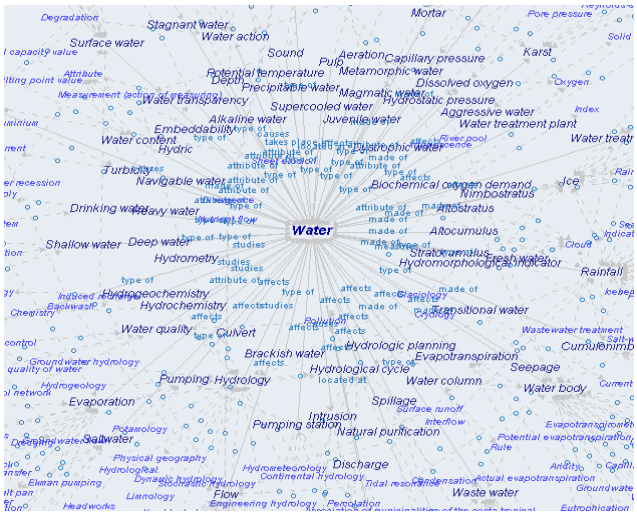


Figure 3. Information overload in the conceptual network of WATER

In EcoLexicon, the incompatibility of facets is applied to all types of semantic relations in a context-dependent way. According to Barsalou (2009: 1283), a concept produces a wide variety of situated conceptualizations in specific contexts. Furthermore, Yeh and Barsalou (2006) state that when situations are not ignored, but incorporated into a cognitive task, processing becomes more tractable. In the same way, any specialized domain contains contextual domains in which certain conceptual dimensions become more or less salient. As a result, a more believable representational system should account for reconceptualization according to the situated nature of concepts.

water supplies in the area. [1] In areas where some water is available for irrigation, shrubs planted on
Vshaped valley has been limited and linear erosion by water has been a major geomorphic process in the alluv
the saturated zone, allowing the plants to transpire water directly from the ground-water system (Figure 7).
in the ocean separated by 50 to 100 km. The flow of water that results from these ocean slopes are called
75 minutes. That's 300 mph (480 kph). These walls of water are capable of inflicting massive damage along
er the land surface, and is a major component of the water cycle.[1][2] Runoff that occurs on surfaces bef
re sampled using a suction sampling apparatus pumping water into 2 l acrylic jars equipped with 2 layers of
ch matches proposed system. (1) Type 5A7 wells inject water used to provide heating or cooling for structures
bly in that time). Every day, a very small amount of water is lost high in the atmosphere, where intense u
times by carefully adding and drawing off distilled water with a pipette, and a drop of glycerine was then
eezing point of water in a similar way.[16] Solutes in water also affect water activity which affects many c
ssures and an accompanying increased demand for fresh water for use by a growing population. One of the earl

Figure 4. Concordances of WATER: contextual domains and linguistic markers

This is done by dividing the environmental field into different contextual domains according to corpus information: HYDROLOGY, GEOLOGY, METEOROLOGY, BIOLOGY, CHEMISTRY, ENGINEERING, WATER TREATMENT/SUPPLY, COASTAL PROCESSES and NAVIGATION.

Figure 4 illustrates how corpus data help to identify different contextual domains through linguistic markers. The activation of concepts, such as IRRIGATION, EROSION and PUMPING frames WATER in diverse settings (natural or artificial, patient or agent, etc), whereas linguistic markers (e.g. *available for*, *used to* or *component of*) help to categorize the concept within each contextual domain through specific conceptual relations (function, meronymy, etc.).

For example, when WATER appears as a functional concept (expressed by markers like *available for*, *used to/for*), it is usually related to the AGRICULTURE or WATER TREATMENT/SUPPLY contextual domains. Consequently, the collocations of WATER disambiguate domain membership, since terms, such as *irrigation* or *plant,* only apply in the AGRICULTURE domain. In contrast, *well* and *population* only apply in the WATER TREATMENT/SUPPLY context.

Thus, domain membership reconceptualizes versatile concepts restricting their relational behaviour. Contextual constraints are neither applied to individual concepts nor to individual relations, since one concept can be activated in different contexts or use the same relations, but with different values. Constraints are instead applied to conceptual propositions (León Araúz and Magaña Redondo, in press). For instance, CONCRETE is linked to WATER through a *part_of* relation. Nevertheless, this proposition is irrelevant if users only want to know how WATER naturally interacts with the landscape or how it is purified of contaminants. Consequently, the proposition WATER *part_of* CONCRETE will only appear in an ENGINEERING context. As a result, when constraints are applied, WATER only shows relevant dimensions for each contextual domain. In Figure 5, thanks to the application of contextual restrictions, WATER is only linked to propositions belonging to ENGINEERING.

However, in Figure 6 the GEOLOGY context shows WATER in a new structure with other concepts and relations. Evidently, the number of conceptual relations varies from one network to another, since WATER is not equally relevant in all contextual domains. Furthermore, relation types also differ, which highlights the changing nature of WATER'S internal structure according to each semantic role. For

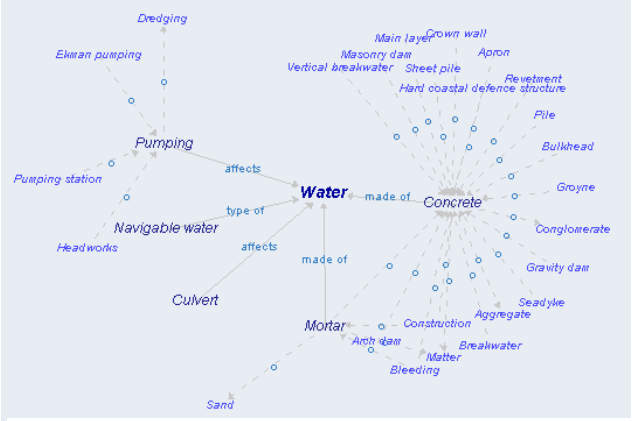


Figure 5. WATER in the ENGINEERING context domain

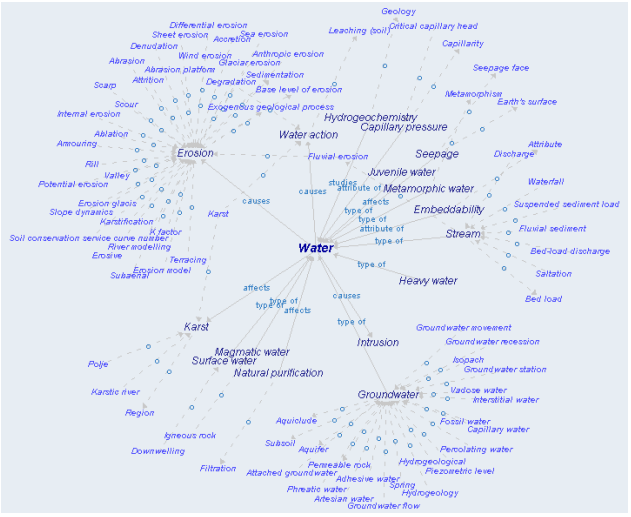


Figure 6. WATER in the GEOLOGY context domain

example, in the ENGINEERING context domain, most relations are *made_of* and *affects*, whereas in the GEOLOGY domain, *causes* and *type_of* stand out. *Affects* is also shared by the GEOLOGY domain, but the arrow direction shows a different perspective. More specifically, in geological contexts WATER is an active *agent,* whereas in ENGINEERING, the concept is more subject to changes (*patient*). Finally, WATER is not always related to the same concept types. In ENGINEERING, WATER is only linked to *artificial* entities or processes (PUMPING, CONCRETE, CULVERT), while in GEOLOGY it is primarily related to *natural* ones (EROSION, GROUNDWATER, SEEPAGE).

Reconceptualization does not involve a clear-cut distinction between different context domains, since they can also share certain conceptual propositions. This is due to the fact that multidisciplinarity gives rise to fuzzy category boundaries and, as a result, contextual domains can form their own hierarchical structure. Moreover, they are also dynamic and flexible structures that should evolve over time according to the type and amount of information stored in our knowledge base (León Araúz and Magaña Redondo).

4. Conclusion

Terminological knowledge bases need certain formal and coherent criteria for conceptual description. In EcoLexicon natural constraints are determined by conceptual nature and the semantic role played by each concept in the environmental domain. This makes the system a consistent resource in its different representational levels. Contextual constraints enrich the system from both a qualitative and quantitative standpoint. On the one hand, they structure knowledge in a similar way to how things relate in the real world, as well as in the human conceptual system. On the other hand, conceptual dimensions are noticeably reduced with a coherent method based on a cognitive approach. As a result, the situated representation of versatile concepts is a viable solution for managing information overload and at the same time enhancing knowledge acquisition processes.

5. Acknowledgements

This research has been partially supported by project FFI2008-06080-C03-01/FILO, from the Spanish Ministry of Science and Innovation and project P06-HUM-01489, from the Andalusian Regional Government.

6. References

Barsalou, L.W. 2009. "Simulation, situated conceptualization and prediction". *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 364: 1281-1289.

Barrière, C. 2004. "Building a concept hierarchy from corpus analysis". *Terminology* 10: 2, 241-263.

Barrière, C. 2001. "Investigating the causal relation". *Terminology* 7: 2, 135-154.

Condamines, A. 2002. "Corpus analysis and conceptual relation patterns". *Terminology* 8: 1, 141-162.

Cruse, D.A. 1995. "Polysemy and related phenomena from a cognitive linguistic viewpoint". In St. Dizier, P. y Viegas, E. (eds.), *Computational Lexical Semantics*. Cambridge: Cambridge University Press. 33-39.

Dancette, J. and Halimi, S. 2005. "La représentation des connaissances; son apport à l'étude du processus de traduction", *Meta* 2, 548-559.

Faber, P., León Araúz, P. and Prieto Velasco, J.A. 2009. "Semantic relations, dynamicity and terminological knowledge bases". *Current Issues in Language Studies*. 1: 1. 1-23.

Faber, P., Montero Martínez, S., Castro Prieto, M.C., Senso Ruiz, J., Prieto Velasco, J.A., León Araúz, P., Márquez Linares, C.F. and Vega Expósito, M. 2006 "Process-oriented terminology management in the domain of Coastal Engineering". *Terminology* 12: 2, 189-213.

Green, R., Bean, C.A. y Myaeng, S.H. 2002. *The semantics of relationships*. Dordrecht: Kluwer.

Hovy, E.H. 2002. "Comparing sets of semantic relations in ontologies". In Green, R., Bean, C.A. y Myaeng, S.H. (eds.), *The Semantics of Relationships*: *An Interdisciplinary Perspective*. Dordrecht: Kluwer. 91-110.

Kageura, K. 1997. "Multifaceted/Multidimensional concept systems". In Wright, S.E. y Budin, G. (eds.), *Handbook of Terminology Management*: *Basic Aspects of Terminology Management*. Amsterdam/Philadelphia: John Benjamins. 119-32.

León Araúz, P. and Magaña Redondo, P.J. (in press). "EcoLexicon: contextualizing an environmental ontology". *Terminology and Knowledge Engineering*, Dublin.

León Araúz, P., Magaña, P.J. and Faber, P. 2009. "Building the SISE: an environmental ontology". In *Proceedings of Towards e-Environment*. Prague.

Marshman, E., Morgan, T. and Meyer I. 2002. "French patterns for expressing concept relations." *Terminology* 8: 1, 1-29.

Murphy, M.L. 2003. *Semantic Relations and the Lexicon*. Cambridge: Cambridge University Press.

Rogers, M. 2004. "Multidimensionality in concepts systems: a bilingual textual perspective". *Terminology* 10: 2, 215-240.

Winston, M.E., Chaffin, R. and Herrmann, D. 1987. "A taxonomy of part-whole relations". *Cognitive Science* 11, 417-444.

Yeh, W. and Barsalou, L.W. 2006. "The situated nature of concepts". *American Journal of Psychology* 119, 349-384.

From the People's Synonym Dictionary to fuzzy synsets – first steps

Lars Borin and Markus Forsberg

Språkbanken, University of Gothenburg, Sweden
lars.borin@svenska.gu.se, markus.forsberg@gu.se

Abstract

We present our ongoing work on creating fuzzy synsets for Swedish using the lexical resources Synlex and SALDO. Synlex is a graded synonym list created by asking members of the public – users of an online Swedish-English dictionary – to judge the degree of synonymy of a random, automatically generated synonym pair candidate. SALDO is a full-scale Swedish lexical-semantic resource with non-classical, associative relations among word and multiword senses, identified by persistent formal identifiers. We discuss two approaches for mapping Synlex synonym pairs to SALDO senses – transitive closure and clique formation – as well as our planned work for including other kinds of classical lexical-semantic relations from various existing free lexical resources, into Swesaurus, a multi-faceted resource for Swedish combining classical wordnet-type relations with the associative thesaurus relations from SALDO.

1. Introduction

The Princeton WordNet (WN; Fellbaum 1998) and other wordnets created in its image are standard items in any modern language technology resource toolkit. Notwithstanding their widespread use and general popularity in language technology research and applications, some of the decisions that shaped WN are debatable at least from a lexicographical and linguistic point of view. Most often, the unclear theoretical status of the notion of synonymy is pointed out (e.g., Ci 2008; Piasecki et al. 2009).

Since the synonymy relation is the basis of the whole wordnet endeavor, defining as it does the central entity of Princeton-type wordnets, the *synset*, any flaw in this concept will call into question the foundations of the whole wordnet enterprise. Relevant in this connection, there is a postulated universal linguistic principle of (full) *synonymy avoidance* (Carstairs-McCarthy, 1999). This being an intrinsic characteristic of human language – so the reasoning goes – a dictionary whose fundamental organization is based on the notion of synonymy almost by definition cannot present a faithful reflection of our lexical knowledge, at least not from a linguistic point of view.

WN synonyms, as originally defined, should be interchangable in some contexts, but not necessarily in all contexts (Miller, 1998, 24); in fact, even one context is enough (Alonge et al., 1998, 22). This indicates that synonymy in the WN sense may not correspond exactly to how linguists and lexicographers understand this term, and further that it may be a matter of degree – for instance expressible as the number of possible substitution contexts of a particular synonym pair. To the best of our knowledge, this interesting notion has never been explored with respect to wordnets. But if this is the case, and if we had some practicable means of quantifying the degree of synonymy among words, then we could actually define a kind of wordnet based on this, where synsets could grow or shrink, depending on the degree of synonymy that we require for a particular purpose.

The work described below represents an attempt to accomplish exactly this. In Språkbanken, a language technology R&D unit at the University of Gothenburg, We have started work on a 'fuzzy wordnet' for Swedish, understood here as a wordnet based on 'fuzzy', or graded, synsets. This endeavor is made feasible by the previous existence of a number of freely available lexical resources on which we can draw in our work. The work is in its initial stages, so what we can offer in this paper are some preliminary results of automatic merging of two unique lexical resources, together with a discussion of a number of interesting theoretical issues that arise from this work.

The two lexical resources under discussion here are Synlex and SALDO.

2. Synlex

Graded synonymy relations for part of the Swedish vocabulary are available in *Synlex* (the People's Synonym Lexicon; Kann and Rosell 2006). This lexical resource has been created by asking members of the public – users of an online Swedish-English dictionary – to judge the degree of synonymy of a random, automatically generated synonym pair candidate, on a scale from 0 (not synonyms) to 5 (fully synonymous). A synonym pair list containing all pairs that average 3.0 or more on a large number of judgements is available for download under an open-source license. The latest version of the list at the time of writing is

dated 2009-05-29, and contains 18,607 graded synonym pairs (37,214 when symmetry of synonymy is taken into account).

The members of these pairs are words (i.e., text word forms) – not even part of speech (PoS) is indicated – mainly dictionary base forms (lemmas), but sometimes inflected forms, and in some cases multiword units (MWUs). One problem then becomes, in the case of a word having as synonyms several other words – because of homonymy and polysemy – to determine how many synsets we are dealing with. Also, for those familiar with WN, we should add that Synlex contains words of all PoS, and synonymy relations are sometimes between words with different PoS, just as in EuroWordNet.[1]

3. SALDO

SALDO (Borin, 2005; Borin and Forsberg, 2009; Borin et al., 2008; Borin and Forsberg, 2008), or SAL version 2, is a free modern Swedish semantic and morphological lexicon intended for language technology applications. The lexicon is available under a Creative Commons Attribute-Share Alike license and LGPL 3.0.

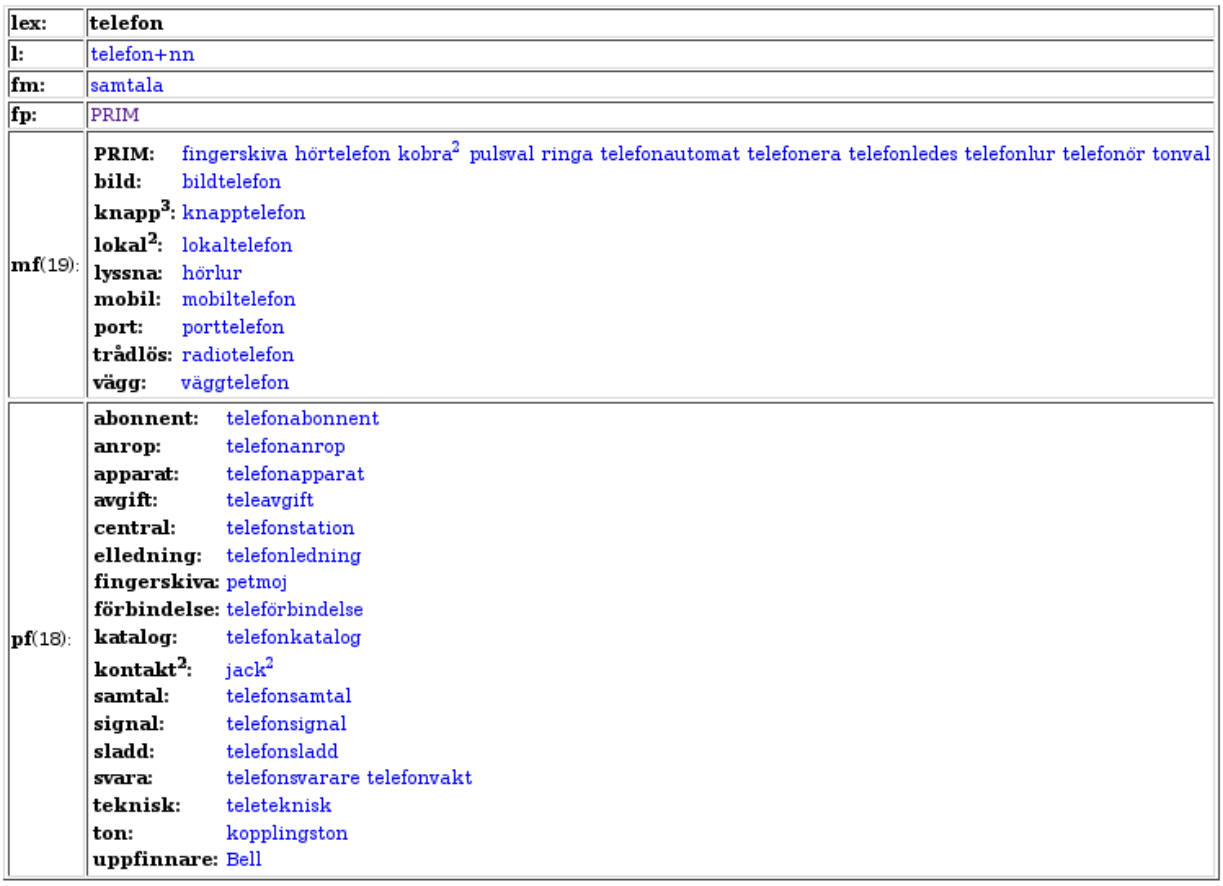
SALDO started its life as *Svenskt associationslexikon* (Lönngren, 1992) – 'The Swedish Associative Thesaurus' – a so far relatively unknown Swedish thesaurus with an unusual semantic organization, reminiscent of, but different from that of WordNet (Borin and Forsberg, 2009). SAL has been published in paper form in two reports, from the Center for Computational Linguistics (Lönngren, 1998), and the Department of Linguistics (Lönngren, 1992), both at Uppsala University. Additionally, the headwords and their basic semantic characterizations have been available electronically, in the form of text files, from the very beginning.

