Lexical and Textual Resources for Sense Recognition and Description

Jerker Järborg, Dimitrios Kokkinakis, Maria Toporowska Gronostaj

Göteborg University

Department of Swedish Language, Språkdata Box 200, SE-405 30, Göteborg, Sweden {jerker.jaerborg, dimitrios.kokkinakis, maria.gronostaj}@svenska.gu.se

Abstract

It is common knowledge that the creation of language resources for Language Engineering (LE) applications is a time-consuming, and hence expensive, enterprise. From this knowledge stems the demand for the re-usability of resources, which always remains essential. In this paper we will, however, concentrate on another, complementary, aspect, namely that of combining and extending existing resources by a variety of means and with a minimum of manual interaction. The resources to be discussed below consist of (i) a large lexical database, (ii) a formalized computational lexicon, and (iii) a sense-tagged corpus for Swedish. Some results concerning the semi-automatic annotation of the corpus and examples of a variety of phenomena analysed, such as compounding, will also be given. The annotation has been performed within the framework of the SemTag project, while part of this material has been successfully used in the SENSEVAL-2 exercise. In addition to these three resources, it can be added the background material of the Swedish Language Bank (some hundred million words) that forms the basis for the creation of (i) and partly (ii). Having been developed at our department, the lexical resources can easily be accessed, and, more importantly, can be systematically improved where necessary. It should be noted that this type of work requires close cooperation between specialists in lexicography and language technology.

1. Introduction

Under the last decade a number of projects and initiatives within LE have been launched, aiming at the elaboration of recommendations and models for harmonized lexical resources with focus on the description of semantic content, e.g. EAGLES (Expert Advisory Group for Language Engineering Standards), ISLE (International Standards for Language Engineering) and SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica). The relevance of these resources for building semantically tagged corpora, semantic concordances and the like, is apparent, as almost all of the applications concerning natural language processing require virtually some sort of sense recognition and the more semantic information that can be supplied the more semantically advanced and complex processes can be performed.

Creation of testbeds for evaluation and improvement of lexical resources is thus a central task within a number of projects such as ELSNET (Corazzari *et al.*, 2000), WASPS (Kilgarriff & Tugwell, 2001) and SemTag (Järborg, 1997; see below), another is the elaboration of procedures for automatic annotation of corpora as applied to the SENSEVAL exercises. Access to sense-tagged corpus, e.g. corpus composed of word occurrences marked with appropriate sense or semantic information taken from some sort of lexical resource, possibly extended with further lexical semantic information (such as semantic types, domain and selectional restrictions), makes a valuable repository of semantic data to be re-used for semi-automatic enhancing of semantic lexicons, machine translation, linking of parallel corpora.

- In this paper we focus on the description of:
- Swedish lexical resources with emphasis on their semantic content (GLDB, SIMPLE; sections 2 and 3);
- 2. the semantic annotation layer of Swedish textual resources (SemTag; section 4);

- 3. a methodology for the treatment of a number of phenomena relevant to the lexical level of annotation, such as names and compounding;
- 4. a few problematic cases.

2. GLDB

The Gothenburg Lexical Database (henceforward the GLDB) was originally developed in the years 1978-1985 and served as the source of a monolingual Swedish defining dictionary (SO, 1986), followed by a series of improved editions (under various titles). The GLDB was intended both for lexicographic and scientific use, including LE. Thus, although the sense descriptions (definitions etc.) are formulated in ordinary language, some care has been taken to avoid circularities in the definition system and a fairly strict format is used in the definitions. The principal lexical unit of the GLDB is the lexeme, regarded as the union of an expressional unit, the *lemma* (a set of morphosyntactically related word forms) and a content unit, a lexicalized (main) sense. The content unit is regarded as having a core sense (more or less corresponding to a prototypical sense) and optionally one or more lexicalized sub-senses, systematically related to and derivable from the core sense. The types of relations of core sense to subsense are described in the GLDB in terms of (i) extended core sense, (ii) restricted or specialized core sense, (iii) other shifts concerning its figurative or metonymic uses. Changes in the referential scope of typical arguments, as compared with the core meaning, are but one indication of some minor sense differentiation noted by subsenses. The inherent flexibility of the lexeme model was supposed to be an advantage for sense-tagging (see section 4). The GLDB lexeme model is similar to the model used in the New Oxford Dictionary of English (NODE, 1998).