The history of SAL has been documented by Lönngren (1989) and Borin (2005). Initially, text corpora were used as sources of the vocabulary which went into SAL, e.g., a Swedish textbook for foreigners and a corpus of popular-scientific articles. A small encyclopedia and some other sources provided the large number (over 3000) of proper names found in SAL. Eventually, a list of the headwords from *Svensk ordbok* (SO, 1986) was acquired from the NLP and Lexicology Unit at the University of Gothenburg, and the second paper edition of SAL (Lönngren, 1992) contained 71,750 entries. At the time of writing, SALDO contains 76,200 entries, the increased number being because a number of new words have been added, but also because a number of entries belong to more than one part of speech or more than one inflectional pattern.

The central semantic relations of SALDO are based on *association*, a "non-classical" lexical-semantic relation (Morris and Hirst, 2004). SALDO describes *all* words semantically, not only the open word classes. It is organized by two primitive semantic (association) relations, one obligatory and one optional. Every entry must have a *mother* (or *main descriptor*), a semantically closely related entry which is more central, i.e., semantically and/or morphologically less complex, probably more frequent, stylistically less marked and acquired earlier in first and second language acquisition, etc. The mother will in practice often be either a hyperonym or synonym of the headword. However, it need not be either: Sometimes it is an antonym, and quite often it is a different part of speech from the headword, which takes us outside the realm of traditional lexical-semantic relations. An artificial most central entry, PRIM, is used as the mother of 50 semantically unrelated entries at the top of the hierarchy, making all of SALDO into a single rooted tree. An entry may also have an additional descriptor, the *father* (or *supplementary descriptor*), which serves to further characterize the entry semantically. The mother and father relations can then form the basis of any number of derived relations. Thus the m-sibling relation – 'having a common mother' – is very interesting, as such sibling groups tend to correspond to natural semantic groupings. Figure 1 shows how the SALDO entry for the Swedish noun *telefon* 'telephone' is associated to a number of other words: *samtala* 'hold a conversation' is the mother of *telefon*, while *telefonledes* 'by phone', *ringa* 'call v.', *mobiltelefon* 'mobile phone', *pulsval* 'pulse dialling' and a number of others are m-siblings having *telefon* as their mother. In the p-sibling group of *telefon* (senses having *telefon* as their father), we find *telefonkatalog* 'phone directory', *telefonsvarare* 'answering machine', the proper name *Bell* and a number of others.

We soon realized that in order to be useful in language technology applications, SAL would have to be provided at least with part-of-speech and inflectional morphological information – both entirely absent from SAL in its original form – and SALDO was created. The morphological component of SALDO has been defined using Functional Morphology (FM) (Forsberg and Ranta, 2004; Forsberg, 2007), a tool that provides a development environment for computational morphologies. It is a tool with a flexible language for

[1] Although in EuroWordNet this kind of synonymy is still formally distinct from within-PoS synonymy, bearing the label XPOS_NEAR_SYNONYM (Alonge et al., 1998, 25ff).



lex:	**telefon**
l:	telefon+nn
fm:	samtala
fp:	PRIM
mf(19):	**PRIM:** fingerskiva hörtelefon kobra[2] pulsval ringa telefonautomat telefonera telefonledes telefonlur telefonör tonval **bild:** bildtelefon **knapp**[3]**:** knapptelefon **lokal**[2]**:** lokaltelefon **lyssna:** hörlur **mobil:** mobiltelefon **port:** porttelefon **trådlös:** radiotelefon **vägg:** väggtelefon
pf(18):	**abonnent:** telefonabonnent **anrop:** telefonanrop **apparat:** telefonapparat **avgift:** teleavgift **central:** telefonstation **elledning:** telefonledning **fingerskiva:** petmoj **förbindelse:** teleförbindelse **katalog:** telefonkatalog **kontakt**[2]**:** jack[2] **samtal:** telefonsamtal **signal:** telefonsignal **sladd:** telefonsladd **svara:** telefonsvarare telefonvakt **teknisk:** teleteknisk **ton:** kopplingston **uppfinnare:** Bell

Figure 1: Semantic (associative) relations for *telefon* 'telephone n.' in SALDO (rendered in blue/non-bold)

defining morphological rules together with a platform for testing, which is used to minimize the risk of resource degradation during development. Furthermore, it has a rich export system, targetting around 20 formats, and supports both (compound) analysis and synthesis.

SALDO is, as one of its distribution channels, published as web services, updated daily. Web services provide clean interfaces and instant updates, but are restricted to small amounts of data because of network latency. Presently available web services include incremental fullform lookup, semantic lookup, compound analysis, and an inflection engine service. See <http://spraakbanken.gu.se/eng/saldo>.

4. From Synlex and SALDO to fuzzy synsets

Importantly to our purposes here, the basic units of SALDO are uniquely identified *word senses*. The current version of SALDO contains some 73,400 senses. Consequently, it is easy to find an answer to the question: "How many senses does a particular base form have?"[2] We can simply make an automated comparison between words in Synlex and word senses in SALDO via the Synlex words. From the point of view of Synlex, such a comparison yields five interesting sets, for a word w_i in Synlex (on the assumption – simplifying but largely correct – that Synlex contains pairs of base forms, including MWUs):

1. w_i is not a base form in SALDO
2. w_i occurs once in Synlex and it has one sense in SALDO
3. w_i occurs once in Synlex and it has several senses in SALDO
4. w_i occurs in several pairs in Synlex and it has one sense in SALDO
5. w_i occurs in several pairs in Synlex and it has several senses in SALDO

Calculating the set of Synlex pairs, such that each member of every pair is in one of set 2 or set 4 above – i.e., pairs where both members have only one SALDO sense – should then allow us to go on to calculate fuzzy synsets of various degrees from this set. Performing the first calculation yielded an initial set of 9,236 pairs, i.e., a bit less than half of Synlex (see the lower half of table 1). In the result set we replace each Synlex word form w_i with the corresponding SALDO word sense identifier l_i. For convenience we also mul-

[2]Base forms in SALDO include multi-word units.

set	size (Synlex entries)
1: 1 – 0	4,909
2: 1 – 1	4,779
3: 1 – many	645
4: many – 1	20,784
5: many – many	6,097
total 1–5	**37,214**

pair type	size (number of pairs)
set 2 – set 2	1,254
set 2 – set 4	2,144
set 4 – set 2	2,144
set 4 – set 4	6,097
total	**18,472**

Table 1: Connecting Synlex to SALDO

tiply the synonymy degree by a factor 20 in these pairs, making it range from 60 to 100.

In this paper, we report on our work on this subset of the Synlex entries, as part of a recently initiated project with the aim of bootstrapping a fuzzy wordnet for Swedish from Synlex and other available lexical resources (see section 9 below).

We have experimented with two different methods for constructing fuzzy synsets from Synlex: transitive closure (next section) and cliques (section 6).

5. Synset construction by transitive closure

Our first algorithm for building fuzzy synsets is a straightforward computation of the transitive closures of the word sense pairs, as follows. For every graded word sense pair with a degree higher or equal to d_{cutoff}, we check membership of the word senses in the current result set of synsets $Synsets$, and make the necessary adjustments to this set based upon their membership; in pseudocode:

$$
\begin{array}{l}
Synsets = \{\} \\
for\ \langle\langle l_i, l_j\rangle, d_k\rangle \in Synlex_{saldo} \\
\quad d_k \geq d_{cutoff} \\
\quad case\ membership(\langle l_i, l_j\rangle, Synsets)\ of \\
\quad\quad \langle S_1, S_2\rangle \Rightarrow Synsets.merge(S_1, S_2) \\
\quad\quad \langle S_1, \{\}\rangle \Rightarrow Synsets.add(l_j, S_1) \\
\quad\quad \langle \{\}, S_2\rangle \Rightarrow Synsets.add(l_i, S_2) \\
\quad\quad \langle \{\}, \{\}\rangle \Rightarrow Synsets.new(\{l_i, l_j\}) \\
return\ Synsets
\end{array}
$$

In other words, we calculate the synsets of degree d_{cutoff} by collecting in the same set all l_i that are connected by some path of graded synonymy relations where no relation has a degree less than d_{cutoff}.

The calculation of the transitive closure carries the hidden assumption that implicitly derived pairs are valid at the same degree as the pairs they are derived from. This assumption turns out to be rather problematic (see section 7 below).

$d \geq$	$\lvert SS\rvert$	$\lvert S\rvert = 2$	$2 < \lvert S\rvert \leq 25$	$\lvert S\rvert > 25$	$\max \lvert S\rvert$
60	1,485	951	530	4	3,893
70	1,641	1,026	602	13	1,245
80	1,640	1,066	566	8	441
90	1,068	800	268	0	18
100	416	362	54	0	6

Table 2: Synsets computed with transitive closure

$d \geq$	$\lvert SS\rvert$	$\lvert S\rvert = 2$	$2 < \lvert S\rvert \leq 25$	$\lvert S\rvert > 25$	$\max \lvert S\rvert$
60	1,533	956	560	17	1,598
70	1,650	1,034	602	14	921
80	1,609	1,047	556	6	397
90	1,016	761	255	0	18
100	394	347	47	0	5

Table 3: Synsets computed with PoS-constrained transitive closure

Applying the transitive closure algorithm to our 9,236 Synlex pairs[3] yields the results presented in table 2, to be read as follows: d is the degree; $|SS|$ is the number of synsets; $|S| = 2$ is the number of synsets of size 2; $2 < |S| \leq 25$ is the number of synsets of a reasonable size; $|S| > 25$ is the number of synsets of a suspiciously large size; and $\max |S|$ is the size of the largest synset.

The largest synset in this table is salient – at degree 60 it is as large as 3,893. This is an indication of a couple of things: first that the basic assumption about the implicit pairs is too strong, but also, that Synlex contains word senses that are missing in SALDO or Synlex has pairs that are simply wrong. In all cases we have pairs that merge reasonable synsets into a huge one. Interestingly, the smaller sized synsets are reasonable on manual inspection, which indicates that Synlex generally provides us with good information.

As a heuristic filter, we kept only same-PoS pairs, which resulted in the removal of 484 pairs, and repeated the experiment with the new set of 8,752 pairs. The result is presented in table 3, where the size of the largest synset at degree 60 has been halved by this filtering process, but the size is still significant.

6. Clique-based synset construction

A *clique* is a graph theoretic notion that describes a subgraph of a graph where all nodes are connected to all other nodes in the subgraph. If we require that all synsets are cliques, then we avoid the assumption about the implicit pairs.

The algorithm for creating cliques is simple: for every SALDO sense occurring in Synlex we create synsets by iteratively adding new lexemes that are connected to all previous ones.

[3]The number of SALDO word sense identifiers in this set is 8,594, i.e., the average synset size is slightly above 2.

$d \geq$	$\lvert SS\rvert$	$\lvert S\rvert = 2$	$2 < \lvert S\rvert \leq 25$	$\lvert S\rvert > 25$	$\max \lvert S\rvert$
60	6,933	5,687	1,246	0	22
70	4,931	4,313	618	0	16
80	4,024	3,673	351	0	16
90	1,582	1,542	40	0	11
100	484	482	2	0	7

Table 4: Synset cliques

$$
\begin{array}{l}
Synsets = \{\} \\
for\, l_i \in SALDO_{Synlex} \\
\quad SS = \{\{l_i\}\} \\
\quad while\, SS.exts\_exists(Synlex_{saldo}, d_{cutoff}) \\
\quad\quad for\, S\, in\, SS.has\_exts(Synlex_{saldo}, d_{cutoff}) \\
\quad\quad\quad ES = S.extensions(Synlex_{saldo}, d_{cutoff}) \\
\quad\quad\quad SS.extend(S, ES) \\
\quad Synsets.add(SS) \\
return\, SS
\end{array}
$$

A comment is in order here: Since it is possible that l_i is in more than one synset, with the operation `extend` we build a new synset for every possible extension. A natural optimization of the algorithm would be to divide the extension set ES into cliques to avoid rebuilding the same synset.

The result of running the algorithm on our material is given in table 4, and looks initially very promising, since there are no oversized synsets.

However, this algorithm has a hidden assumption, namely that all relevant pairs have been graded. This is not true for Synlex, which has the effect that the synsets end up being small, and worse, that some senses appear in many synsets, which strictly speaking is a contradiction in terms, since synsets and senses are two sides of the same coin on the WN view of things, so that a particular sense of a lemma should appear in only one synset. In some cases this is an indication that Synlex has a more fine-grained sense description than SALDO (see below), but in many cases it is an invalid split caused by a missing pair, i.e., a synonymy judgement missing from Synlex.

7. Degree computation for implicit pairs

A natural question at this point is: Is it possible to calculate new degrees for the implicitly derived pairs computed by transitive closure, i.e., given (l_1, l_2, d_1) and (l_2, l_3, d_2), could we not calculate a reasonable degree for (l_1, l_3) from d_1 and d_2?

In general, this is not a simple problem, and we will illustrate why with two pairs taken from Synlex:

integration	anpassing	60
anpassning	integrering	60

Here, *integration* 'integration' is related with the lowest Synlex degree to *anpassning* 'adaptation', which in turn is related to *integrering* 'integrating n.', again with the lowest Synlex degree.

What would be a reasonable degree for the derived pair *integration – integrering*? As already indicated by the glossing, they are completely synonymous, or nearly so, but there is no way to calculate this information from these two pairs.

Naturally, we have discussed various possible ways of performing this calculation. Since synonymy is normally considered both symmetric and transitive, there ought to be a standard way of calculating the degree of transitively derived synonymy even with graded synonyms. Interpreting degrees as distances in the plane and calculating the Euclidian distance under the assumption that the two synonymy links are at right angles to each other could possibly yield good results, on average (but not in this particular case, of course).

Another possibility, proposed here, is to annotate every synset with some standard statistical measures that reflect the composition of the grades, which can be used as basis for the calculation of the implicit links in the synset. Currently, we use *mean*, *standard deviation*, *min*, and *max*, as defined in this figure:

$$
\begin{array}{ll}
\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \mu)^2} & \mu = \frac{1}{n}\sum_{i=1}^{n} d_i \\
min = \min_i\{d_i\} & max = \max_i\{d_i\}
\end{array}
$$

As indicated above in section 5, we use the min measure in calculating the transitive closure of Synlex pairs, but we calculate all the measures for the resulting synsets (see figure 2, from <http://spraakbanken.gu.se/eng/swefn>)

8. Discussion

As we saw above, both methods for fuzzy synset construction have some drawbacks. However, the clique method has the more serious drawback – at least in our view – that in order for it to be used for synset construction, where a sense only occurs in one synset, we would need to somehow add information that is missing from Synlex, namely about the synonymy of unjudged pairs. For this reason, we have decided to adopt the transitive closure approach in our continued work. In fact, this approach offers a way of computing the synonymy of such pairs, as we discussed in the previous section, so that clique computation could be added as a refinement on top of it.

In working with the initial set of Synlex pairs and the fuzzy synset candidates automatically derived from this set by the transitive closure approach,

avg: 100 % **dev**: 0 % **min**: 100 % **max**: 100 %	anlita..1 leja..1
avg: 72 % **dev**: 0 % **min**: 72 % **max**: 72 %	annars..1 eljest..1
avg: 60 % **dev**: 0 % **min**: 60 % **max**: 60 %	annonsering..1 reklam..1
avg: 70 % **dev**: 14 % **min**: 60 % **max**: 90 %	integrering..1 integration..1 adaptation..1 anpassning..1
avg: 72 % **dev**: 8 % **min**: 64 % **max**: 80 %	tillreda..1 anrätta..1 tillaga..1
avg: 78 % **dev**: 16 % **min**: 64 % **max**: 100 %	nylle..1 nuna..1 anlete..1 ansikte..1
avg: 72 % **dev**: 0 % **min**: 72 % **max**: 72 %	anskri..1 stridsrop..1

Figure 2: Some synsets at threshold 60 (from <http://spraakbanken.gu.se/eng/swefn>)

we have been subjecting the result to constant manual evaluation, drawing upon our long experience of Swedish lexicography.

In its subdivision of lemmas into senses, SALDO reflects a well-established Swedish lexicographical tradition, which for practical reasons has tended to avoid excessively fine-grained sense distinctions. In paper dictionaries compiled in this tradition, definitions tend for this reason to be couched in very general terms, in order to cover as many different usages as possible.

WN, on the other hand, is all about different usages. The practice of defining synonymy as substitutability in at least one context, and then for all practical purposes defining senses using synsets, has the practical consequence that whenever we find that a word w_i can be substituted for another word w_j in one particular context, we will probably have to postulate a new sense both for w_i and w_j. Hence, WN lemmas by design will tend to have many senses (Vossen, 1998, 9). It seems that if we want manageable synsets, we have to accept a fine granularity of senses (as they are understood in WN).

The result presented above actually comes out of an iterative process where we have tried to identify problematic pairs using two simple diagnostics:

1. We examined senses with many connections to other senses, since many connections may be an indication that two or more senses have been collapsed into one in SALDO.
2. We inspected pairs that merged already large synsets, as a result of lowering the threshold for synset inclusion. The pairs that connect two large synsets together are not problematic in themselves, but they have the potential gain of reducing the size of the synsets drastically.

In both cases, the action to be taken is one of: 1. add a word sense to SALDO; 2. remove the pair from Synlex; 3. do nothing. In practice, all three have been necessary. The work with Synlex has thus allowed us to refine the semantic structure of SALDO in the direction of actual usage, which should be beneficial in a resource intended to be used in language processing. Reconciling the senses of SALDO – reflecting deep lexicographic thinking about words – with those of Synlex – a noisy and 'anarchistic' resource – will certainly raise many tricky and theoretically interesting problems, for lexicography and language technology alike.

9. Conclusion: Towards Swesaurus – a fuzzy wordnet for Swedish

A few thousand synsets do not a wordnet make. The work described above represents the first steps towards a Swedish lexical-semantic resource which we call Swesaurus, where our goal is to add classical lexical-

semantic relations and fuzzy synsets to the existing associative thesaurus structure of SALDO, thus combining classical and non-classical lexical-semantic relations in one resource.[4]

In this ongoing work, we can draw upon a number of other existing lexical resources, e.g.:

- Thus, we have extracted the lexical-semantic relations encoded in a conventional print dictionary, or rather, the database underlying this dictionary, where we find about 12,000 sense pairs explicitly labeled with one of five classical lexical-semantic relations: synonymy, hyponymy, hyperonymy, antonymy, cohyponymy, plus the 'lexicographic' relation often rendered as "see" in dictionaries. Computing the transitive closure of these sense pairs yields some 20,000 additional pairs, i.e., about 30,000 in total.
- The Swedish Wiktionary (close to 50,000 entries) provides lexical-semantic relations – e.g., synonymy, antonymy, "related words" – for a subset of its entries. This free resource also contains definitions of the senses, which we cannot get from other sources.
- We have further a lexical resource consisting of pairings of word senses from the same dictionary database mentioned above, with automatically extracted headwords of their dictionary definitions. Even though there are many invalid items in this extensive list (52,800 pairs), we believe that we can clean it with mostly automatic processing, using the other resources that we have at our disposal.

In fact, a decided advantage for our work is the fact that we can utilize several lexical resources containing overlapping information. This means that one resource can be used to disambiguate ambiguous information in another resource. For example, as mentioned above, the mother, or primary descriptor of a SALDO sense is in practice often a synonym or hyperonym of the sense. For those synonym pairs in Synlex – almost half – which have not been mapped to SALDO sense identifiers, because of a one-to-many or many-to-many mapping between the Synlex string and SALDO senses, we will use the heuristic that if the pair can be matched to a child-mother configuration in SALDO, the corresponding sense(s) will be chosen for the ambiguous member(s) of the pair, the reasoning being that if (the lemmas of) the senses of a SALDO child-mother relation map to a Synlex pair, then it is very likely that this particular mother happens to be a synonym of the child (although we could be wrong).[5]

Using such a strategy, we believe that we will be able to bootstrap a wordnet-like extension to SALDO, achieving a decent-quality resource with a fairly modest amount of manual work, because we are in the fortunate position of actually being able to reap the fruits of a large collected human effort that has gone into the creation of the existing resources we have at our disposal.[6]

10. Acknowledgements

The research presented here was supported by the Swedish Research Council (the project *Safeguarding the future of Språkbanken* 2008–2010, VR dnr 2007-7430) and the University of Gothenburg through its support of the Centre for Language Technology (the *Swedish FrameNet++* project) and through its support of Språkbanken (the Swedish Language Bank).

We would like to express our gratitude to the anonymous referees for their constructive remarks and insightful questions, which have contributed substantially to improving the presentation in this paper.

11. References

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti, and Wim Peters. 1998. The linguistic design of the EuroWordNet database. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 19–43. Kluwer, Dordrecht.

Lars Borin and Markus Forsberg. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, Göteborgs universitet.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. Odense.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a

[4] Incidentally, this is the reverse of what is going on with the Princeton WordNet at this moment, where associative relations are being added (called "evocation" on the WN website).

[5] Another possibility is to first map the print dictionary database synonym pairs (and possibly its "see" relation pairs as well, since we are dealing with graded synonymy in Synlex) to SALDO, and then match this list against the remaining Synlex pairs. Again this would be an example of 'synergistic' merging of several existing lexical resources.

[6] Unfortunately, an existing embryonic Swedish wordnet is not one of those resources, since it has not been developed since about 2002, and further its licensing format prohibits its inclusion in the open source resource that we wish to develop.

freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beata Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

Lars Borin. 2005. Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat. *LexicoNordica*, 12:39–54.

Andrew Carstairs-McCarthy. 1999. *The origins of complex language*. Oxford University Press, Oxford.

Jiwei Ci. 2008. Synonymy and polysemy. In Patrick Hanks, editor, *Lexicology: Critical concepts in linguistics. Vol. III: Core meaning, extended meaning*, pages 191–207. Routledge, London. Reprinted from Lingua 72 (1987): 315–331.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.

Markus Forsberg and Aarne Ranta. 2004. Functional morphology. In *ICFP'04. Proceedings of the ninth ACM SIGPLAN international conference of functional programming*, Snowbird, Utah. ACM.

Markus Forsberg. 2007. *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract*. Ph.D. thesis, Göteborg University and Chalmers University of Technology.

Viggo Kann and Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, pages 105–110. Department of Linguistics, University of Joensuu.

Lennart Lönngren. 1989. *Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi*. Centrum för datorlingvistik. Uppsala universitet. Rapport UCDL-R-89-1.

Lennart Lönngren. 1992. *Svenskt associationslexikon. Del I-IV*. Institutionen för lingvistik. Uppsala universitet.

Lennart Lönngren. 1998. A Swedish associative thesaurus. In *Euralex '98 proceedings, Vol. 2*, pages 467–474.

George A. Miller. 1998. Nouns in WordNet. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 23–46. MIT Press, Cambridge, Mass.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In Dan Moldovan and Roxana Girju, editors, *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 46–51, Boston. ACL.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

SO. 1986. *Svensk ordbok*. Esselte Studium, Stockholm.

Piek Vossen. 1998. Introduction to EuroWordNet. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Kluwer, Dordrecht.

PredXtract, a generic platform to extract in texts predicate argument structures (PAS)

Elisabeth Godbert, Jean Royauté

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
CNRS UMR 6166 - Université de la Méditerranée
Parc Scientifique et Technologique de Luminy, case 901
13288 Marseille Cedex 9
Elisabeth.Godbert@lif.univ-mrs.fr, Jean.Royaute@lif.univ-mrs.fr

Abstract

Verbal and nominal predicate structures present interesting properties for information extraction. We show how to study these predicate structures in a uniform way, using the fact that the nominalization of a verb has the same arguments as the verb. We then describe the extraction platform (PredXtract) which we have developed in order to extract predicate argument structures and which highlights relations between biological entities in biological texts. We present and discuss our results.

1. Introduction

This paper focuses on the extraction of verbal and nominal predicate structures, which can be expressed in a great variety of forms (Meyers et al., 2004a). Defining a uniform representation for these structures is decisive to converge on a VerbNet, PropBank or FrameNet representation (Kipper et al., 2000; Wattarujeekrit et al., 2004; Miyao et al., 2006; Levin, 1993) and to acquire semantic relations.

In predicate-argument representation, verbs and their nominalizations are the most productive predicates and have the same argument relations, where arguments play precise conceptual roles: subjects and complements, which are core arguments, and adjuncts. With a nominalization, it is possible to build complex noun-phrases (NPs), in which the head noun is bound to prepositional phrases (PPs) with specific prepositions which mark core arguments or adjuncts. For example, the NP *milk concentration by ultrafiltration* is related to the sentences *ultrafiltration concentrates milk* and *milk is concentrated by ultrafiltration*: the NP is built with the predicate head *concentration*, preceded or followed with its arguments *ultrafiltration* and *milk*, whether or not it is introduced by a preposition. In these structures, the core arguments are preserved and it is possible to insert an adjunct (*in the manufacture of cheese*).

Verbal and nominal structures are closely correlated; we will show in Section 2. how to link an NP built with a nominalization, to a core sentence. We use the following notation: N_0 *V W*, where N_0 is the subject of the verb, V the verb and W, a sequence of complements (N_1 ... N_n) linked to the verb (Gross, 1986).

Our objective is to define a underspecified semantic representation where (i) the predicate (nominal or verbal) expresses the action, (ii) the subject is the Agent (who performs the action) (iii) the object is the Patient (who is involved by the action) and (iv) possible adjuncts express context of the action ; this semantic role is named here Circumstance. Thus, this representation is situated at the syntax-semantics interface. Distinguishing the core arguments from the adjunct arguments in predicate structures (Tesnière, 1959) is important in information extraction and particularly in scientific sublanguage. Later, this semantic representation will be enriched by including more complex roles derived, for example, from semantic frames of VerbNet.

At present, we have developed a robust platform, PredXtract, based on the Link Parser (Sleator and Temperley, 1991). This platform is a generic tool which extracts verbal and nominal predicate argument structures (PAS) in English texts. More specifically, it exhibits relations between biological entities.

2. Nominal and verbal argument structures

We present here a typology of seven classes of verbal and nominal structures, defined from their core arguments:

- Verbs accepting a direct object are grouped together in Class 1 and 2; in the corresponding predicate noun phrases (PNPs), the preposition *of* marks the direct object.
- Verbs that do not accept a direct object are grouped together in Class 3 to 5; in the corresponding predicate noun phrases (PNPs), the preposition *of* marks the subject.
- Symmetric predicates with interchangeable arguments concern Class 6 and 7.

This classification has been elaborated, from scientific texts of the web, and from the grammar of English described in (Quirk et al., 1987), as well as from the data of "The Specialist Lexicon", which gives, for all verbs, their nominalizations and the different prepositions that can introduce core arguments (www.nlm.nih.gov/pubs/factsheets/umlslex.html).

2.1. The preposition *of* as marker of the object

Class 1: N_0 *V* N_1 = N^{pred} *of* N_1 *by* N_0. This class groups together predicates with a direct object and which accept passive voice (N_1 *is* V_{ed} *by* N_0). This is the most important class with more than 1,000 couples of verbs/nominalizations. For example, the couple *activate/activation* belongs to this class : *IFN-gamma activates protein kinase C delta / activation of protein kinase C delta by IFN-gamma*.

Class 2: N_0 *V* N_1 *Prep* N_2 = N^{pred} *of* N_1 *Prep* N_2 *by* N_0. This class concerns constructions with a direct object and with a second complement introduced with a preposition inherited from the verbal construction. This preposition is the same in the verbal construction and the nominal construction. These constructions also accept passive voice. Example: N_0 *attributes a protein fragment* ***to*** *a sequence / attribution of a protein fragment* ***to*** *a sequence by* N_0.

2.2. The preposition *of* as marker of the subject

Class 3: N_0 *V* = N^{pred} *of* N_0. This class concerns constructions without complement. In the NP construction, the preposition *of* introduces the subject argument. Example: *the femoral head necroses / necrosis of the femoral head*.

Class 4: N_0 *V Prep* N_1 = N^{pred} *of* N_0 *Prep*N_1. This construction can appear without a prepositional complement but if the complement is present, the same preposition introduces it in the sentence and in the NP. As in Class 3, the preposition *of* marks the subject argument in the NP. Example: *tryptophans fluctuates in gramicidin / fluctuation of tryptophans in gramicidin*.

Class 5: N_0 *V Prep* N_1 *Prep* N_2 = N^{pred} *of* N_0 *Prep* N_1 *Prep* N_2. In this class, the two prepositions which appear in the sentence also appear in the NP. Example: *temperature decreases from 200 K to 70 K / decrease of temperature from 200 K to 70 K*.

2.3. Predicates with permutable arguments

Class 6: N_a *V with* N_b = N^{pred} *of* N_a *with* N_b = N^{pred} *of/between* N_a *and* N_b. This is a special class because the arguments can permutate without a change in the meaning. For that reason we noted them N_a and N_b. Examples: *genes interact with proteins*; *interaction of genes with proteins / interaction of/between genes and proteins*.

Class 7: N_0 *V* N_a *Prep* N_b = N^{pred} *of* N_a *with/to* N_b *by* N_0 = N^{pred} *of/between* N_a *and* N_b *by* N_0. We consider that this class is a variant of Class 6 because N_a and N_b are in the complement position in the sentence. For example, from the sentence N_0 *connects a new sequence with/to a cluster*, it is possible to derive several NPs : *connection of a new sequence with/to a cluster / connection of/between a new sequence and a cluster*. In these different constructions, the N_0 argument can be absent in the sentence or in the NP.

In all classes, the arguments introduced by prepositions *of* or *by* can be in the position of left modifier of the nominalization (*regulation of VEGF by TGFbeta1 / VEGF regulation by TGFbeta1 / TGFbeta1 Regulation of VEGF*).

3. PredXtract, an extracting platform

The PredXtract platform produces the representation of a sentence in a set of complex predicate argument structures. PredXtract uses the Link Parser (LP) and its English native Link Grammar (LG), a variant of dependency grammars (Sleator and Temperley, 1991). The sentence processing of the LP produces a set of graphs where words are linked in pairs by labeled arcs with grammatical functions; each graph corresponds to a possible analysis. In LG, generic links attach verbs (`MVp` link) or nouns (`Mp` link) to any preposition which introduces an NP.

In order to mark the precise role of each argument of the predicates, we have: these six sentences receive the (i) defined specific argument links, in order to distinguish core arguments from adjunct arguments during the extraction process; (ii) integrated in the native grammar of the LP, a grammatical module to parse predicate NPs with specific argument links; (iii) post-processed the parse to align argument links of the verbs to the argument links defined for nominalizations; (iv) modified the classification heuristics of the LP parses because they are not always adapted to biomedical texts (Pyysalo et al., 2006) and because the predicate NP attachments are often not correct.

Besides, to enhance the accuracy of the parsing, we have followed Szolovits (2003) and added in the grammar all of the words of "The Specialist Lexicon" (SL), which includes UMLS terms. We have also added a lexicon of genes and proteins extracted from corpus. The lexicon contains about 400,000 lexical items (500,000 inflected forms).

We describe below the different processes and components of PredXtract.

Link Grammar of nominalizations.

Several teams in biomedecine use the LP but without modifying its grammar (Ding et al., 2003; Hakenberg et al., 2009). This parser is also used in other domains as the information extraction in Reuter corpus (Madhyastha et al., 2003). According to our classification of the nominalizations, we have added to the native LP a grammar module of PNPs in which about 3,900 nominalizations are divided into 89 subclasses. Each subclass corresponds to a syntactic pattern with core arguments (including clauses with *that*) and adjuncts.

All of the words in the LG appear in the same format: just the inflected form or the inflected form followed by a dot and an extension. The extension (a short sequence of alphanumeric characters) allows to re-use the same word in different disjoint linguistic descriptions. Each nominalization belongs to one or more subclasses and can accept one or more syntactic descriptions; in these cases, specific extensions are used.

Figure 1 shows several examples of extensions: the `nt0` extension corresponds to the nominalizations of transitive verbs (*regulate / regulation, product / production, accumulate / accumulation*), `ni2` (*respond / response*) corresponds to the nominalizations of prepositional verbs with the preposition *to*, and `ndt7` (*treat / treatment*) corresponds to the nominalization of the verb with a direct object and a complement introduced by the *with* preposition.

We see in Figure 1 parses of two short sentences with five nominalizations. In the first sentence (example 1), *response* has two arguments : the `MSI` link marks the subject introduced by the preposition *of*, while the `MCITO` link marks the complement introduced by *to*. The second NP shows the prepositional use of *treatment*, not saturated in this case: it has only one argument, introduced by the preposition *with* (link `MCDTWI`) inherited from the verb.

In Example 2, the two nominalizations *production* and *accumulation* have a left argument marked by the ASOT link. This link means that the argument can be subject or object. In this case, the argument role remains underspecified in this modifier position, because it is not possible to specify the argument role when a prepositional position is lacking.

Verb-noun alignment.
The native grammar of the LP does not distinguish core arguments and adjunct arguments for verbs: it marks all prepositional complements with the same link (MVp).
Rather than writing a grammar for verbs, which would have been very complex, we have defined a module that aligns verb arguments to nominalization arguments during a post-processing step. This module therefore produces a representation of verbs similar to the representation of nominalizations. For this, we use the data of "The Specialist Lexicon" (SL) which gives, for all verbs, the prepositions that can introduce a core argument. This module performs several tasks: (i) distinguish complements from adjuncts of verbs, by using the data of SL, and substitute the generic MVp link with a specific argument link when appropriate; (ii) identify each "verbal sequence" (compound with a verb and a set of possible auxiliaries, negation, and modal verbs); (iii) identify arguments in passive or active voice, and interchangeable arguments.

Recognition of syntactic arguments.
For each parse of a sentence, all of the predicates and their arguments are identified. Each argument link points on the heads of core arguments or on a word which introduces it (a preposition or a conjunction). Then the surface structure of each argument is reconstructed via the links, by using linguistic criteria. The reconstructed arguments can be NPs (most cases), clauses or adverbs.

Filtering of parses.
For each sentence, the parses (often several thousands) are re-ordered by attributing to each parse a score defined through several criteria. Among the main criteria:
(i) in the case of multiple prepositional attachments to verbs or nouns, we favor parses whose number of argument links is maximum - a higher score is given to these parses;
(ii) for the treatment of PNPs containing several nominalizations, we favor prepositional arguments attached to the head of the PNP; a specific score is calculated in the case of these PNPs.
This second point is illustrated in (Figure 1, examples 2-a and 2-b) with the two parses of the same sentence. In this sentence, with three nominalizations derived from transitive verbs, the preposition *by* can be attached either to *regulation* or to *production* with the MST argument link. We favor the parse given in example 2-a, because the first nominalization (*regulation*) is in a saturated form, ie. with all core arguments: the subject argument (*matrilysin production*) is marked with MST link and the object argument (*beta-catenin accumulation*) is marked with MOT link.

Syntax-semantic interface.
For each sentence, PredXtract produces an underspecified semantic representation, which is close to the syntax. As we have seen, we separate core arguments from adjunct arguments. On this basis, we identify core arguments in several alternation forms. Following Cohen et al. (2008), we extend the paradigm of alternations to the nominalizations. For example, the following sentences illustrate different surface forms of similar PAS of the verb *regulate*:
(i) *Fatty acids and eicosanoids regulate gene expression*;
(ii) *telomerase activity is mainly regulated by hTERT*;
(iii) *a DNA binding protein regulated by IL-4 (...)*;
(iv) *a unique mechanism regulating gene expression (...)*;
(v) *the regulation of eIF4E by 4E-BP phosphorylation is performed at its free state*;
(vi) *this study reports the first evidence of VEGF regulation by heregulin in cancer cells*.
To unify the semantic representation, two macro-roles have been defined: Agent and Patient. These two macro-roles are present in these examples with the same PAS. Sentence (i) is in the active voice and the two NPs *Fatty acids* and *eicosanoids* are the Agent and *gene expression* the Patient; sentence (ii) is in the passive voice: the Agent is *hTERT* and *telomerase activity* is the Patient; sentences (iii) and (iv) show this verb in participial modifier forms (past and present) with respectively *IL-4* and *A unique mechanism* as Agent and *A DNA binding protein* and *gene expression* as Patient; finally, the last sentences (v) and (vi) show two nominal forms of *regulate* with respectively *4E-BP phosphorylation* and *heregulin* as Agent and *eIF4E* and *VEGF* as Patient.
At present, PredXtract does not take into account all possible syntactic alternations, which correspond to all the different ways in which verbs can express their arguments (Levin, 1993).

4. Results and discussion

4.1. PredXtract outputs

In this section, we present results obtained with PredXtract, by showing, for each sentence, the analysis which has obtained the best score. In the first and the two last examples, the extraction of all the predicates and their arguments were identified, and the analysis is correct. In the two others, almost all the predicates and their arguments were identified, but the analysis is not completely correct.

Example 1: from the sentence *Hyperoxic exposure induced an S-phase arrest associated with acute inhibition of Cdk2 activity and DNA synthesis*, 9,168 parses were found and PredXtract outputs:

```
-----------------------------------------------------
Nominalization 1: exposure

Nominalization 2: arrest
 Agent or Patient: S-phase

Nominalization 3: inhibition
 Patient: Cdk2 activity
 Patient: DNA synthesis

Nominalization 4: synthesis
 Agent or Patient: DNA

Verb 1: induced (verbal sequence: induced ; active)
 Agent: hyperoxic exposure
 Patient:  an S-phase arrest associated
    with acute inhibition of [...] synthesis

Verb 2: associated (verbal sequence:
                              associated ; passive)
```

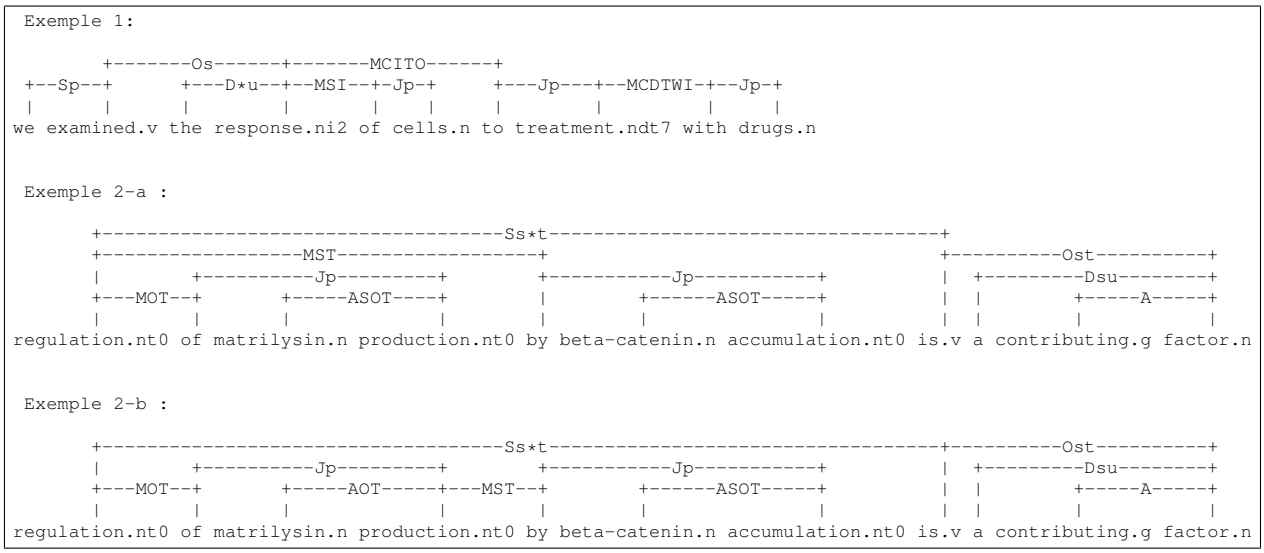


Figure 1: LP parses with several nominalizations.

```
 Patient A: an S-phase arrest
 Patient B: acute inhibition of Cdk2 activity
    and DNA synthesis
-------------------------------------------------------
```

This example shows a short sentence with six predicate structures. We can notice that (i) *exposure* has no argument, (ii) *inhibition* has two coordinated Patient roles, (iii) the role of the argument of *arrest* and *synthesis* is underspecified (Agent or Patient), and (iv) the verb *associated* has two interchangeable arguments (Patient A and Patient B).