At present, the GLDB contains over 70 000 lexemes with very extensive lexicographic information and is thus a rich, albeit non-formalized, source for computational lexicons. The types of semantic information which can be semi-automatically extracted from the GLDB's structured sense description include semantic data corresponding to Pustejovsky's (1995) qualia, domain information and to some extent information supporting assignment of the lexicon entry to the ontological type.

3. SIMPLE

In the course of Språkdata's participation in the EU projects PAROLE (Preparatory Action for Linguistic Resources Organisation for Language Engineering) and SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica) the structure of formalized computational lexicons was elaborated and model lexicons were built for 12 languages including Swedish. In the SIMPLE project, a computational lexicon with focus on semantic information was produced, comprising some 10 000 semantic units. The units are based on and linked to the syntactic units from the PAROLE lexicon, which means that formalized syntactic information from the latter is directly accessible for the SIMPLE semantic units. These semantic units are described in a strict format which is a cluster of semantic features with predefined values, whenever possible (see Lenci et al. 1999 for full documentation on the model and format). In the Swedish SIMPLE lexicon, the following subset of semantic features has been used to encode the semantic properties and behaviour of nouns and verbs. We exemplify them with values for the noun katt 'cat' and the two core senses of the verb måla 'paint': måla 1 'paint 1, to cover the surface of (something) with paint, as decoration or protection' and måla 2 'paint 2, to depict (an object, person, or scene) with paint'.

- semantic type, whose value is an element in the SIMPLE ontology list (e.g. Semu <katt> 'cat' EARTH ANIMAL; <måla 1> 'paint 1' PURPOSE ACT; <måla 2> 'paint 2' PHYSICAL CREATION);
- domain, whose value is an element in the LexiQuest's domain list (Semu <katt> GENERAL, ZOOLOGY;
 <måla 1> GENERAL; <måla 2> GENERAL, GRAPHIC ARTS);
- semantic class, whose value is an element in the LexiQuest's semantic class list for nouns and EWN classification of verbs. (Semu <katt>: MAMMAL;
 <måla 1>, <måla 2> CREATION);
- *glossa*, a definition taken from GLDB;
- semantic argument structure, list of arguments assigned by the predicative expression. (Semu <måla 1> & <måla 2> arg 1, arg2, arg 3, arg 4);
- selectional restrictions/preferences on arguments, whose values are either semantic types or particular representations of Semus. The latter are chosen whenever the preference is restricted to a unique realisation, e.g. for the verb *mjau* 'miaow' the first argument is specified as <katt>. Otherwise <måla 1> arg1 HUMAN, arg2 CONCRETE ENTITY, arg 3 INSTRUMENT, arg 4 MATERIAL (paint);
- *status of the argument*, the arguments can take one of the following values: *true*, *default* or *shadow*. The true value is chosen when the arguments are obligatorily realized; the default value is for semantically optional realisations and the shadow value is for those arguments which are incorporated in the meaning of a lexical item (Pustejovsky, 1995). Thus, for the verb *måla* the true arguments are: arg1,

arg2, the default argument is arg 3 and the shadow argument is arg 4;

- *link to the syntactic unit* (Synu). The Synus in the Swedish PAROLE lexicon are linked to the Semus in the Swedish SIMPLE lexicon, which is effected in a robust information block with a coherent and exhaustive morphological, syntactic and semantic description. The linking of these units is one-to-one, one-to-many or many-to-one;
- *link to a corresponding lexeme in GLDB*, which not only provides access to all the lexical information encoded in GLDB, but also relates these two resources to each other.

The latter linkage is substantial for further utilization of GLDB and for enhancing of SemTag with new layers of semantic information, which in turn augment the scope of possible applications.

3.1. Extending SIMPLE

Promising experiments have been made to automatically extend the coverage of the relatively small SIMPLE lexicon by taking into consideration the compounding, a distinctive feature of the Swedish language, and semantic similarity in noun phrases of enumerative type.