Example 2 : with the sentence *Moreover, overexpression of dominant negative SHP2 blocked the protective effect of IL-6 against Dex-induced apoptosis*, the parser produces 64 parses and the output is:

```
-------------------------------------------------------
Nominalization 1: overexpression
 Patient: dominant negative SHP2

Nominalization 2: effect
 Agent: IL-6
 Patient: {against} Dex-induced apoptosis

Nominalization 3: apoptosis
 Agent: Dex-induced

Verb 1: blocked (verbal sequence:  blocked ; active)
 Agent: overexpression of dominant negative SHP2
 Patient: the protective effect of IL-6 against
     Dex-induced apoptosis
-------------------------------------------------------
```

In this example, the identification of all predicates and arguments are correct except for *apoptosis* where the Agent argument (*Dex-induced*) is not correct. The grammar of predicate NPs does propose an adjective as argument, but in this case *Dex-induced* is a compound adjective and was not registered as adjective in the grammar. In its present state, our system does not handle these compounds well.

Example 3: for the sentence *ET-1 expression and increased permeability may occur secondary to PKC isoform activation and may be modulated by VEGF and nitric oxide*, the parser produces 24 parses and the PredXtract output is:

```
-------------------------------------------------------
Nominalization 1: expression
 Agent or Patient: ET-1

Nominalization 2: activation
 Agent or Patient: PKC isoform

Verb 1: increased  (verbal sequence:
                      increased ; passive)
 Patient: permeability

Verb 2: occur  (verbal sequence:
                     may occur ; active)
 Agent: ET-1 expression
 Agent: increased permeability
 Circumstance: {to} PKC isoform activation

Verb 3: modulated (verbal sequence:
                      may be modulated ; passive)
  Agent: VEGF
  Agent: nitric oxide
  Patient: ET-1 expression
  Patient: increased permeability
-------------------------------------------------------
```

We can note: (i) the use of the modal *may* which operates on the verbs *occur* and *modulated* and which is included in the verbal sequence, (ii) the identification of the coordinate arguments of these two verbs, and (iii) an error with the Circumstance argument of the *occur* verb which is incomplete: *secondary* was ignored because the idiom *secondary to* was not recognized.

The following two short examples illustrate the presence of Circumstance roles in the verbal (Example 4-a) and nominal structures (Example 4-b) and their identification in these two structures.

Example 4-a: in this sentence *Characterization of these essential modules in transcription factors has been hampered by their low sequence homology*, the parser produces eight parses and the PredXtract output is:

```
-------------------------------------------------------
Nominalization 1 : characterization
 Patient: these essential modules
 Circumstance: {in} transcription factors

Nominalization 2 : transcription
```

```
Verb 1: hampered (verbal sequence:
                  has been hampered ; passive)
 Agent: their low sequence homology
 Patient: characterization of these essential
           modules in transcription factors
-------------------------------------------------------
```

We can see that the nominalization *characterization* has two arguments: a Patient role (*these essential modules*) and a Circumstance role (*in transcription factors*) .

Example 4-b: in this other sentence *An association between cyclin D3 and the C-terminal domain of pRb2/p130 was demonstrated using the yeast two-hybrid system* the parser produces 124 parses and the PredXtract output is:

```
-------------------------------------------------------
Nominalization 1 : association
 Agent A: cyclin D3
 Agent B: the C-terminal domain of pRb2/p130

Verb 1: demonstrated (verbal sequence:
                       was demonstrated ; passive)
 Patient: an association between cyclin D3 and
           the C-terminal domain of pRb2/p130
 Circumstance: using the yeast two-hybrid system
-------------------------------------------------------
```

In this last example, we focus on Circumstance role (*using the yeast two-hybrid system*) in the verbal structures (*demonstrated*). This structure has another argument which is a Patient role (*an association between cyclin D3 and the C-terminal domain of pRb2/p130*). We can also notice the two co-agents : *cyclin D3* and *the C-terminal domain of pRb2/p130* of the nominalization *association* derived from the symetric verb *associate*.

4.2. Evaluation

PredXtract has been evaluated with a corpus of 335 Medline [1] abstracts given by biology researchers. From the 3,500 sentences of this corpus, we have selected 700 random sentences; 300 of them have been used to finalize our system and the evaluation has been done on the 400 others. In this evaluation we take into account the false positives, which are the PAS produced by the system, but which are false, and the true negatives which are the PAS that are not extracted. Because of the possibility of wrong segmentation of arguments, we have calculated two values for recall, precision and F-measure, with:
(i) [Case 1] only the true and complete arguments (the true but incomplete arguments are scored as missing arguments),
(ii) [Case 2] the true and complete arguments and the true but incomplete arguments.
The 400 sentences contain 708 nominalizations and 965 verbs; thus, nominalizations represent 42.3% of all predicates. Besides, the length of the sentences ranges from 10 to 60 words.

Table 1 presents the evaluation results for the nominalizations (N) and the verbs (V).
These results show a very small difference between values for nominalizations and verbs (at the most 0.04). So we can say that PredXtract identifies the arguments in an uniform way.

[1] Medline : a bibliographic database of biomedical information

	N	V
True and complete arguments	508	1668
True but incomplete arguments	46	225
False arguments	86	254
Missing arguments	108	260
Case 1		
Recall	0.77	0.77
Precision	0.79	0.78
F-measure	0.78	0.77
Case 2		
Recall	0.84	0.88
Precision	0.87	0.88
F-measure	0.85	0.88

Table 1: Evaluation of verbal and nominal PAS.

Our system obtains rather good results in the identification of arguments in case of multiple possible prepositional attachments. The main problems in parses come from long distance attachments and coordination.
Besides, we have also calculated the recall for each sentence. We observed that there is no clear relation between sentence length (from 10 to 60 words in our evaluation) and recall values.

4.3. Related research

Much research has been published on predicate argument structures but it is difficult to compare research because objectives are often different: as for PredXtract, it is a generic system which extracts PAS of all predicates (nominal and verbal) in the sentences processed ; the other systems, in general, aim to extract specific templates.

In biomedicine, research focuses on PAS dedicated to gene /protein interaction, where two genes or proteins are in a subject and a complement position in a proteomic relation. For example, McDonald et al. (2004) work on the specific sublanguage of gene-pathway relations, and obtain a precision rate of 89% and a recall rate of 61% with a complete parsing ; Huang et al. (2004), on protein-protein interactions, have a precision rate of 80.5% and a recall rate of 80% with a pattern-matching processing.

As Cohen et al. (2008) observe, research on nominalizations in biomedicine is very limited. Current research has rarely handled nominalizations extensively. Leroy et al. (2003) use templates built around a small set of prepositions (*of*, *in* and *by*) to capture relations with genes, proteins, gene locations, diseases, etc., they use a shallow-parsing with finite state automata and obtain 90% of precision. A specific work on PP attachments on nominalizations (Schuman and Bergler, 2006) in proteomic texts achieves good results (precision: 82%) with linguistic heuristics using information of "Specialist Lexicon" nominalizations, but the system does not produce information on the PP roles (subject, object or adjunct).

Concerning nominalizations in other texts than biology, the

first version of NOMLEX (Macleod et al., 1998) is used in information extraction (Meyers et al., 1998). The NOMBANK project (Meyers et al., 2004b) annotates automatically, semi-automatically and manually, in corpus (the Wall Street Journal Corpus of the Penn Treebank), predicate nouns (verbal, adjectival and other) with their argument relations and improves the lexical base of predicate nouns (NOMLEX-PLUS). These annotated corpora are particularly used for automatic learning.

5. Conclusion

PredXtract is a robust platform organized around the Link Parser. It parses long sentences and extracts verbal and nominal predicate argument structures. For the parsing of verbal structures as well as nominal structures, the recall, precision and F-measure values are around 0.78 without a significant difference between its three measures. This is interesting because nominalizations represent 43% of all predicates of the corpus, and thus bring a large added amount of information. These results confirm our choice to make an appropriate and effective processing of the nominalizations, as shown by Miyao et al. (2006): these authors work on similar texts and observe that their system has difficulties in processing the prepositional phrases, especially when they appear in predicate noun phrases.

As PredXtract is based on very large lexicons, it can be considered as a platform which extensively recognizes PAS, independently from the predicate type. At present, we use it for extraction of PAS in biomedical texts. To adapt it to another domain would require the addition of possible sets of specific lexical items.

To refine PredXtract outputs, we are considering extending our description of verbs with Verbnet (Kipper et al., 2000), giving special attention to the description of diathesis alternations. In our description of verbs, the classical example with *spray* (Levin, 1993) *Jack sprayed paint on the wall / Jack sprayed the wall with paint* is not taken into account at present with our system. A more precise description of these syntactic frames will also allow to improve the syntactic frames of predicate noun phrases. For a more accurate semantic description of predicates we will add semantic roles and predicate classes derived from VerbNet. In the biomedical domain, the next step will require the annotation of arguments with UMLS or other biomedical term resources.

Acknowledgments

Many thanks to Christine Brun and Bernard Jacq of LGPD-CNRS for having supplied us with their corpus of Medline abstracts tagged with gene nouns.

6. References

K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9):e3158, 09.

Jing Ding, Daniel Berleant, Jun Xu, and Andy W. Fulmer. 2003. Extracting biochemical interactions from medline using a link grammar parser. In *ICTAI '03: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, page 467, Washington, DC, USA. IEEE Computer Society.

M. Gross. 1986. Lexicon-grammar: the representation of compound words. *International Conference On Computational Linguistics. Proceedings of the 11th conference on Computational linguistics*.

Jörg Hakenberg, Illés Solt, Domonkos Tikk, Luis Tari, Astrid Rheinländer, Quang Long Ngyuen, Graciela Gonzalez, and Ulf Leser. 2009. Molecular event extraction from link grammar parse trees. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 86–94, Morristown, NJ, USA. Association for Computational Linguistics.

M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li. 2004. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

Karin Kipper, Hoa Trang Dang, and Martha Stone Palmer. 2000. Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696.

G. Leroy, H. Chen, and J. D. Martinez. 2003. A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, 36:145–158.

B. Levin. 1993. English verb classes and alternation: A preliminary investigation. *The University of Chicago Press*.

C. Macleod, R. Grishman, A. Meyers, L. Barret, and R. Reeves. 1998. Nomlex: A lexicon of nominalizations. *In Proceedings of the Eighth International Congress of the European Association for Lexicography.*, pages 187–193.

Harsha V. Madhyastha, N. Balakrishnan, and K. R. Ramakrishnan. 2003. Event information extraction using link grammar. *Research Issues in Data Engineering, International Workshop on*, 0:16.

D. M. McDonald, H. Chen, H. Su, and B. B. Marshall. 2004. Extracting gene pathway relations using a hybrid grammar: the arizonarelation parser. *Bioinformatics*, 20(18):3370–3378.

A. Meyers, C. Macleod, R. Yangarber, R. Grishman, L. Barrett, and R. Reeves. 1998. Using nomlex to produce nominalization patterns for information extraction. *Proceedings of the COLING-ACL '98 Workshop on Computational Treatment of Nominals, Montreal, Canada.*

A. Meyers, R. Reeves, C. Macleod, R. Szekeley, V. Zielinska, B. Young, and R. Grishman. 2004a. The cross-breeding of dictionaries. *In proceedings of LREC-2004, Lisbon, Portugal*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004b. The nombank project: An interim report. *In proceeding of HLT-EACL Workshop: Frontiers in Corpus Annotation.*

Y. Miyao, O. Tomoko, M. Katsuya, T. Yoshimasa, Y. Kazuhiro, N. Takashi, and J. Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. *In Proceedings of the COLING-ACL, Australia*, pages 1017–1024.

Sampo Pyysalo, Filip Ginter, Tapio Pahikkala, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2006. Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions. *International Journal of Medical Informatics*, 75(6):430–442, June.

Randolph Quirk, Sydney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1987. *A Comprehensive grammar of the English Language.* Longman.

Jonathan Schuman and Sabine Bergler. 2006. Postnominal prepositional phrase attachment in proteomics.

D. Sleator and D. Temperley. 1991. Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science technical report, CMU-CS-91-196, Carnegie Mellon University,USA.*

P. Szolovits. 2003. Adding a medical lexicon to an english parser. *In Mark Musen, editor, Proceedings of the 2003 AMIA Annual Symposium*, pages 639–643.

Lucien Tesnière. 1959. *Eléments de syntaxe structurale*. Klincksieck, Paris.

T. Wattarujeekrit, P. K. Shah, and N. Collier. 2004. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5: 155.

Semantic relations of Adjectives and Adverbs in Estonian WordNet

Kadri Kerner, Heili Orav, Sirli Parm
University of Tartu
Liivi 2-308
E-mail: kadri.kerner@ut.ee, heili.orav@ut.ee, sirli.parm@ut.ee

Abstract

After several years of developing Estonian WordNet, we have realized that the list of semantic relations proposed within EuroWordNet project is not quite compatible with Estonian WordNet. Up to the present day we were following the structure of English according database, without finding the resources for Estonian approach. Now we are able to deal with this approach – we have the common knowledge of the deficiency and needs of the thesaurus. We still need a fixed list of semantic relations which have to be Estonian language specific. This paper gives an overview of semantic relations in Estonian WordNet. We will focus mainly on semantic relations among adjectives and adverbs. We discuss the problems of creating possible hypernymy/hyponomy relation between adjectives. Then we will describe which types of adverbs can be connected to adjectives by using derivational relations. Also, we propose different types of semantic relations to adverbs. Finally, some issues about relations between nouns are described.

1. Introduction

It is useful to create an adequate set of semantic relations in Estonian WordNet (EstWN), because some natural language applications using wordnet as a valuable language technology resource would benefit from it, for example Semantic Webs, ontologies, word sense disambiguation systems, machine translation and question-answering.

Since there are not many wordnets created for Finno-Ugric languages which examples to follow, then we have to create semantic relations for adjectives and adverbs only leaning on the specifics of Estonian language.

The most known handbooks for theoretical background about semantic relations in linguistics are the ones of Alan Cruse (1986, 2000), where the overview of different semantic relations for English language is given. Semantic relations of English adjectives and adverbs have been dealt by Kathrine J. Miller (1998). The same problems have been dealt with in several publications which analyze the essence of language from a theoretical aspect. The important example is from the founder of conceptual semantics, Ray Jackendoff's book (2002). He discusses through many chapters about the essence of meaning; most of all about the relations between pure language meaning and conceptual knowledge, also about the relations between meanings and about relations between different features of word classes. On the other hand many researchers have been dealing with semantic relations considering only some particular relation. So we have different studies all over the world, for example Vanhatalo (2005) looked for the synonymy relation in Finnish language, Õim (1991) in Estonian language and Apresjan (2006) in Russian language; the expression of the relation of antonymy in Estonian is studied by Õim (1995); the expression of Sweden adjectives by Vogel et al (2004).

Also, semantic relations have been studied in connection of creating wordnets and thesauri. For example the study of lexical and semantic aspect of English nouns is analyzed by Miller (1988); the semantic aspect of verbs is analyzed by Fellbaum (1988) and of Estonian by Vider (1999) and Orav (1998); also of adverbs and adjectives by Miller (1988), Fellbaum and others (1990); same kind of studies of Estonian adjectives are carried out by Orav (2006) and of Estonian adverbs by Parm (2007). In addition to previously mentioned studies, the Proceedings of Global WordNet Conference is being published (see 2002, 2004, 2006, 2008, 2010)[1], in which it is possible to find topics about the problems of semantic relations in other languages, while creating thesauri.

2. Estonian WordNet

There are two thesauri available for Estonian. First thesaurus (Saareste, 1979) has more of an historic value (compiled by Andrus Saareste as war refugee in Uppsala in 1979) and second, the modern one is the *wordnet*-type thesaurus of Estonian. The creation of the Thesaurus of Estonian Language[2] was started in 1998 within the project EuroWordNet (EWN, see also Vossen 1998)[3]. Estonian WordNet is created considering the idea of Princeton Wordnet (more in Miller et al 1990), where words are gathered to concepts according to their meanings. These concepts are connected with different semantic relations like hyponymy-hypernymy, antonymy, meronymy etc. Considering the EWN ideas and priorities, more attention is paid to synonymy as a fundamental semantic relation. Based on this, it is possible to assemble different words into one single database unit and into hypernymy-hyponymy relation. There are 43 different semantic relations in EuroWordnet. In EstWN there are now (February 2010) about 28 000 concepts. Estonian WordNet's basic idea is the creation of theoretically systematic and applicably proper network of meanings because it is useful for some natural language applications.

[1] See for detailed bibliography http://www.globalwordnet.org/gwa/gwa_conferences.htm

[2] Also Estonian WordNet, EstWN, see http://www.cl.ut.ee/ressursid/teksaurus/

[3] There are currently around 50 wordnets to different languages in the world (see more http://www.globalwordnet.org/).

3. Current semantic relations in Estonian WordNet

Our chosen approach so far for enlarging our thesaurus has been manual and domain-specific, i.e we have added semantic fields like architecture, transportation, personality traits and so on. There are 45 different types of semantic relations present (see also Table 1). The most frequent relation among nouns and verbs is hypernymy/hyponymy; near_synonymy and near_antonymy are more frequent among adverbs and adjectives.

Relation	Amount
has_hypernym/ has_hyponym	20069
near_synonym	5298
near_antonym	2054
Antonym	1088
Fuzzynym	756
has_mero_part/has_holo_part	571
has_instance/belongs_to_class	474
state_of/be_in_state	358
is_caused_by/causes	346
role/involved	333

Table 1. Amount of frequent semantic relations used for almost 28 thousands concepts.

Since one person is dealing with one domain at the time then it makes the relations between different concepts (in one domain) easier to determine. For example from the domain of architecture the concept *antiiktempel* ('antique tempel') has 1 hypernym, 11 hyponyms, 1 has_holo_part and 8 has_mero_part relations. For the most part the specific domains in EstWN are covered with many types of different semantic relations.

3.1. Semantic relations of adjectives

The most thoroughly examined domain in EstWN is the adjectives of personality traits. This specific domain includes in Estonian around 1200 words or expressions, which accordingly form around 400 synsets in EstWN. Work with this field showed us the problems which are connected with adjectives. One of the main problems concerns also (as other word-classes do) semantic relations. In analyzing Estonian, one of the main problems that arises is that the number of compounds is indefinite. It is easy for a speaker of Estonian to create new compounds (also adjectives) that are not listed in any dictionary but are, nevertheless, easily understood. In compounds, the first element functions as an attribute and the second as the head. For example: *arenemis**võimeline*** 'develop-capable', *otsustus**võimeline*** 'decision-capable', *õpi**võimeline*** 'learn-capable', *armastus**võimeline*** 'love-capable', *vastutus**võimeline*** 'responsibility-capable', etc. Does it mean that the second element of these words, i.e., – ***võimeline*** 'capable', acts as a hypernym in Estonian? The answer is no, because in these cases it is only via the first element of the words that the conceptually correct lexical relation is expressed. Thus, for example *armastus**võimeline*** 'love-capable' is synonymous with *hooliv* 'regardful', and *vastutus**võimeline*** 'responsibility-capable' is antonymous to *vastutustundetu* 'irresponsible', etc.

At the same time the examples below show that some concepts can act as hyponyms for others where the situation is precisely the opposite. Consider the following examples:

kade'envious' -> hyponym is *armu**kade*** 'jealous',
ahne 'greedy' > hyponym is *võimu**ahne*** 'power-greedy', hyponym is *raha**ahne*** 'money-greedy/money-grubber', hyponym is *kasu**ahne*** 'benefit-greedy'***;***
lahke 'kind' -> hyponym is *külalis**lahke*** 'guest-kind/hospitable', etc.

The examples above point to the fact that only some part of character terms in Estonian is hierarchical and show that even the lexical relations themselves can be language- or culture-specific.

Also, the semantic analysis of adjectives needs much broader and much more language-specific relations. On of the examples is the necessity to relate some particular concrete and abstract objects to the property characterizing them (SHAPE –> *round, square*; TEMPERATURE –>*hot, cold*; PERSONAL TRAIT –> *tender, kind).* This kind of approach has used already for dealing some other languages, for example German language in GermaNet (http://www.sfs.uni-tuebingen.de/GermaNet/) and Russian language in RussNet (Azarova & Yarovskaya, 2010).

3.2. Semantic relations of adverbs

Adverbs as a word class are morphologically and semantically very complex and their division in Estonian language is not very strict. It is possible though to distinguish adverbs by their meanings to four different groups: adverbs of place which denote the spatial relations; adverbs of time which denote time relations that characterize the events; adverbs of manner which denote the manner of events or the state or the position of the participant in an event; adverbs of degree which denote the amount of objects or extent of property. (EKK, 2000)

EstWN in present includes all different groups of meaning. While defining meanings we have followed the previously mentioned division, as the result of what the definition contains also the hint to according group of meaning. Since the Handbook of Estonian Language says that the scale of meanings of adverbs is quite broad, then many of adverbs tend to appear in multiple senses (for example the meanings of Estonian adverb '*vee*l' (English '*yet', 'still', 'more'*) can show time (*I'm still at home*) or quantity (*I'll pour some more water*) and also even modality (adverbs with empty meaning). In order to make the definitions of adverbs more clearer there should be created a semantic relation between adverbs which points to a different group of meaning. This semantic relation can be named as a so called base-category which has four possibilities: category of space, category of time, category of quantity and category of state. So, it should be possible to use semantic relations to indicate the base-category

(time-category, space-category, quantity-category, state-category etc) in thesaurus or/and to find out the semantic base what unites different usages of these words. The semantic relation of base-category supports the definitions of adverb's different senses.

Among adverbs there are words which from the viewpoint of nowadays Estonian do not have any morphological word parts except the stem, for example *kohe* ('now'), *otse* ('straight'), *veel* ('yet') etc (Kasik, 2009). In Estonian and generally in Finno-Ugric languages the derivation is quite notable while creating lexis (Kasik, 2009). In Estonian there are plenty of adverbs which are derived from other word classes, especially from adjectives, for example *ahne* > *ahne/lt* ('greedy'>'greedily'), but also from substantives (liik > liigi/ti ('sort'>'by sort') and from verbs (ärka/ma > ärk/vel 'to wake' > 'awake'). Also EstWN contains mostly adverbs which are derived from adjectives. The suffix changes the word's meaning in way that creates a completely new independent word (Kasik, 2009) and most of the derived adverbs belong to adverbs of manner, often denote derivations place and state and more rarely they denote place or time.

Considering the structure of wordnet it is possible to determine semantic relations (also derivational relation) between the same word class and also across word classes. In this paper we will firstly address the semantic relations between adverbs.

Adverbs of manner and state can occur as sets similar to declinable word's cases of place, for example *välja-väljas-väljast* ('out'-'outside'-'from outside'); *kaugele-kaugel-kaugelt* ('to far'-'far'-'from far'), but the meaning of these different forms does not change. It was decided that these sets can be arranged into hypernym/hyponym relation where inessive and adessive forms act as hypernyms. For example, *väljas* 'outside (inessive)' -> hyponym is *välja* 'out (illative)', hyponym is *väljast* 'from outside (elative)'; *kaugel* 'far (adessive)' -> hyponym is *kaugele* 'to far (allative)'; hyponym is *kaugelt* 'from far (ablative)'. Also, some adverbs can form sets of comparative, for example *hästi-paremini-kõige paremini* ('well'-'better'-'best'), *kaugel-kaugemal-kõige kaugemal* ('far'-'farther'-'farthest'). Basically these are different forms of the same word and we can also determine the base form as a hypernym. For example, *hästi* 'well (base form)', hyponym is *paremini* 'better (comparative)'; hyponym is kõige *paremini* ' (superlative)'.

Next we highlight the obligatory polaritive particle *-gi*/*-ki* (gi-suffix), which does not directly belong into the category of derivational suffixes. One of the meanings of gi-suffix is intensifying another meaning indicates polarity (Paldre, 1998), so the gi-suffix complements the word or the sentence with a shift of meaning. For example, adverbs *veel* ('more') and *veel/gi* can't be in the same synset because they are not synonymous and can't be explained with one and the same definition. As a solution we created a derivational relation between adverbs and now the two synsets are differently defined but still connected. For example, *vee/lgi* 'increasingly, progressively, more and more' -> is_derived_from *veel* 'else, further, other, more, in addition'.

Since adverbs are often derived from adjectives then the derivational relations between adjectives and adverbs are also important to determine. Deriving adverbs from adjectives is considered as a syntactic derivation (Kasik, 2009) which means that by adding a derivational suffix only the word class changes (from adjective to adverb), for example adverb *abitu*/*lt* 'helplessly, impotently, unable to help' -> is_derived_from adjective *abitu* 'helpless'.

One of the most productive and most frequent in Estonian is the lt-suffix. According to lexical semantics the adverbs derived with lt-suffix are considered adverbs of manner and the meaning of adjective carries over to adverb. (Kasik, 2009) This makes it possible then to derive adverbs with lt-suffix automatically since in EstWN there are more adjectives present. Also it is possible to carry over all the semantic relations already present with an adjective and also to carry over the definition of adjective. For example, adjective *aeglane* 'slow' -> xpos_near_synonym is adjective *aeglus* 'slowness'; state_of is noun *kiirus* 'speed, swiftness'; has_derived adverb *aeglase/lt* 'slowly'. In some cases semantic relations carried over from the adjective need to be corrected or added. For example, adverb *aeglaselt* 'slowly' -> near_synonym is adverb *pikkamööda* 'leisurely'; near_synonym is adverb *raskesti* 'ponderously'; near_synonym is adverb *loiult* 'languidly'; antonym is adverb *käbedasti* 'hurriedly, hastily, in haste, hotfoot'.

3.3. Semantic relations of Nouns

In our thesaurus we have mostly added nouns. The main semantic relation is the hypo-/hypernym relation, but work with specific domains have shown to us that very important is the mero-/holonym relation as well. For example in the vocabulary of building materials several mixtures (like cement and so on) involve the same substance. So the question here is – do we have to add it to every synset? Especially it is interesting when the substance or something have obtained via different chemical processes.

For nouns, the so called relation of association needs to be created (for example, how the nouns 'land' and 'land-tax' are associated). There are many examples, because the semantic marginal relations are not studied, as said before.

Since a PhD thesis about the systematic polysemy of nouns in Estonian was completed (Langemets, 2009), it is now possible to check the representation of systematic polysemy in EstWN thoroughly. The systematic pattern encoding enables to systemize and unify the representation of systematic polysemy, also it enables to relate logical and regular relations between word senses (Langemets, 2009). In EstWN the systematic polysemy is not marked explicitly, also it has been marked quite arbitrarily, for example 'school' has both BUILDING-INTITUTION senses, 'theatre' is only in a

INSTITUTION sense, 'university' in only in a INSTITUTION sense etc. In this thesis around 80 different types of systematic polysemy patterns are presented. It is proposed that systematic polysemy could be marked with a special kind of a polysemy relation, but this possibility needs more investigations.

4. Conclusion

Semantic relations are important in various NLP applications and therefore it is important to provide EstWN with rich and systematic network of different types of relations. While increasing EstWN with new concepts the need for more specific relations became clear.

It is necessary that semantic relations of adjectives need to be examined more carefully and we should seek for ways to use more hierarchical semantic relations. Also there is a need for more specific connection between adjectives and a corresponding semantic field to relate some particular objects to the property characterizing them. Same problems need to be solved among adverbs.

As for adverbs, the relation of base-category between adverbs is useful in order to indicate the group of particular meaning. It is possible also in some cases to determine hypernym/hyponym and derivational relation between adverbs. Since adverbs are often derived from adjectives then one of the important semantic relations to use is the derivational relation between adjectives and adverbs.

A systematic list of systematic polysemy patterns could be useful in automatic semantic analysis of Estonian, also for example in information extraction, where underspecification of senses is considered.

4. References

Apresjan**,** J. (2000). Systematic Lexicography. Translation, Oxford University Press.

Azarova, I., Yarovskaya M. (2010). Hierarchy of Perceptional Adjectives in RussNet. In Principles, Construction and Application of Multilingual Wordnets. Ed. P. Bhattacharyya, Ch. Fellbaum, P. Vossen. Proceeding of the 5th Global Wordnet Conference. Narosa Publishing House, India.

Beckwith, R., Fellbaum, Ch., Gross, D., Miller, G. (1990). WordNet: A Lexical Database Organized on Psycholinguistic Principles. - *Using On-line Resources to Build a Lexicon*. Toim. Zernik, U. Hillsdale, NJ: Erlbaum, ptk 9, 211-231.

Cruse, D.A. (1986). Lexical semantics. Cambridge University Press.

Cruse, D. A. (2000). Meaning in language: an introduction to semantics and pragmatics. Oxford. Oxford University Press.

EKG=Eesti Keele Grammatika. Erelt, M., Kasik R., Metslang H., Rajandi H., Ross K., Saari H., Tael K., Vare S. (1993). *Eesti keele grammatika* [Grammar of Estonian language]. Tallinn: Keele ja Kirjanduse Instituut. EKK 2000= Eesti keele käsiraamat. Tallinn: Eesti Keele Sihtasutus.

Fellbaum, Ch., Gross, D., Miller, K. J. (1990) (revised 1993). Adjectives in WordNet – International Journal of Lexicography 3 (4).

Gross D., Miller K.J. (1990). Adjectives in WordNet. *International Journal of Lexicography*, 3, 265-277.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford University Press.

Kasik, R. (2009). Eesti keele sõnatuletus (3. trükk). Tartu: Tartu Ülikooli Kirjastus.

Langemets, M. (2009). *Systematic Polysemy of Nouns in Estonian and its Lexicographic treatment in Estonian Language*. Phd Thesis. University of Tallinn.

Miller, G., Beckwith, R., Fellbaum, Ch., Gross, D., Miller, K.J. (1990). Introduction to WordNet: An On-line Lexical database. - *International Journal of Lexicography* 3, 235–312.

Miller, K. J. (1998). Modifiers in WordNet. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Massachusetts 47–67.

Orav, H. (1998). Eesti keele direktiivverbide semantilise välja struktuur tesaurusena. Tartu Ülikool.

Orav, H. (2006). Isiksuseomaduste sõnavara semantika eesti keeles. Dissertationes Linguisticae Universitatis Tartuensis. Tartu.

Paldre L. (1998). Eitustundlikud üksused eesti keeles. Magistritöö üldkeeleteaduse alal. Tartu Ülikool, Filosoofiateaduskond, eesti filoloogia osakond.

Parm, S. (2007). Partiklite veel, juba, alles, jälle tähendused eesti kirjakeeles. Magistritöö. Tartu Ülikool.

Proceedings of Global WordNet Conference: 2002, 2004, 2006, 2008, 2010 – see more bibliography: http://www.globalwordnet.org/gwa/gwa_conferences.htm

Saareste, A. (1958—1968). *Eesti keele mõisteline sõnaraamat I-IV. Dictionnaire analogique de la Estonienne I—IV*. Kirjastus Vaba Eesti, Stockholm.

Vanhatalo, U. (2005). Kyselytestit synonymian selvittämisessä. Sanastotietoutta kielenpuhujilta sähköiseen sanakirjaan. Suomen Kirjallisuuden Seura.

Vider, K. (1999). Sagedamad eesti verbid semantilises andmebaasis. Magistritöö. Tartu Ülikool.

Vogel**,** A. **(**2004**).** Swedish Dimensional Adjectives. Acta Universitatis Stockholmiensis. Stockholm : Almqvist & Wiksell International.

Vossen, P. (1998). Introduction to EuroWordNet. - *Computers and the Humanities*, 32 (2-3), 73--89.

Õim, A. (1991). Sünonüümisõnastik. Tallinn: Oma kulu ja kirjadega.

Õim, A. (1995). Antonüümisõnastik. Tallinn.

Consistency of Sense Relations in a Lexicographic Context

Peter Meyer, Carolin Müller-Spitzer
Institut für Deutsche Sprache (IDS)
Mannheim, Germany
E-mail: meyer@ids-mannheim.de, mueller-spitzer@ids-mannheim.de

Abstract

The representation of semantic relations between word senses of different entries in a dictionary is subject to a number of consistency requirements. This paper discusses the issue of maintaining and accessing consistent information on cross-references between sense-related items in electronic dictionaries from a mainly text-technological point of view. We present a number of consistency criteria for cross-referencing related senses and propose a practical approach to handling sense relations in an online dictionary. Our proposal is currently being tested in a large ongoing online dictionary project for German called *elexiko*. We focus on three different aspects of the dictionary development and editing process where consistency is an important issue: lexicographic data modelling, implementation of a lexicographic database system for an electronic dictionary, and development of practical tools for the lexicographer's workbench.

1. Introduction

Semantic relations between lexicographic items, such as synonymy and hyponymy between specific senses of different lexemes, are typically encoded as cross-references in the respective entries in a dictionary.

The necessity of keeping the reference structure of a dictionary *consistent* raises a number of conceptual and practical issues. In the context of describing lexical-semantic relations in dictionaries, consistency may require that, among other things, bidirectional relations, as existing in paradigmatic sense relations, are given for both reference points between which a specific relation holds. For example, if *require* is given as a synonym in the entry *demand*, then *demand* should also be listed as a synonym in the entry *require*. This is a form of consistency that is important for the underlying lexicographic data model as well as for the dictionary user.

As a matter of fact, however, consistency in bidirectional references is rarely met. In Figure 1, three entries taken from Duden: "Das Synonymwörterbuch" (2007), a conventional German dictionary of synonyms, are shown: *arbeitsunfähig* (unfit or unable to work), *dienstunfähig* (disabled, unfit for service), and *erwerbsunfähig* (unable to work, incapacitated). The meaning descriptions of these three entries are semantically very close. The terms constitute a set or cluster of synonyms. Nevertheless, there are striking inconsistencies. For example, in the entry *arbeitsunfähig*, the synonym *erwerbsunfähig* is missing although *arbeitsunfähig* is given as a synonym of the head word *erwerbsunfähig*. In addition, *dienstunfähig* is not listed as a meaning equivalent to *arbeitsunfähig*, whereas in the entry *dienstunfähig*, both *arbeitsunfähig* and *erwerbsunfähig* are listed as synonyms (cf. Müller-Spitzer 2010).

arbeitsunfähig
bettlägerig, krank, unpässlich; (*bildungsspr.*): indisponiert; (*oft emotional*): malade.

dienstunfähig
arbeitsunfähig, erwerbsunfähig, invalide, krank, nicht arbeitsfähig, nicht dienstfähig, nicht einsatzfähig, untauglich.

erwerbsunfähig
arbeitsunfähig, behindert, dienstunfähig, invalide; (*Amtsspr.*): schwerbehindert, schwerbeschädigt.

Figure 1: Entries *arbeitsunfähig*, *dienstunfähig*, and *erwerbsunfähig* from Duden: "Das Synonymwörterbuch" (2007).

It could be argued that consistency is not of particular importance here. Presumably most lexicographers attempting to compile a reference dictionary of synonyms chiefly aim to provide an abundance of words with similar meanings that can be substituted for each other: Their intention is not to depict theoretical lexical-semantic structures as lexicographic information, cf. also (Lew, 2007). However, it is argued here that, as the entry *arbeitsunfähig* in particular illustrates, a more consistent approach would help to provide the dictionary user with better information. Presumably any lexicographer would have added *erwerbsunfähig* as a synonym of *arbeitsunfähig* to this dictionary, if the incomplete listing had been noticed.

More generally, consistency of cross-references means that, depending on the overall design and purpose of the dictionary, its reference structure should reflect certain formal properties of the underlying lexical and semantic structure. A simple example of such a property is a symmetry constraint on synonymy: If word sense *S1* of lexeme *L1* is synonymous with word sense *S2* of lexeme *L2*, then, trivially, *S2* is also synonymous with *S1*. This

implies, as we have seen above, a possible corresponding requirement on the cross-reference structure of a dictionary: In many lexicographic contexts, if the section on *S1* in the entry for *L1* contains a synonymy reference to the section on *S2* in the entry for *L2*, then there should be a corresponding reverse reference in the *L2* entry. Similarly, if *S1* stands in a hyponymy relation to *S2*, then *S2* is a hypernym of *S1*. In this case, however, enforcing the corresponding possible requirement on reference structure is not feasible in conventional print dictionaries since this would imply that each and every hyponym of a lexeme must be included in its entry. But, as already noted above, not even the symmetry of the synonymy relation is usually enforced in standard dictionaries, cf. also (Müller-Spitzer, 2007).

Compared to print dictionaries, users of electronic dictionaries are much more likely to be confused by missing reverse links for a synonymy reference to another article because following links to sense-related items in an electronic dictionary is faster and more straightforward than looking them up by leafing through a printed dictionary. If a synonym is given for a specific sense in an entry and in the link-targeted entry this headword is not mentioned as a synonym, users are probably surprised by the lack of reverse linking. Here, a formal inconsistency at the level of data modelling easily leads to an inconsistency (in a less formal sense of the word) on the level of presentation and, hence, in user experience. Moreover, keeping track of all semantic relations represented in a lexicographic database is an elementary and essential prerequisite for lexicographic work on an electronic dictionary. It would be very useful if lexicographers were automatically informed that the entry is already mentioned as a target in another entry when they start to write a dictionary entry. Protecting dictionary authors from producing inconsistencies this way calls for extensive computer assistance, particularly when large amounts of data are involved.

On a terminological note, we will say that in both the synonymy and the hyponymy case two *unidirectional* references may stand in a *reverse relation* to each other and then together form a *bidirectional* reference. Provided that the unidirectional components of a bidirectional reference are stored in separate places, they must *correspond* to each other in that they (a) encode reverse semantic relations and (b) the target item of one unidirectional reference is the source item of the other and vice versa. This will be called the *correspondence requirement* for bidirectional links. Obviously, this is a different kind of consistency since the correspondence requirement for an actually bidirectional synonymy reference must be satisfied regardless of the question whether *all* synonymy references should be bidirectional.

2. XML modelling of Sense-Relation References: The Case of *elexiko*

We will discuss conceptual and implementational aspects of maintaining and controlling referential consistency in a concrete case, namely, the German corpus-based monolingual online dictionary *elexiko* that is accessible free of charge under *www.elexiko.de* and forms part of a long-standing and ongoing research project of the Institut für Deutsche Sprache (Institute for the German language), cf. (Haß, 2005), (Klosa et al., 2006). *elexiko* is still in progress (elexiko, 2003 seqq.); thus, this dictionary is not a complete reference book following an alphabetical compiling procedure.[1]

The lexicographic data pertaining to each *elexiko* entry are realised as a single XML document. All documents conform to a highly granular structural layout as defined in a complex XML Document Type Definition (DTD). The structural layout is strictly based on lexicographic content; any presentational aspects, such as typographic details, are taken care of by XSL transformations that generate HTML documents from the XML data.