In the first approach, we have assumed that a considerable number of casual or on the fly created compounds can inherit relevant parts of semantic information, provided that the heads of lexemes occur in the SIMPLE lexicon, and make their incorporation into the lexicon feasible. However, in order to restrain automatic incorporation of lexicalised compounds having idiomatic, metaphoric or metonymic meaning, we check whether a compound is included as a separate entry (i.e. whether it is lexicalised) in the GLDB defining dictionary. If this is the case, the compound is not subjected to automatic inheritance. We apply compound segmentation only to the productive compounds and only the heads of these are matched against the SIMPLE lexicon and annotated with information on their semantic class. For instance, none of the words in the phrase: färjor, kryssnings|fartyg, tankers och ro-ro-lfartyg 'ferries, cruise liners, tankers and ro-ro-vessels' are in the SIMPLE lexicon. During segmentation, however, the second and fourth words' heads get the label VEHICLE since they match the entry fartyg 'vessel' (which is in the SIMPLE lexicon) and, based on a number of conditions satisfied, the remaining nouns can be matched with the same class; see (Kokkinakis et al., 2000) for details on this method.

The above approach can be even combined with parsing of specific types of noun phrases. After the compound segmentation process has been performed and the semantic matching with items in the SIMPLE lexicon accomplished, one can take into consideration the semantic similarity of the noun types occurring in enumerative noun phrases. Thus, in this approach, we investigate how, and to what extent a partial parser can be utilized to automatically extend existing semantic lexicons. The method rests on the assumption that words entering into the same enumerative, syntagmatic relation with other words are perceived as semantically similar. Given the semantic information on a few category members, we automatically collect and examine surrounding contexts and thus try to identify other words that might also belong to the same semantic class. Large quantities of partially parsed corpora are an important ingredient for the enrichment and further development of the semantic resources. For instance, the common nouns: jurist 'lawyer', optiker 'optician' and läkare '(medical) doctor' have being manually coded in the (original) SIMPLE lexicon, with the OCCUPATION-AGENT semantic class ("individuals or groups of humans identified according to a role in professional, social or religious disciplines"). In the noun phrase: jurister, läkare, optiker, psykologer och sjukgymnaster the three first nouns get the OCCUPATION-AGENT label, while the two last, namely 'psychologist' and 'physiotherapist' will also get the same label by the system, since they satisfy a number of condition, such as that they have not received a semantic class annotation and the rest of the members of the phrase (at least two) have been assigned the same semantic class; see (Kokkinakis, 2001) for details on this approach.

Both methods are of equal importance: through parsing we allow the incorporation of new, mainly noncompound words; through compounding we allow new compounds of existing words. The experiments carried out so far have shown that input data of 1000 entries can be expanded to 25 000 (22 000 through compounding and 3000 through parsing).

It may also be possible to equate some SIMPLE semantic classes with certain headwords (found by pattern matching) in the GLDB definitions and thus extend the SIMPLE lexicon with new entries described partially, e.g all the words in the GLDB defined as *person, man, woman, child* etc. might be automatically assigned to the semantic class HUMAN. A similar procedure can apply to many other concrete and abstract noun classes, such as for example animals, body parts, different types of artifacts or properties. In other words, hyperonymy relations between the lexicon entries and their headwords (genus proximum) in GLDB's definitions can not only be utilized to enrich the SIMPLE lexicon but also to make this type of semantic information explicit and accessible in the GLDB.

There is no doubt that systematic enhancement, harmonization and integration of the semantic information in these two lexical resources can be carried through even on other semantic layers. The domain layer is one of these, as the domain inventory in these two resources shows some overlap. Labels referring to sciences, like mathematics chemistry, zoology etc., arts and its subbranches, like literature, music, or other subject areas, like gardening, knitting, heraldry, are but some examples which make the transfer of domain information in the both directions feasible.