In order to demonstrate the internal makeup of *elexiko* entry documents, we present a fragment of a typical XML representation. To ease comprehension, we will not use the original element names used for *elexiko* documents, but some hopefully self-explanatory English equivalents. The XML structure presented here is slightly simplified where this does not affect the topic under discussion. Boldface type is used to indicate data that is used to uniquely specify a particular reference to a sense-related item.

```
<elexiko-article id="1234">
 <general>
  <lemma-sign>Familie</lemma-sign>
 </general>
 <sense id="relatives">
  <usage>
   <paraphrase>
    Mit Familie bezeichnet man eine Gruppe von
Personen, die durch Geburt oder durch Heirat
miteinander verwandt sind. In engerem Sinn
bezieht sich der Sprecher mit Familie auf eine
Lebensgemeinschaft, die aus Eltern und
Kindern besteht, in weiterem Sinn auch auf
eine Gemeinschaft, die mehrere Generationen
umfasst und zu der z. B. die Großeltern, die
Geschwister der Eltern und Großeltern ein-
schließlich deren Angehörige usw. gezählt
werden.
   </paraphrase>
   <paradigmatic-relations>
    <partonymy>
     <item  articleID="9999"
            senseID="female descendant"
            subsenseID="0">
       Tochter
```

[1] In this paper, we will not discuss the linguistic and lexicographic foundations for the kind of XML modelling and for the treatment of sense relations in *elexiko*. However, there is ample literature that relates *elexiko* to other approaches in electronic lexicography, cf. Storjohann 2009 and 2010 and www.owid.de/elexiko_/pgProjektveroeffentlichungen.html resp. http://www.owid.de/elexiko_/pgVortraege.html for up-to-date references.

```
      </item>
      <item  articleID="3737"
             senseID="mother and father"
             subsenseID="0">
        Eltern
      </item>
     </partonymy>
    </paradigmatic-relations>
    <subsense id="dynasty">
     <subsense-paraphrase>
 Mit Familie bezeichnet man eine angesehene,
wohlhabende, einflussreiche bzw. adlige
Personengruppe, deren Mitglieder durch Geburt
oder Heirat miteinander verwandt sind.
     </subsense-paraphrase>
     <paradigmatic-relations>
      <synonymy>
       <item articleID="5678"
             senseID="dynasty"
             subsenseID="0">
        Haus
       </item>
       <item articleID="1066"
             senseID="dynasty"
             subsenseID="0">
        Dynastie
       </item>
      </synonymy>
     </paradigmatic-relations>
    </subsense>
   </usage>
  </sense>
  <sense>
  </sense id="biological taxon">
   …
</elexiko-article>
```

The root element of each entry document has an attribute @id, its *article ID* – a string representation of an integer number uniquely identifying the entry. It contains one <general> element with sense-independent information (relating to, e.g., orthography and morphology) and arbitrarily many <sense> elements representing different word senses. No distinction is made between polysemy and homonymy.

The lemma sign [for terminology cf. (Hausmann & Wiegand, 1989)] is part of the general, that is, sense-independent information in the article as specified within the <general> element. In our sample entry with an article ID of "1234", the German equivalent to 'family' has the citation form (nominative singular) *Familie*.

Each word sense is represented by a <sense> element with an attribute (a *sense ID*) that identifies this sense uniquely within the article. The most salient word sense of *Familie* might be paraphrased as 'group of close relatives of a person'. Using English IDs for the purpose of this article, we might choose "relatives" as the ID. The ID is not supposed to be a concise hint at the semantics of a sense; it just serves as a convenient mnemonic. In the XML document, a short explanation of the contexts associated with the word sense "relatives" is stored in a <paraphrase> element. For illustration purposes, a second word sense of *Familie* used in biology is shown in the XML fragment above.

The word sense with the ID "relatives" is assumed to have a specialized *subsense* in German, namely, 'group of relatives who play an important role in society'. This subsense appears nested inside the appropriate <sense> element as a <subsense> element with a *subsense ID* attribute "dynasty".

Figure 2 shows a partial view of the *elexiko* entry on *Familie* as it is presented to the user in a web browser. The sense and subsense IDs – here in their original German appearance, for example, "Verwandte" for "relatives" and "Dynastie" for "dynasty" – serve as headings for the different senses and subsenses. In this particular view, the meaning explanations as stored in the <(sub-sense-)paraphrase> elements are given.

Figure 2. A part of the HTML-based online presentation of the entry *Familie* in *elexiko* (screenshot).

In *elexiko*, references to sense-related items, henceforth *paradigmatic references*, always relate specific word senses of two entries. The "dynasty" subsense of *Familie* contains a synonymy reference to the corresponding sense of the entry *Haus* ('house'). The type of sense relation (named *paradigmatic relation type* here) is encoded as a <synonymy> element inside the <subsense> element. A word (sub)sense may have more than one synonym, so each synonymous word sense is to be given as a separate <item> element enclosed by <synonymy>. Our sample fragment lists another synonym for the same word sense, namely, *Dynastie*. In addition, two partonyms for the "relatives" sense of Familie are given, namely: *Tochter* ('daughter') and *Eltern* ('parents'). The attributes and text content of each <item> element provide a complete specification of the end point or *target address* of the reference, that is, the lemma sign and article ID of the entry *Haus* as well as the sense and subsense IDs of the word sense referred to. If the target address concerns a word sense but not a subsense, "0" is used as subsense ID.

To sum up, three observations may be made at this point. First, in *elexiko*, three strings are used to uniquely identify the target address in a reference to a related item, namely,

- the ID of the target entry as a whole;

- the ID of the target sense; and, where applicable,
- the ID of the target subsense.

Second, all lexicographic information on sense relations targetting (sub)senses of other dictionary articles is stored in the individual entries' subsection (XML element) pertaining to the source address. Specifically, the information on the three source IDs as well as on the type of sense relation is distributed among different nested ancestor elements of the element containing the target specification.
Third, outgoing references are stored in a strictly local fashion, that is, only in the source article. There is no indication in the source XML document as to whether a consistent reverse reference exists in the corresponding place in the target article.
In order to obtain all necessary information on a particular reference, the XML document containing the reference must be parsed, which is an expensive operation in terms of database operations. Ensuring consistency of cross-reference information in one document *D* inevitably requires parsing all documents referred to in *D* as well as all documents referring to *D*.

3. Criteria for Consistency in Cross-Referencing Sense-Related Items

From a lexicographer's point of view, there are many different ways in which a cross-reference might fail or be inconsistent. This section enumerates important criteria for evaluating the consistency of paradigmatic references in *elexiko*. As was said before, we concentrate on aspects of the data structure, not on general lexicological-lexicographic considerations.
In a well-formed and valid XML document representing an entry in *elexiko*, any paradigmatic reference must meet the following requirements:

- It must be **complete**: All necessary pieces of information related to a particular type of reference must be present. To a certain degree, completeness can be enforced through an appropriately specified DTD or XML schema. However, a common and unavoidable problem in the process of compiling a dictionary is the need to make preliminary incomplete references to target addresses that do not exist yet. In *elexiko*, paradigmatic references to a word sense in an entry not yet edited use the dummy word sense ID "0".
- It must be **well-formed**: All required pieces of information must conform to formal specifications. Again, certain aspects of well-formedness cannot be captured by means of an XML schema, for example, conventions regarding allowed formats for different ID types.
- It must be **valid**, that is, it must point to an address that really exists in the lexicographic database. Note that validity presupposes both well-formedness and completeness of the reference. A particular prerequisite for validity is factual **consistency**: Different parts of a paradigmatic reference must not contradict each other. For instance in *elexiko*, a target address contains both the ID of the target entry and its lemma sign. Of course, the lemma sign specified in the target entry with that ID itself must be identical to the one given in the reference.

Let us call a reference that fulfills all of the above criteria a *correctly specified* unidirectional reference. Correctly specified references might still be lexicographically *inadequate* in relating wrong addresses or picking the wrong paradigmatic relation. Lexicographic adequacy cannot be checked by an automated procedure and constitutes yet another, very important kind of consistency requirement.
For a given unidirectional reference *R*, additional conditions are needed in order to define whether another unidirectional reference *R'* counts as a *potential reverse reference*, such that *R* and *R'* together form a bidirectional reference. The necessary requirements may be stated as follows:

- The type of paradigmatic relation must be a candidate for a bidirectional reference.
- Both *R* and *R'* must be correctly specified. Special provision must be made for the case that one or both of the references are specified correctly only on the dictionary entry level, but are not (yet) *complete* on the word sense level. This situation typically arises when the target article has not been edited yet.
- The correspondence requirement stated above must be met. If one of the references is not yet complete, this requirement must be relaxed to state that the target address of the incomplete reference must either be identical to the source address of the reverse reference or refers to a larger part of the entry that contains this source address.

If there are no potential reverse references for a given paradigmatic reference *R*, this might count as an instance of inconsistency in case bidirectionality is specified to be compulsory for the given paradigmatic relation. For example, *elexiko* employs a very narrow lexicographic concept of synonymy for which compulsory reciprocity is indeed a sound requirement.[2] If potential reverse references can be found in the lexicographic database, different cases may be distinguished according to which of these references are completely specified and whether there is more than one candidate reverse relation in the database. In case both unidirectional references *R* and *R'* are correctly specified and fulfill the correspondence requirement, we may classify the resulting bidirectional reference as correctly specified. Again, a correctly specified bidirectional reference might still be lexicographically inadequate.
This brief overview should suffice to demonstrate some of the intricacies of managing consistency issues in dictionaries. These problems must be dealt with at several

[2] In this paper, we simply use synonymy as a typical example candidate for a symmetric sense relation. Actual decisions on how to model sense relations will depend on the lexicographic setting and are independent of the conceptual and implementational points of the paper; hence, our approach can just as easily be applied to other sense relations such as antonymy: *elexiko* distinguishes between five categories of antonymy several of which are candidates for compulsory reciprocity.

stages of the process of conceiving, implementing, and editing dictionaries. The following sections will examine some of these stages in turn and discuss the merits and pitfalls of possible solutions.

4. Making Dictionary Entries Consistent: Considerations on Data Modelling

At first glance, a conceptually clear and simple solution to inconsistency threats in a lexicographic database seems to commend itself: Detach all reference-related information from the entry documents and put it in a separate table. After all, such a table (which we will call a *reference table* for short) would be the standard solution for modelling many-to-many relationships in a relational database. Each row in a reference table corresponds to a unidirectional or bidirectional paradigmatic reference. The columns specify the paradigmatic relation type and the three ID strings of source and target adress. The relational table might just as well be represented in an XML format. A sample entry for a *unidirectional* paradigmatic reference could then roughly look like this (cf. Section 2):

```
<reference relation="synonymy">
  <srcLemmaSign>Familie</srcLemmaSign>
  <srcEntryID>1234</srcEntryID>
  <srcSenseID>relatives</srcSenseID>
  <srcSubSenseID>dynasty</trgSubSenseID>
  <trgLemmaSign>Haus</srcLemmaSign>
  <trgEntryID>5678</trgEntryID>
  <trgSenseID>dynasty</trgSenseID>
  <trgSubSenseID>0</trgSubSenseID>
</reference>
```

In a similar XML representation, compulsory *bidirectional* references can be coded in a redundancy-free way that compliance with the correspondence requirement is guaranteed:

```
<reference relation="synonymy">
  <entry>
    <lemmaSign>Familie</lemmaSign>
    <entryID>1234</entryID>
    <senseID>relatives</senseID>
    <subSenseID>dynasty</subSenseID>
  </entry>
  <entry>
    <entryID>5678</entryID>
    <senseID>dynasty</senseID>
    <subSenseID>0</subSenseID>
  </entry>
</reference>
```

In ontology-based systems, this approach might be a sensible choice for modelling sets of synonymous senses since consistency is enforced when each *set* of *n* word senses is indeed represented as a *set* of XML elements instead of a group of $n(n\text{-}1)$ separate unidirectional references. Still, non-overlap of different sets of synonyms cannot be enforced this way. Aside from that, all entry and sense IDs in a reference table entry must themselves be correctly specified. This constitutes yet another consistency problem.[3]

As soon as other kinds of sense relations have to be considered for the data model, such as paradigmatic relations that are only *potentially* bidirectional, the disadvantages of a separate reference table will, in most cases, outweigh the benefits.

To begin with, a serious drawback of a separate data model for reference-related information becomes apparent when *entry-specific* information on paradigmatic relations is to be provided. In *elexiko*, for instance, sense-related items belonging to a given word sense in an entry are ordered according to corpus salience and discourse relevance. In such situations, the individual entries would have to include references to locations in the reference table, which would mean replacing one consistency issue with another. This problem is an indication for a more general need to separate two concerns, that is, to provide lemma-specific and lexicographically relevant information on sense relations on the one hand and to infer or keep track of all existing sense-relations between dictionary items on the other.

Introducing a separate reference table considerably complicates the editing process for dictionary entries since two tables must be modified concurrently and kept in agreement. As a consequence, manually editing the XML representation of an article becomes virtually impossible because it is too confusing and error-prone. A separate software tool would be needed just to keep the two database tables in synch at any time and to present all relevant entry-related reference table information in a perspicuous way to the lexicographer. As a final point, deciding which *types* of cross-references to pull out into a reference table and which to leave in the entries can be a delicate decision that cannot easily be changed later.

Everything considered, we believe that in most cases a minor improvement in handling compulsory bidirectionality will not justify the numerous administrative and conceptual complications induced by the introduction of a separate reference table. As a consequence, we strongly favour a maximally parsimonious data model for electronic dictionaries that leaves all reference-related information strictly within the respective entries.

5. Handling References on the Implementation Level

For *elexiko*, the 'local' alternative outlined above has been opted for so that all unidirectional references are encoded solely within the respective entry documents and no separate data structure for bidirectional links is needed. This

[3] Partitioning all word senses of a language into equivalence classes presupposes transitivity of the synonymy relation. However, a range of philosophical, semantic, lexicological, and lexicographic arguments against the transitivity of synonymy have been advanced. Quine's insistence on the context-specific nature of synonymy springs to mind, cf. (Bosch, 1979) for a succinct overview. See (Storjohann, 2006) for a range of lexicological and lexicographic observations on synonymy that bear on this important issue.

means that, at least in principle, all management and information access tasks concerning (paradigmatic) references could be processed through queries on the XML representations of the dictionary entries. However, performance considerations regarding the underlying database system suggest a different strategy. As noted above, checking for inconsistencies in an entry's references would entail (a) searching the database for XML documents that contain certain information, that is, references to a given entry and (b) parsing these XML documents. Compared to a standard search operation in a relational database table, searching through hundreds of thousands of complexly structured XML documents is already a very expensive database operation, in terms of both time and CPU load, even if highly optimized indices (cf. Müller-Spitzer & Schneider, 2009) are used. Parsing the relevant XML documents is even more costly, no matter whether the parsing is done in the database system itself or on a client system.

As long as merely individual entries are checked for reference inconsistencies by a lexicographic tool (see next section), the necessary searching and parsing processes on the XML instances in the Oracle-based *elexiko* system take a few seconds at most. More demanding tasks such as the following ones are out of the question without a separate handling of reference information:

- searching all dictionary entries for inconsistent references, paradigmatic or other;
- processing complex queries requiring a recursive traversal of a possibly large number of referential links, such as for
    - visualising link trees starting from a given word sense;
    - finding minimal link paths between addresses;
- enabling end users of the dictionary to formulate and process complex queries on referential structure.

However, a simple and effective solution to the performance bottleneck of XML processing is available: One can simply *copy* all information pertinent to paradigmatic references to a separate relational database table. Afterwards, complex queries on cross-reference structure can be processed on this relational table using fast standard SQL queries. Initial construction of the additional table – which will be called *link table* in this paper only to distinguish it terminologically from a *reference table* as defined above – can be accomplished using a rather simple XQuery construct. This can be a time-consuming operation, but it needs to be done only once. Afterwards, the link table must automatically be updated each time an entry is altered, added, or deleted. To this end, a so-called trigger is installed in the database. The trigger starts a stored update procedure on the link table whenever the main table that contains the XML documents undergoes a change.

A link table may have exactly the same structure as a reference table. The difference to notice is that a link table does not contain any new information over and above the table of dictionary entries; it simply mirrors reference-related aspects of the dictionary entries. In other words, a link table is not part of the data model.

Even though the link table does not contain any information that is not already present in the XML instances, it offers several distinct advantages. It abstracts from the particularities of representing information in the XML format of the entries; specifically, as noted in Section 2, source and target of a reference are necessarily encoded in completely different ways within the entries while they can be represented in a simple and uniform format in the link table. Accessing the link table does not require parsing: it only requires standard relational database queries. Even if the information is represented as XML, modern database systems can transparently map it to an underlying relational representation, rewriting XPath expressions as SQL queries. In the Oracle database system used for *elexiko*, this is called "XML/SQL duality". Even though exact figures depend on a wide variety of factors, information extraction from an underlyingly relational link table may very well be 100 times faster than parsing dictionary entries. Oracle uses a dimension-less quantity named *cost* to measure the database system load for a query; and indeed, in terms of cost, looking up and parsing complex XML-based entries access might easily be more than 1000 times more expensive than a link table query.

Overall, modelling references in a strictly 'local' fashion as an integrated part of the pertinent source entry is an approach both theoretically sound and pragmatically viable. Database performance can be enhanced dramatically through the use of a relational link table that provides fast access to the reference structure. The solution is robust in that it does not necessitate additional software tools for the editing process or a refactoring of existing database resources. The question which cross-reference relations should be included in the link table does not amount to a vital decision that is difficult to change afterwards.

A further decision has to be made as to whether bidirectional links should be encoded as two different and independent unidirectional entries (table rows) in the link table or rather be handled in a separate and possibly less redundant way, for example, as shown in the second XML example of Section 4. However, there are reasons to prefer the more redundant representation. In a typical setting where the database system has to process large numbers of potentially complex user queries on cross-reference structure, the time penalty induced by having to look up one more table row for a consistency check hardly matters. On the other hand, editing links in the entries has more complicated reverberations for a link table with a separate storage format for bidirectional links. If, for instance, a newly added article contains a paradigmatic reference that is reverse to an already existing one, the latter has to be deleted from the table while a new bidirectional reference is added.

6. Aiding the Lexicographer: Tools for Safeguarding Consistency

The implementation aspects that we focused on in the previous section obviously have no immediate bearing on the consistency topic of this paper. However, a link table can form a vital part of an assistive IT environment for the working lexicographer whose virtual workbench might include a software tool to help him safeguard reference consistency. Such a reference management tool is currently under development for *elexiko*; this section will present some of its functionality in the light of the preceding remarks.

In what follows, let *D* be the XML document of the dictionary entry currently being edited. In the most basic case, work on the article is done in a generic XML editor. Without a reference management tool, editing references in *elexiko* looks as follows:

- The lexicographer inserts a sense-related item in the entry *D*.
- In the online version of *elexiko*, the lexicographer has to check which senses and subsenses constitute the correct reference target.
- The corresponding IDs of the reference target (lemma/sense/subsense) must be looked up in the *elexiko*-database and manually copied into the entry *D*.
- After completing the entry, the lexicographer has to check the consistency of sense-related items in D in correspondence to the ones in the target entries; this procedure has to be done in the online version.

A reference management tool as a separate application facilitates lexicographic work in a significant way: When entry *D* is opened in the XML editor, the tool enumerates all paradigmatic references in other articles to word senses in *D* (incoming references) as well as all paradigmatic references in *D* to other articles (outgoing references). For each incoming reference in the list, the management tool displays current status information regarding to what extent the consistency criteria given in Section 3 are met for unidirectional as well as bidirectional references. Where an incoming reference is not yet complete because the source article was compiled before editing *D* so that the appropriate target word sense IDs are missing, authors can update the source document with only a few mouse clicks just by choosing from a list of all the word senses in *D*.

In a similar vein, the management tool automatically checks whether all currently outgoing references are correctly specified. The lexicographer can select any of these references and let the program fill in missing details on the desired target word sense by simply choosing from a list. Additionally, the table of outgoing references can be used to speed up navigation within *D* in the editor. A sample screenshot of the management tool developed for *elexiko* is shown in Figure 3 (see below).

Apart from securing consistency with respect to references from and to individual dictionary entries, a reference management tool should also provide tools to scan an entire lexicographic database for

- inconsistent (incorrectly specified) references, in particular references pointing to inexistent entries or word senses within entries; or
- missing reverse references for unidirectional references of an obligatorily bidirectional type.

In *elexiko*, article editing is done in a standard XML editor with a Java API that is used by the reference management application for obtaining the current contents of the active document, navigating within the document, inserting data into it, and so on. On the other hand, the reference manager communicates with the Oracle database system using a standard JDBC interface. The management tool parses the active XML editor document in order to obtain a list of outgoing references. For incoming references, the link table of the database system is used.