4. SemTag

The sense-tagging project is called *Lexikalisk* betydelse och användningsbetydelse "Lexical Sense and Sense in Context", henceforth informally referred to as SemTag. The theoretical aim of the project was to test the hypothesis of the GLDB, namely that it is possible to describe the lexical senses of words in a way that predicts their senses in actual linguistic usage. The practical goal was to produce a large, semantically tagged, text corpus, to serve as an empirical base for language engineering. The tagging is carried out in a KWIC format interface (see figure 1) and proceeds alphabetically, although some

articles in the corpus (over a million tokens, balanced according to the principles of the Brown corpus) have already got each text word tagged¹. At least 250 000 text words have been tagged. Many of the text words consists of predictable, non-lexicalized compounds, both elements of which are represented in the GLDB; such words are subject to special treatment (see below).

The SemTagged material can be accessed and sorted in various ways, including a sense-ordered version, in which the semantic description from the GLDB is given, followed by all the concordance lines containing the lexeme or sub-sense in question. This version can be regarded as a kind of lexicon with a large set of authentic examples or as a collection of text passages in which the words have been annotated semantically; cf. Miller et al. 1993 and Landes et al. (1998) and the notion of semantic concordance exemplified in these works and implemented in the SEMCOR corpus, as well as the work by Ng & Lee (1996) resulting into the DSO corpus. In this form, the material is being extensively used in a project investigating lexical relations from a cognitive perspective (Norén, 2002). In Language Engineering, SemTagged material has already been used in the Swedish part of the SENSEVAL-2 competition (Kokkinakis et al., 2001) and the full material will, in time, be made generally available for research via the Internet.

The SemTag project, which has been carried out by several persons, seems to prove that manual sense-tagging can be carried out with a great deal of confidence, which partly can be explained by the dynamic lexeme model of the GLDB. Most uncertain cases are due to avoidable deficiencies in the GLDB (i.e. where the semantic descriptions do not adhere to the GLDB principles, see section 5) and can thus be eliminated after systematic adjustments of units or definitions. The process allows for qualitative and quantitative evaluation of manual sensetagging, as well as providing a valuable feedback for the GLDB.

It may be thought that the principal computational use of the SemTagged material would be as a 'gold standard' for semantic disambiguation experiments. However, since the GLDB and the SIMPLE lexicon have been linked, it follows that a text word tagged as a GLDB sense unit can be automatically marked with the semantic classes and domains of SIMPLE, if it belongs to the central vocabulary. Since the SIMPLE lexicon is formalized, many other uses become feasible. (Here it should be noted that the final element of a compound noun, which is the semantic head in most cases, will often belong to the central vocabulary, which means that the coverage of the SIMPLE lexicon in the SemTagged material is greater than would appear at first sight, cf. section 4.1). Thus, many sentences and constituents in a sense-tagged text can be analysed in some basic semantic dimensions. It is expected that the corresponding features bundles will prove to be sufficient e.g. for checking syntactic parsing or for roughly classifying the content of a text. Although the corpus used in SemTag is not tagged in its entirety yet, it may also be possible, in some cases, to define the

¹ At this stage of the project the Stockholm-Umeå Corpus (SUC) Ejerhed *et al.* (1992) is used for the lexical semantic tagging. SUC is already part-of-speech annotated and manually verified.

'semantic context' for a given lexeme in the SIMPLE categories. For instance, the two senses of the verb *träna* 'to train' namely 1. improve one' s particular skill[...] and 2. teach (a person or animal) a particular skill, restrict the semantic context in different ways. For the *träna 1* the subject argument is restricted to the semantic class HUMAN and the object argument to the class PURPOSE ACT (e.g. *hon tränar löpning* 'she practises running') and for *träna 2* the restrictions on human subject are weakened, as the nouns subcategorized as HUMAN or ACT can appear in this position (e.g. *läraren/räkning*)

tränar skolbarnen i abstrakt tänkande 'the teacher/ counting trains the pupils in abstract thinking'). The direct object position can be filled only with the nouns of the type HUMAN and it can be followed by a prepositional object referring to PURPOSE ACT or some PROPERTY. It might be worth to mention that the subsense of the latter points out the BODY PART as a possible realisation of the direct object (e.g. *Han tränar sina muskler* 'He trains his muscles').

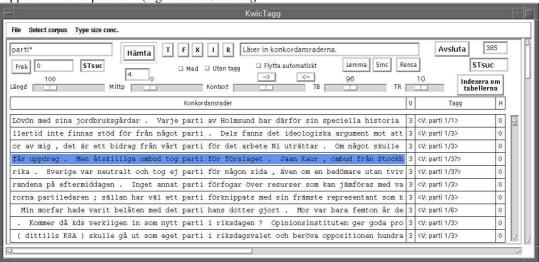


Figure 1: The KwicTag Interface

4.1. SemTag: Compounding

Swedish is a strongly compounding language (like most other Germanic languages) and many compound text words in any Swedish material are lexically not compounds, represented, temporarily constructed generally with predictable meanings. In almost all cases, such compounds are essentially binary and, in most cases, both elements are represented in the GLDB. In SemTag, the lexically represented elements are separately tagged with the respective GLDB units, whereupon the semantic relation between the elements is described by a schematic definition, selected from a limited set (at present comprising some 50 definitions; see table 1). Thus, the compounds bomullstyg and klänningstyg are analysed as follows (in translation):

1st element (X): bomull 1/1 'cotton'
2nd element (Y): tyg 1/1 'fabric'
Relation: 'Y that consists of X'
1st element (X): klänning 1/1 'dress'
2nd element (Y): tyg 1/1 'fabric' Relation: 'Y that is intended for X'
Relation: 'Y that is intended for X'

Even from these elementary examples, it can be seen that it would be possible, given a sizable number of such analysed compounds, to automatically establish a "semantic compounding profile" for all lexemes in predictable compounds. Here, the lexeme *tyg 1/1* coud be described as having the tendency to be the semantic head of attributes referring to material or to intended use. Similarly, the lexeme *bomull 1/1* would probably almost always appear in compounds as "X"in the relation 'Y that consists of X', i.e. as an attributive material. Such semantic information would seem to be a useful complement to formal semantic classification of the SIMPLE type; examples are given in the table below (the last definition is considered the default one).

Semantic Definition	Example
Y that is located in/at	<i>klassrumsdörr</i> classroom+door
Y that is made up of X	<i>kanalsystem</i> canal+system
Y that originates from X	<i>smutsfläck</i> dirt+stain
Y that is aimed at X	kaninjakt rabbit+hunt
Y that is about X	<i>partikelfysik</i> particle+physics
Y that produces X	<i>batterifabrik</i> battery+factory
Y that prevails in X	<i>partiideologi</i> party+ideology
Y that contains X	<i>kaffetermos</i> coffee+thermos
Y that consists of X	kaffepulver coffee+powder
Y that has to do with X	klädbesvär clothes+trouble

Table 1: A Selection of Semantic Definitions with Examples

Obviously, compounding profiles would enable precise automatic analyses of new compounds with known

elements and good approximate analyses of compounds with an unknown first (or even second) element. They may also assist in analysing the semantic relation between the elements used as free units in syntactic constructions (*köpa tyg till en klänning* 'buy fabric (to be used) for a dress'; *cf. köpa tyg till en släkting* 'buy fabric to (be used by) a relative'). At the very least, the compounding profiles could be used for semantically verifying the results of morphologically based compound segmentations.

4.2. Semantic Annotation: Names

Names, both proper names and single or multiword names of various types are appropriately marked. Such names include: surnames and undecidable person names (e.g. *Martinsson*), female first names (e.g. *Anna*) and male first names (e.g. *Magnus*), place names (e.g. *Kaliningrand, Litauen*) organizations, object and brand names (e.g. *Volvo*), institution names (e.g. *Konstfack*), titles of films, books and radio programs (e.g. *Klarspråk*), place names of various types, such as bodies of water (e.g. *Klappmarksbäcken, Långsjön*), addresses (e.g. *Klippgatan*).

Sometimes names, particularly those of firms, institutions and titles will be multiword names (next section).

4.3. Semantic Annotation: Multiwords and Idiomatic Expressions

While manually annotating text corpora, special attention has been paid to lexemes occurring in idiomatic expressions. These lexemes are tagged in a twofold way, that is, with the lemma and sense number to show their affiliation to a particular meaning and also with a label idiom to show that they are part of an idiom listed in the comment field. Thus, an inventory of the idiomatic expressions instantiated in the tagged corpora can be compiled. Its relevance for NLP tasks like automatic annotation of corpora, machine translation and summarization is obvious.