7. Conclusion and Prospects

In this paper, we have presented a robust, conceptually parsimonious, and linguistically sound solution to handle cross-references between sense-related entries in an electronic dictionary. We have argued that in typical cases, modelling cross-references with separate data structures simply shifts the sources of possible inconsistencies to another place and merely introduces additional conceptual complexity. Therefore, we suggest to keep information on cross-references strictly local to the respective source entries. To enhance performance of database retrieval, information related to cross-references is additionally kept in a separate, relationally stored link table that is automatically updated whenever entries are altered or added. Taking advantage of such a table, reference management software can then continually screen for referential conflicts while a dictionary entry is being edited and easily check the overall referential consistency of a dictionary database.

Our approach is well suited to a setting where several independent dictionaries are to be gradually integrated into a global database environment with cross-dictionary references. It can easily be extended to other kinds of cross-references between and even within dictionary entries.

The task of visualising lexicographic reference structure is a lucid example of the practical use to which our approach can be put. Figures 4.1 and 4.2 are based on the output of a visualisation tool developed for *elexiko*. Figure 4.1 shows the paradigmatic relations given in the entry for the three word (sub)senses of the entry *Familie* as a directed graph. The program is able to traverse long chains of cross-references from one word sense to the next. In this way, graphs with several thousands nodes (word senses) can be constructed recursively. Calculating such huge graphs on the basis of parsing dictionary entries alone would hardly be feasible; with the use of a link table, it becomes a matter of seconds. In Figure 4.2, some incoming references for the word sense “relatives” are displayed with a recursion depth of 2.

Such a visualisation of paradigmatic structures may be useful for lexicographers for checking a longer chain of

paradigmatically associated entries as well as for navigational tasks provided for dictionary users.

To sum up, our proposal is founded on a fine-grained division of labour: On the one hand, lexicographic reference information that is specific and relevant to an individual entry is represented in the entry itself; on the other hand, further facts about sense relations, such as chains of ever more specific hyponyms of a word sense, can then be inferred efficiently through the use of a link table. This link table not only allows for fast and comfortable consistency checking routines but also for more flexible ways to make use of reference information in an electronic dictionary.

Figure 3. GUI of the reference management software for *elexiko*.

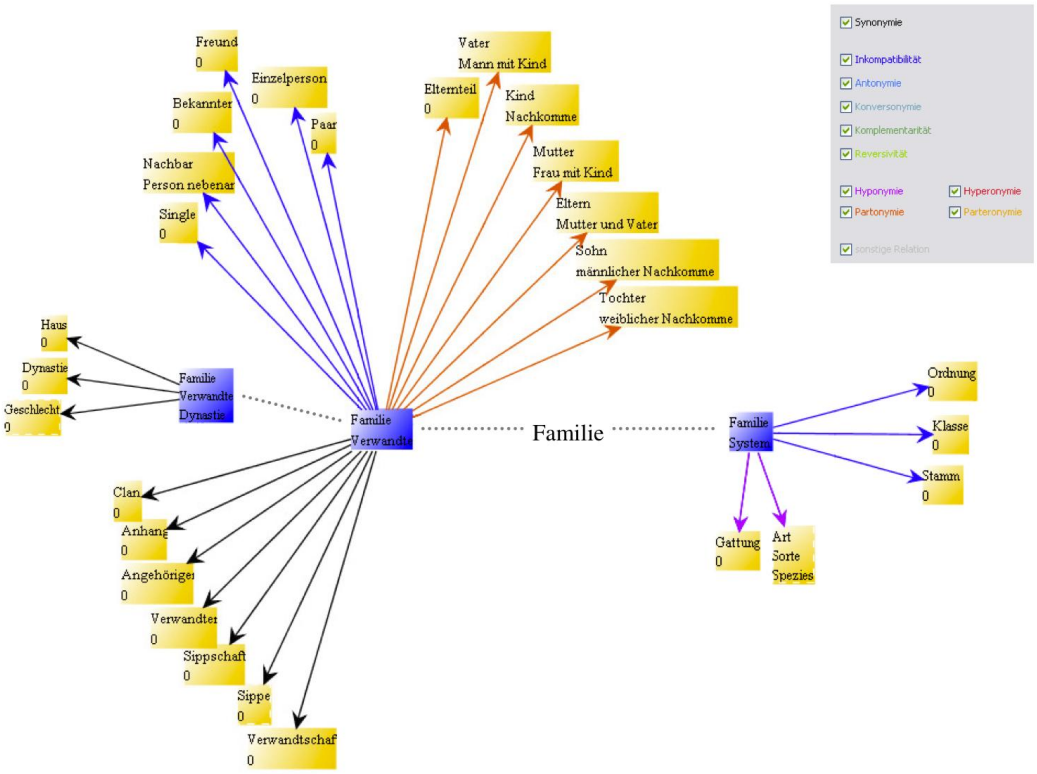


Figure 4.1. Visualization of outgoing references of the *elexiko* entry *Familie*, sense "Verwandte" ('relatives') (recursion depth of 1). The boxes represent word (sub)senses and indicate lemma sign, sense ID, and, where applicable, subsense ID. Arrows stand for unidirectional paradigmatic (sense) relations whose type is marked by colour.

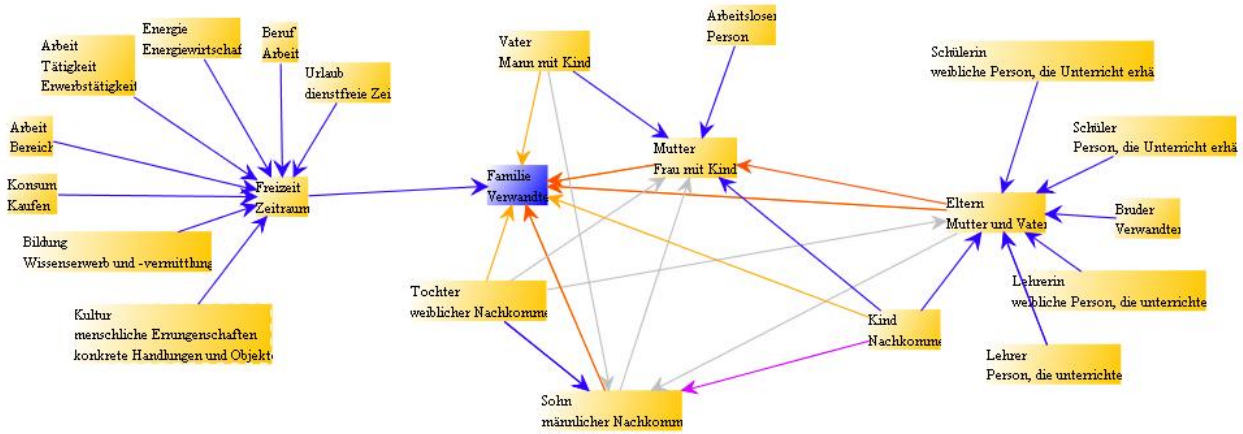


Figure 4.2. Visualization of incoming references of the *elexiko* entry *Familie*, sense "Verwandte" ('relatives') (recursion depth of 2). The boxes represent word (sub)senses and indicate lemma sign, sense ID, and, where applicable, subsense ID. Arrows stand for unidirectional paradigmatic (sense) relations whose type is marked by colour.

8. References

Bosch, P. (1979). Synonymie im Kontext. Ein Nachwort. In W.V.O. Quine, *Von einem logischen Standpunkt*. Berlin: Ullstein, pp. 161--172.

Duden (2007). *Das Synonymwörterbuch.* 4th edition, Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

elexiko (2003 seqq.). In *OWID – Online Wortschatz-Informationssystem Deutsch*, Mannheim: Institut für Deutsche Sprache, www.owid.de/elexiko_/index.html

Haß, U. (Ed.) (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Schriften des Instituts für Deutsche Sprache. Berlin, New York: de Gruyter.

Hausmann, F.J., Wiegand, H.E. (1989). Component Parts and Structures of General Monolingual Dictionaries: A Survey. In F.J. Hausmann et al. (Eds.), Wörterbücher / Dictionaries / Dictionnaires. An International Encyclopedia of Lexicography. Berlin, New York: de Gruyter, pp. 328--360.

Klosa, A., Schnörch, U., Storjohann, P. (2006). ELEXIKO - A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim. In E. Corino, C. Marello, C. Onesti (Eds.), *Atti del XII Congresso Internazionale di Lessicografia. Torino, 6-9 settembre 2006* (Proceedings of the 12th EURALEX International Congress). Vol. 1. Alessandria: Edizioni dell'Orso, pp. 425--430.

Lew, R. (2007). Linguistic semantics and lexicography: A troubled relationship. In M. Fabiszak (Ed.), *Language and Meaning. Cognitive and Functional Perspectives*. Frankfurt a.M.: Peter Lang, pp. 217--224.

Müller-Spitzer, C. (2007). Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. *Hermes* 38, pp. 137--171.

Müller-Spitzer, C. (2010). The Consistency of Sense-Related Items in Dictionaries. Current Status, Proposals for Modelling and Potential Applications in Lexicographic Practice. In P. Storjohann (Ed.), *Lexical-semantic relations from theoretical and practical perspectives*. Lingvisticæ Investigationes Supplementa. Amsterdam/New York: Benjamins (forthcoming).

Müller-Spitzer, C., Schneider, R. (2009). Ein XML-basiertes Datenbanksystem für digitale Wörterbücher – Ein Werkstattbericht aus dem Institut für Deutsche Sprache. *it-Information Technology* 51(4), pp. 197--206.

Storjohann, P. (2006). Kontextuelle Variabilität synonymer Relationen. OPAL – Online publizierte Arbeiten zur Linguistik 2006(1), Mannheim: Institut für Deutsche Sprache.

Storjohann, P. (2009). Plesionymy: A case of synonymy or contrast? *Journal of Pragmatics* 41(11), pp. 2140--2158.

Storjohann, P. (2010). Colligational patterns in a corpus and their lexicographic documentation. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference 2009 in Liverpool*. (published online under: http://ucrel.lancs.ac.uk/publications/CL2009/)

Wheels for the mind of the language producer: microscopes, macroscopes, semantic maps and a good compass

M. Zock
LIF- CNRS, UMR 6166,
Case 901 - 163 Avenue de Luminy
F-13288 Marseille
E-mail: michael.zock@lif.univ-mrs.fr

Abstract

Languages are not only *means of expression*, but also *vehicles of thought*, allowing us to *discover* new ideas (brainstorming) or *clarify* existing ones by refining, expanding, illustrating more or less well specified thoughts. Of course, all this must be learned, and to this end we need resources, tools and knowledge on how to use them. I will be mainly concerned here with the produtive mode, language generation in the mother tongue or a foreign language.

We all are familiar with microscopes, maps, and navigational tools which we normally associate with professions having little to do with NLP. I will argue here that this does not need to be so. Methaphorically speaking, our brains or computers use the very same tools both for comprehension or expression. I will illustrate this claim mainly for language generation. Particular emphasis will be given to global structures (patterns) and navigational tools. I will argue that patterns can be compared to *macroscopes*, accounting for important aspects of language production and language learning (fluency acquisition). The idea of the *lexical compass* is used to show how one could help people to navigate in a huge, multi-dimensional space, in order to find the word they are looking for.

1. Introduction

Languages are not only *means of expression*, but also *vehicles of thought*, allowing us to *discover* new ideas (brainstorming) or *clarify* existing ones by refining, expanding, illustrating more or less well specified thoughts. Of course, all this must be learned, and to this end we need resources, tools and knowledge on how to use them. I will be mainly concerned here with the produtive mode, language generation, be it in the mother tongue or a foreign language.

We all are familiar with microscopes, maps, and navigational tools which we normally associate with professions having little to do with NLP. I will argue in this paper that this does not need to be so. Methaphorically speaking, our brains or computers use the very same tools, regardless of the task (analysis vs. generation).

Paper- or electronic dictionaries are not only resources, but also *microscopes*, revealing details concerning a given word: meaning spelling, grammar, etc. They are particularly appreciated if one is looking for the meaning of a word (comprehension, section 1), or if one is hunting for an elusive word (production, section 4).

Macroscopes are tools to reveal the great picture. Even though badly needed, they are not yet available in hardware stores, but they do exist in some scientists' minds. They are known under the headings of pattern recognition, feature detectors, etc. The resulting abstractions, models, schemata or blueprints (frames, scripts, patterns) are useful for a great number of tasks. I will illustrate this point for patterns via two examples related to *language production* in real-time (section 2) and *foreign language learning* (section 3).

Semantic maps (ontologies, encyclopedias, thesauri, wordnets) are excellent tools for organizing knowledge and words in a huge multidimensional meaning space. Nevertheless, in order to be truly useful, i.e. to guarantee access to the stored and desired information, maps are insufficient — we also need some navigational tool(s). To illustrate this point I will present some of my ongoing work devoted to the building of a *lexical compass*. (section 4).

2. Microscopes

Words are pointers to information, that is, they are economic means to express and nutshell complex thoughts. Dictionaries are storehouses, containing information associated to words. This being so, they have the potential to function like microscopes: blow up and show in detail (hidden) information.

Next to powerful search mechanisms electronic dictionaries have nowadays various facilities to display information[1]. Hence, interfaced with a text editor they allow for active reading. Clicking on a word allows at least partially to display associated information: *translation*, *definition*, *usage* (in the current context), *grammatical information*, *spoken* form, etc. Figure 1 illustrates such an interface, displaying information concerning some japanese text given as input. Of course, the script conversion (transliteration from kana to latin characters) needs to be done by a dedicated component.

[1] A corpus query system like *Sketch Engine* (Kilgarriff, 2004) can reveal additionally, very precious information: a word's grammatical and collocational behaviour in texts.

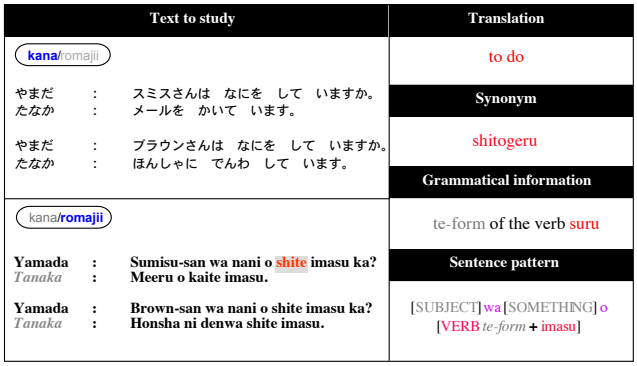


Figure 1. Interface revealing hidden information

Such kind of tools exist nowadays, at least for some of the mentionned features. They do exist both for western and oriental languages. For example, LiveDictionary works on a Mac in conjunction with Safari[2]. It allows to load various dictionaries (including WordNet) into your machine and to reveal the words' meanings. Hovering the mouse over a given word will yield its translation or other information associated with it. Elda[3] is a reading aid for people wishing to learn Italian or German. The following sites deal with oriental languages[4]. Obviously, all these tools are very precious for language users, no matter whether they are readers, writers, learners or experts.

3. Macroscopes

3.1 Sentence production in real time

Spontaneous speech is a cyclic process involving a loosely ordered set of tasks: conceptual preparation, formulation, articulation (Levelt, 1989). Given a goal one has to decide what to say (conceptualization) and how to say it (formulation), making sure that the chosen elements, words, can be integrated into a coherent whole (sentence frame) and do conform to the grammar rules of the language (syntax, morphology). During vocal delivery (articulation), in itself already a quite demanding task, the speaker may decide to initiate the next cycle, that is, start to plan the following ideational fragment.

It is clear, to produce language in real-time is quite a challenge. Nevertheless, despite individual differences nearly everyone seems to manage. The question is how is this done? Linguists describe languages in terms of rules, but people hardly ever learn such descriptions, leave alone apply all of them, at least not at the initial stages of acquiring a new language. What people do learn though are patterns complying with these rules. Of course, people do use rules, but in conjunction with patterns. In other words, we believe in processing at two levels or speeds[5]: the skeleton, i.e. global sentence structure is taken care of by patterns, while local accomodations (morphology) are based on rules.

If our hypothesis is correct, than we must show what these patterns looks like. Put differently, we must show how people manage to recognize potential syntactic structures on the basis of formal characteristics of the input (goal and conceptual structure). Our approach is based on the following two assumptions: (a) people do not process word by word or concept by concept, they rather operate on larger chunks (typically clauses); (b) skilled speakers know immediately which syntactic structure correspond well to a given conceptual structure (message). This means, that they have acquired some kind of competency of structure mapping. We will try to operationalize this kind of competency by spelling out some of the rules or formal characteristics. Since input is represented in terms of graphs or semantic networks, our rules will be expressed in graphical terms: *shape* of the nodes (rectangle vs. ovals), *direction* of the arcs (incoming vs. outgoing), *type of link* (case role vs. attribute), *type of argument* (oval, rectangle or a combination of both). For more details see Zock (1997).

Note that two kinds of processes are possible: syntactic structure is determined prior to lexicalization or words are chosen prior to their (full) syntactic determination. In the first case we have concepts rather than words in the nodes, the output being a syntactic structure, setting constraints for the lexical items to be inserted. This has certain similarities with phrase structure grammars of the early 60ies. In the second case (our approach), lexicalization precedes structure determination, hence we have a lexicalized graph waiting for complete syntactic specification. Once this is done, we have an almost complete lexicalized syntactic structure. What is still lacking are morphological adjustments (agreement, inflexion, etc.) and insertion of function words (determiners, prepositions). Once these are determined we can hand the result to the articulator for converting the symbols into graphemes or sounds.

The following three figures present some prototypes of basic and more complex structures of English.

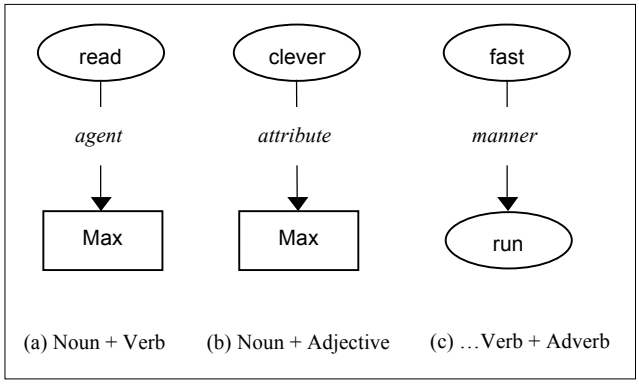


Figure 2. Basic patterns

Taking a look at figure 2 reveals that *entities* are always

[2] http://www.eloquentsw.com/livedictionary.html

[3] http://dev.eurac.edu:8081/MakeEldit1/Eldit.html (login required, but the use is free)

[4] http://www.rikai.com/perl/Home.pl (Japanese-English)
http://www.popjisyo.com/WebHint/Portal_e.aspx (Chinese-Japanese-Corean-English)
http://www.popupchinese.com/tools/newsinchinese (Chinese)

[5] There is a well-known trade-off between storage and time.

presented in rectangles. They typically map on nouns. *Predicates* appear in ovals, mapping either on verbs, adjectives or adverbs depending on the *type of link* (case role vs attribute; case 'a' vs. 'b') and *argument* (entity vs process; verb/adjective vs adverb; case 'a/b' vs. 'c').

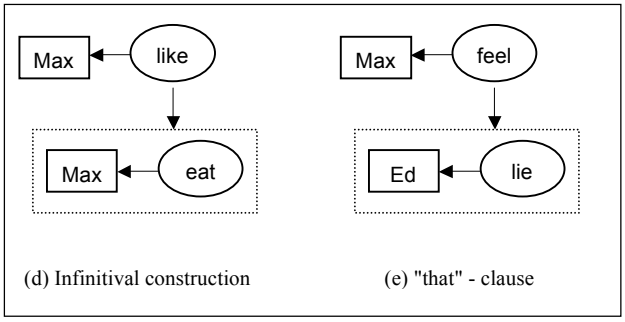


Figure 3. More complex patterns (A)

Figure 3 presents typical representatives of *infinitival constructions* (d) and *that-clauses* (e). The two are somehow similar to 'f' (figure 4), except that in 'd' there is coreference of the deep-subject.