Thus, in the example *kusten var klar* 'the coast was clear' both the lexeme *kust 1/1* 'coast' and *klar 1/4* 'clear' will be marked with the additional information: *i uttr.* "*kusten är klar*" 'in the expression "the coast is clear"'. In the case where a multiword unit is also some kind of name, the following format is used: *i uttr.* "*Kalle på Spången*" (name, title).

4.4. SemTag: Problematic Aspects

The chief linguistic problem of semantic tagging is that the lexical base, in this case the GLDB, is semantically inadequate in some way. This problem is discussed in section 5. A minor problem is when authors are consciously ambiguous, e.g. using punning and similar meta-linguistic devices. Instances of such things are very rare.

Problems of a more technical kind are more frequent (although probably negligible, statistically speaking). The two most typical cases are: (i) The text word is a regular derivative of a lexeme in the GLDB but is not represented there. Thus, due partly to lexicographic tradition, deverbal nouns derived with the suffixes *-ning* or *-ande* and abstract nouns derived from adjectives with the suffix *-het* (corresponding to English *-ness*) are normally not

represented in the GLDB. Such cases are treated as a kind of compounds, lacking overt last elements. Thus, the text words mjukhet are tagged as mjuk 1/1 and mjuk 1/2/a, as adequate, and are secondarily marked as having a second, nominal element -het. (ii) The text word is a component of non-lexicalized compounds. Thus, the text word kultursponsringsfrågor has as first element kultursponsring 'sponsoring of culture', which of course is not represented in the GLDB, and can only be tagged as 'unknown entity' (such as names). The second element is tagged fråga 1/2 'issue'. The internal composition of the first element (the element of which kultur and sponsring are represented in the GLDB) is described in a secondary comment.

5. Enhancement of the Lexical Resources

In the course of applying the lexical units of the GLDB to a large text corpus, several types of cases will be encountered where the database does not cover the text words. Some of these cases are of a technical nature and are described in section 4.4. Then there are words that cannot reasonably be represented in a lexical database, since they are of a more or less encyclopedic nature (and also unlimited in number); to this category belongs the proper nouns and others. When it comes to " real", not compound, words not represented in the GLDB, one has to make a decision for each word whether it is sufficiently familiar and lexically central for it to be added to the database (obviously, highly specific scientific terms and the like have a low priority for a general lexical database). Nevertheless, many new lemmas and lexemes have passed the tests and been included in the GLDB, mostly because they reflect a changing reality. This type of enhancement, simply the addition of new lexical entities, is an obvious result to be gained by testing the database on new texts. There are, however, more subtle (and interesting) changes that are motivated by the sense-tagging experience: the establishing of new lexemes under the same lemma, the splitting of a lexeme into two or, conversely, the collapsing of two lexemes into one, the addition of new sub-senses to existing lexemes, the re-organisation of core senses and sub-senses into a new local semantic network, the re-formulation of definitions (changing headwords and hence semantic category), and so forth. Many of these types are described in Järborg (1999). It should be emphasized, though, that changes in the database must be systematic and consistent with general semantic principles: a single text instance with deviating semantics is not enough to motivate the establishment of (say) a new sub-sense.

5.1. New Senses

Describing and motivating lexical changes in detail requires a good deal of space; here we will restrict ourselves to one simple example. In the case of the lexeme *plast* 'plastic', the core sense is described as being that of a type of material with certain properties and uses (there is also a sub-sense focusing on connotations of low quality and so forth). However, in the texts there are several instances of *plast* being used for a film of this material, used for covering something. A comparable metonymic use is lexicalized in the case of *järn* 'iron'; the core sense is also that of a material but a sub-sense 'implement of iron' has been established. It is thus logical to establish a new sub-sense of *plast*, 'film of plastic' (with some further refinements). It is quite likely that the lack of this obvious and necessary sub-sense would not have been discovered, had not the GLDB been used for sense-tagging. Very many similar examples could be cited.