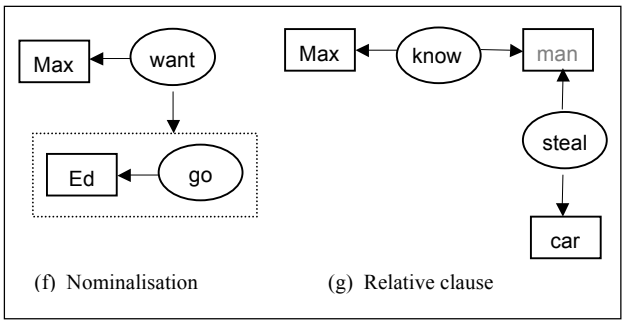


Figure 4. More complex patterns (B)

Figure 4 shows typical structures for *nominalisations* (f) and *relative clauses* (g). The following points are worth mentionning:

- a given conceptual structure may have various syntactic correspondances. For example, 'f' could be realized both as a *nominalisation* (Max wants Ed's departure) or as an *infinitival construction* (Max wants Ed to leave).
- mappings may need to be lexically sensitive. For example, not all bi-transitive verbs allow for passivization. Likewise, not all verbs can be nominalized. These sorts of constraints are well captured in Gross' grammar lexicon (Gross, 1975).
- Patterns are language dependant. This is why learners tend to make mistakes when they switch from their mother tongue to a foreign language. For example, learners of French may have difficulties with nominalizations, which are much less common in this language than, let's say, in English or German.

More complex structures may require chunking, which may alter though the rhetorical effect. For example, relative clauses can be transformed into a sequence of simple clauses, making the text longer, but eventually also more readable. Complexity becomes particularly an issue for multiply, center-embedded relative clauses: "Beer [students [policemen follow] buy] comes from very many different places" (Hudson, 1996).

Having quickly shown how patterns may account for the apparent ease with which people manage to produce language in real-time, let's take a look and see how patterns may be useful in another domain, language learning.

3.2 A multilingual phrasebook augmented with an exercice generator

As we have seen, to speak fluently is a complex skill. If reaching this goal in one's mother tongue is already quite a feat, to do so in a foreign language can be overwhelming. Not only does one have to juggle many constraints —(determine what to say, find the corresponding words, perform the required morphological adjustments, and continue to plan the next segment while articulating, i.e. externalizing, the result produced so far),— but one also has to learn how to do all this in a new language. This requires the acquisition of vast amounts of knowledge, declarative- (words, rules) and procedural, i.e. skills (Levelt, 1975; deKeyser, 2007a).

The project here described is still in its initial phase. Its focus is on learning a new language. More precisely, our goal is to help the learner reach the level of fluency needed to express their basic needs via language (survival level): ask for information; answer a question; solve a concrete problem by using language, etc.

To achieve our goal we have started to build a multilingual phrase-book, an open, customizable, web-based exercise generator and study tool. While the current emphasis is on Japanese, we also work on other languages (English, French, Chinese), the approach being generic. By and large, we'd like to help someone to learn the basic stock of phrases and expressions that are generally taught in the classroom, or that are acquired via a *phrase book*, the latter being structured by tasks a tourist is likely to perform: ask for information in public places, do shopping, etc. Yet, we would like to go beyond this and help the learner not only to learn literally a stock of phrases and words (instance-based generation), but also (or, more importantly) the underlying principles (structures) to build similar sentences. To reach this kind of open ended generativeness we propose an electronic version of a method called *pattern drill* (Chastaing, 1969). There are two good reasons for this:

- in order to be able to perform *automatically*, that is, without having to think about them, a whole set of tasks, we must *exercice* them, as otherwise we will forget or be unable to integrate them into a well staged whole, a prerequisite for *fluency* (deKeyser, 2007);
- different people have *different needs*. This being so, we propose to build an *open system*, allowing the user to tailor the tool to make it fit his or her needs. The

tool is meant to be accessible via the web, but it could also be used on a PDA or a mobile phone.

Pattern drills (PD) are a special kind of exercise based on notions like: analogy, task decomposition (small steps), systematicity, repetition and feedback. Important as they may be, these kind of exercices, are but one of the many tools teachers rely on for teaching a language. Dictionaries, grammars, video and textbooks being supplementary resources. PDs are typically used in audio-oral lessons. Such lessons are generally composed of the following steps:

1. Presentation of a little *story*, where people try to solve a communication problem (hotel reservation, train station, barber shop). The student hears the story and is encouraged to play one of the roles;
2. Contrastive presentation of examples to allow *rule induction*;
3. Use of pattern drills for *rule fixation*;
4. Rule-transposition, i.e. re-use of the learned patterns in a similar but different situation.

These four stages fulfill, roughly speaking, the following functions (a) symbol grounding, i.e. illustration of the pragmatic usage of the structure; (b) conceptualization, i.e. explanation/understanding of the rule; (c) memorization/automatization of the patterns [6], and (d) generalization/transposition/consolidation of the learned material.

The purpose of PDs is to learn words in the context of a sentence, i.e. express quickly more or less complex thoughts. To allow for this kind of exercice by computer, we have built a pattern library whose elements are indexed in terms of goals. This allows not only quick navigation, i.e. finding a pattern in the library, but also to convey roughly what one would like to say (conceptual input). Patterns have constants and variables. Since the value of the latter can be changed by the student we have an open ended system, allowing to enrich patterns with lexical and morphological values in line with the users' needs.

The process works roughly as follows. The student provides a *goal* (eg. *introduce somebody*), to which the system responds with the *patterns* it knows for achieving it. The user chooses one of them, specifying then the lexical and morphological values he'd like the pattern to be instantiated with. The system has now basically all it needs in order to create a set of sentences based on the users' inputs (goal and lexical values). For more details see (Zock & Afantenos, 2009; Zock & Lapalme, 2010).

[6] Psychologists (Posner and Snyder, 1975) draw a clear line between *automatic* and *attentional* processes. The former are fast, parallel, and mandatory. They do not tax memory, neither do they interfere with other tasks or are available to introspection (consciousness). On the other hand, *attentional processes* are slow, serial and they can be observed and controlled, i.e. stopped. They do tax memory, they can interfere with other tasks, and their (intermediate) results can be accessible to our consciousness.

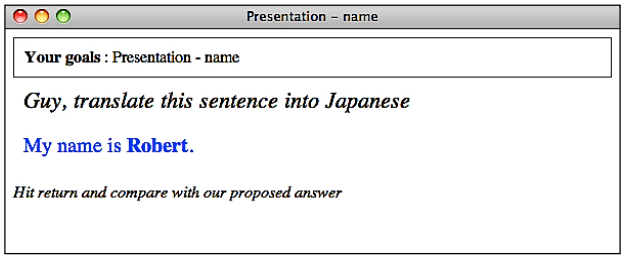


Figure 5. Find all structures containing the term *name*.

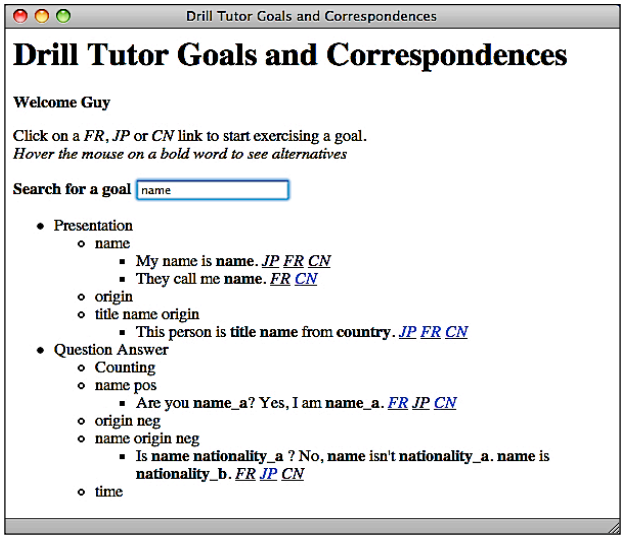


Figure 6. translate "My name is Robert" into the chosen language and hit return

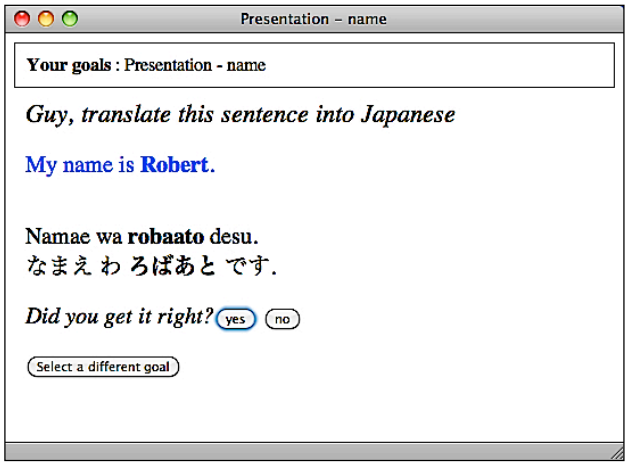


Figure 7. Compare your answer to the system output and tell me whether your answer was right

The number of correct or wrong answers to each goal is saved by the system and the user is shown another instance of the same pattern. It is also possible to choose another pattern as in Figure 5.

Having linked patterns to goals should help users to perceive the function of a given structure (i.e which goal(s) can be reached by using a particular pattern). Yet, most importantly, this linkage offers the possibility to get instances of the pattern from a document (corpus). This is interesting not only for *data acquisition* (building the resource by feeding it with lexical entries likely to occur in a given pattern), but also for *remembrance*. In addition,

presenting patterns with new material allows expanding the learner's experience of the language. The fact that most goals are associated with multiple patterns allows to extend the range of the exercise, reducing thus boredom. Instead of drilling one single pattern in response to a chosen goal, the system can prompt the user by presenting him various patterns.

Obviously, PDs are not a panacea. They can even be harmful it not used properly (paroting, mindless repetition), but used in the right way, that is, at the right moment, with the right goals and at the right proportion, they can do wonders. Just like a tennis player might want to go back to the court and train his basic strokes, a language learner may feel the need to drill resisting patterns. Whoever has tried to become skillfull in a language very different from his own can't but agree with deKeyser (2001) when he writes : "Without automatization no amount of knowledge will ever translate into the levels of skill required for real life use".

Still, true as it may be, we must beware that patterns are but one element of a long chain, i.e one of the many elements of the speaker's toolbox. They need to be learned, but once interiorized they must be placed back into the context where they have come from, an authentic communicative scene. Without this additional experience they will simply fail to produce the wanted effect, that is, help us achieve our communicative goals.

Computers are a medium escaping many of the constraints (rigidity, closedness) other media (tapes or books) are condemned to. They allow for variable order of presentation, dynamic updating of words and much more. Learning a language does not mean memorizing sentences, actually, we tend to forget those sooner or later. What usually remains are ideas, words and patterns, rather than full fledged sentences and rules. Hence, forgetting sentences is not a problem anymore, since we know now how to build them, and this is precisely one of the goals of the tutor described here.

We will now turn to our last tool, a lexical compass

4. Semantic maps and a lexical compass

We spend a large amount of our lifetime searching : ideas, names, documents, and "you just name it". I will be concerned here with the problem of words, or rather, how to find them (word access) in the place where they are stored: the human brain, or an external resource, a dictionary.

No doubt, words play a major role in language production, hence finding them is of vital importance, be it for writing or for speaking (spontaneous discourse production, simultaneous translation). Words are stored in a dictionary, and the general belief holds, the more entries 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 navigational means to reveal the stored information.

Words are basically objects. Like any other object (books, goods in a supermarket, etc.) they pose a storage and access problem. Put differently, once having reached a critical size we must take care to organize and index them, as otherwise we will not find them when needed, despite the fact that they are stored.

I will present here some ideas of how to enhance an existing electronic dictionary, in order to help the user to find the word he is looking for. The goal is to support quick and intuitive navigation based on the users' habits and needs (see also Zock & Schwab, 2006).

What strikes when considering natural observation (introspection) and empirical work is that people having wordfinding problems always know something concerning the target word (meaning, syllables, origine, etc.). This being so, I suggest to start from the known and build a bridge between this point (source-word(s) and the target word.

There are at least two things that people usually know before opening a dictionary[7]: the word's **meaning**, or at least part of it (i.e. part of the *definition)* and its relation to other words or concepts: `x` is *more general* than `y`, `x` is the *equivalent* of `y`, `x` is the *opposite* of `y` (in other words, x being the hypernyme/synonyme or antonym of y), etc. where `x` could be the *source word* (the one coming to one's mind) and `y` the *target word* (the word one is looking for). This is basically conceptual knowledge. Yet, people seem also to know a lot of things concerning the lexical **form** (lexeme): number of *syllables*, beginning/ending of the target word, its *part* of *speech* (noun, verb, adjective, etc.), and sometimes even the *gender* (Brown & McNeill, 1996; Burke et al. 1991; Vigliocco et al. 1997). While, in principal all this information could be used to constrain the search search space, hence, the ideal would be multiple indexes, I will deal here only with the conceptual part (meaning, i.e. partial definition, and the words' relations to other concepts or words).

The yet to-be-built (or to-be-enhanced) resource is based on the age-old notion of association: every idea, *concept* or *word* is connected. In other words, I assume that people have a highly connected conceptual-lexical network in their mind. Finding a word amounts thus to entering the network at any point by giving the word or concept coming to their mind (*source word*) and to follow then the links (associations) leading to the word they are looking for (*target word*). In other words, look-up amounts to navigation in a huge lexical-conceptual space and this is not necessarily a one-shot process.

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

[7] Bear in mind that I am dealing here only with the productive side of language : speaking/writing.

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 wanted to find some word (target word: t_w), yet the only token coming to your mind were *a* somehow related word (source word: s_w). Starting from this input the system would build internally a graph with the s_w at the center and all the words connected to it at the periphery. The graph would be built dynamically depending on the demand. If the list contains the t_w, search stops, otherwise navigation continues, taking either one of the proposed candidates as the new starting point or a completely new token.

Let's take an example. Suppose you were looking for the word ***mocha*** (t_w), yet the only token coming to your mind were ***computer*** (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, 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*, he would probably choose *Java*, which is not only a *programming language* but also an *island*. Taking this latter as the new starting point he might choose *coffee* (since 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 the argotic word of coffee in French.

As one can see, this approach allows word access via multiple routes (there are many ways leading to Rome). In addition, it takes very few steps to make quite substantial leaps, finding a link (or way) between apparently completely unrelated terms. In sum, this is approach is both fast and flexible, at least way more flexible than navigation in a conceptual tree (type hierarchy, ontology) where terms are organized via ISA links, that is hierarchically. In this latter case, navigational mistakes can only be repaired via backtracking.

Of course, one could also have several associations (quasi) simultaneously, e.g., 'black, delicious, strong, coffee, beverage, cappuccino, espresso, Vienna, Starbucks, espresso...' in which case the system would build a graph representing the intersection of the associations (at distance 1) of the mentioned words.

Obviously, the greater the number of words entered and associated to a s_w, the more complex the graph will be. As graphs tend to become complex, they are not optimal for navigation. There are at least two factors impeding readability: *high connectivity* (great number of links or associations emanating from each word), and *distribution* (conceptually related nodes, that is, nodes activated by the same kind of association, do not necessarily occur next to each other, which is quite confusing for the user). This being so, I suggest to display by category (chunks) all the words linked to the source word. Hence, rather than displaying all the connected words as a huge flat list, I suggest to present the words in hierarchically organized clusters, the links of the graph, becoming the nodes of the tree (figure 8). This kind of presentation seems clearer and less overwhelming for the user, allowing for categorical search, which is a lot faster than search in a huge bag of words, provided that the user knows which category a word belongs to.

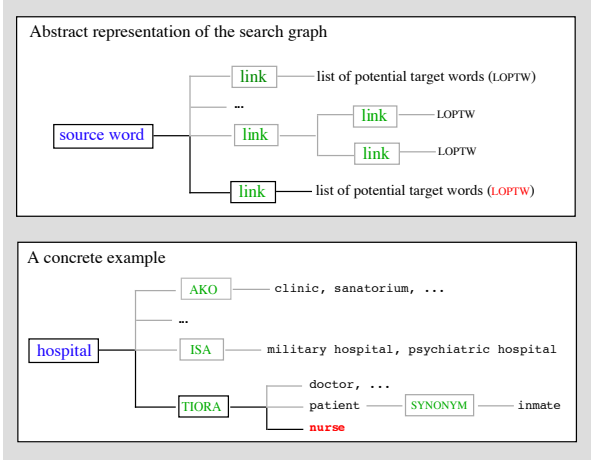


Figure 8: Proposed candidates, grouped by family, *i.e.* according to the nature of the link

5. Conclusion

I have argued in this paper that language processing or language learning require certain tools. Of all these tools macroscopes seem particularly attractive as they reveal structures. Once this is done, we can use them not only to boost performance (for example, produce complex sentences in no time), but also to augment our databases. Indeed, patterns can be used for data mining.

While maps and navigational instrument are badly neeeded, they have to be built with the user in mind. How does s/he organize information? What indexing terms does s/he use? How does s/he build and explore a multidimensional search space? In order to answer these questions, we need to integrate the user in the development cycle right from the start.

6. References

Brown R.; McNeill, D. (1996). The tip of the tongue phenomenon. In: Journal of Verbal Learning and Verbal Behaviour, 5:325-337.

Burke, D.M.; MacKay, D.G.; Worthley, J. S.; Wade, E. (1991): «On the Tip of the Tongue: What Causes Word Finding Failures in Young and Older Adults?». In: *Journal of Memory and Language* 30, 542-579.

Chastain, K. (1969). The audio-lingual habit learning theory vs. the code-cognitif learning theory. *IRAL*, 7(2), 97–107.

deKeyser, R. (Ed.), (2007). Practicing in a second language: Perspectives from applied linguistics and cognitive psychology. Cambridge University Press

deKeyser, R. (2007a). *Skill acquisition theory*. In J. Williams & B. VanPatten (Eds.), *Theories in Second Language Acquisition: An introduction* (pp. 97-113). Mahwah, NJ: Erlbaum.

deKeyser, R. (2001). *Automaticity and automatization*. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125-151). New York: Cambridge University Press.

Gross, M. (1975). Méthodes en syntaxe. Hermann.

Hudson, R. (1996). The difficulty of (so-called) self-embedded structures. UCL Working Papers in Linguistics 8 (1996)

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004). The Sketch Engine. Proc Euralex. Lorient.

Levelt, W. (1989). *Speaking*. MIT Press, Cambridge, Mass.

Levelt, W.J.M., (1975). *Systems, skills and language learning*. In A. van Essen & J.P. Menting (Eds.), *The Context of Foreign Language Learning* (pp. 83-99). Assen: Van Gorcum.

Posner, M., & Snyder, C. (1975). *Facilitation and inhibition in the processing of signal*. In S. Rabbitt& S. Dornic (Eds.), *Attention and performance* (pp. 669–682). New York: Academic Press.

Vigliocco, G.; Antonini, T.; Garrett, M. F. (1997): Grammatical gender is on the tip of Italian tongues. *Psychological Science*, 8, 314-317.

Zock, M. (1997) Sentence Generation by Pattern Matching : the Problem of Syntactic Choice. In : R. Mitkov & N. Nicolov (Eds.) Recent Advances in Natural Language Processing. Series: Current Issues in Linguistic Theory, Benjamins, pp. 317-352

Zock, M. & D. Schwab (2008). *Lexical access based on underspecified input*. Cogalex-08, Coling workshop

Zock, M. & S. Afantenos (2009). *Using e-Learning to achieve fluency in foreign languages*. In Tzanavari A. & N. Tsapatsoulis (Eds.). Affective, Interactive and Cognitive Methods for e-Learning Design : Creating an Optimal Education Experience. IGI Global, Hershey, Pennsylvania

Zock, M. & G. Lapalme (2010). A generic tool for creating and using multilingual phrasebooks. Paper submitted to NLPCS, Funchal, Madeira