6. Applications

The lack of high quality as well as the slow progress on a number of LE applications has been blamed on wordsense ambiguity and the almost non-existence of appropriately annotated material with lexical semantics. Therefore, a semantically tagged corpus is an important and urgently required ingredient for training and evaluation within a large spectrum of applications (text categorization, query-based Information Retrieval, WWW search engines natural language understanding and Machine Translation, to name a few). Semantically annotated texts serve also as a natural accompaniment to lexical databases for the sake of facilitating and qualitatively improving the information already present in such resources. For instance, as a useful testbed to evaluate the lexicon's coverage and disambiguation power, as a repository of corpus examples for the attested senses, as means for checking the frequency of lexicon senses and their co-occurrences.

7. Conclusions

In this paper, we have described lexical and textual resources capable of providing a variety of useful input (semantic layers) for the description and recognition of lexical semantic information for the Swedish core vocabulary. These semantic layers allow a diversity of semantic profiles to be detected, which range from defining a domain area, determining a text structure profile of a material to extraction of semantic patterns for particular lexemes. Domain information encoded in SIMPLE and GLDB resources provides relevant support for the recognition and classification of texts with respect to their domains. Information on semantic types and in particular ontological classification of verbs, with a minutious subcategorization of Speech Act verbs contribute to recognition of the dialogue text from the running prose texts. This information is of importance for both semantic text annotation, syntactic parsing and semiautomatic extraction of subcategorization patterns with integrated semantic and syntactic information. A large semantically analyzed corpus allows us not only to identify new senses and to more precisely describe existing ones, but it also provides us with the appropriate mechanisms to capture shifts in meaning, identify more general senses or even collapse two or more into a single one. Continuous efforts to lexical-semantic tagging of a large Swedish corpus performed within the framework of the SemTag project have been reported.

Semantically annotated corpora opens up exciting opportunities for linguistic analysis, contributing with very important information for the precise assignment of lexical semantic knowledge to polysemous and homonymous content words. The existence of sense or semantic ambiguity is one of the major problems affecting the usefulness of basic corpus exploration tools and a number of LE applications.

- Corazzari, O., N. Calzolari and A. Zampolli, 2000. An Experiment of Lexical-Semantic Tagging of an Italian Corpus. *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC)*. Athens, Greece.
- Ejerhed E., G. Källgren, G. Wennstedt and M. Åström, 1992. *The Linguistic Annotation of the Stockholm-Umeå Corpus project*, Tech. Rep. 33, Univ. of Umeå.
- Järborg J., 1997. Lexikalisk betydelse och användningsbetydelse (LBAB). *Project Description. Technical Report.* Språkdata, Göteborg University. (In Swedish).
- Järborg J., 1999. Lexikon i konfrontation. Rapport från LBAB:4, Lexikalisk betydelse och användningsbetydelse. *Research Reports from the Department of Swedish GU-ISS-99-6*. (In Swedish) (http://svenska.gu.se/~svedk/resrapp/konfront.pdf)
- Kilgarriff A. and D. Tugwell, 2001. WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-theart Lexical Disambiguation. *Proceedings of the MT Summit VII*. Santiago de Compostela,187-190.
- Kokkinakis D., M. Toporowska Gronostaj and K. Warmenius, 2000. Annotating, Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon. *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC)*. Athens, Greece.
- Kokkinakis D., J. Järborg and Y. Cederholm, 2001. SENSEVAL-2: The Swedish Framework. *Proceedings* of the SENSEVAL-2 Workshop. Part of the 39thACL/10thEACL Joint Conference. Yarowsky D. and Preiss J. (eds). Toulouse, France.
- Kokkinakis D., 2001. Syntactic Parsing as a Step for Automatically Augmenting Semantic Lexicons. *Proceedings of the Student Workshop in the 39th ACL/10th EACL Joint Conference.* Toulouse, France, 13-18.
- Miller G.A., C. Leacock, R.I. Tengi, and R.T. Bunker, 1993. A Semantic Concordance. *Proceedings of the ARPA WorkShop on Human Language Technology*. San Francisco, Morgan Kaufman.
- Norén K., 2002. Polysemityper i ordförrådet. Nordiska studier i lexikografi 6. Rapport from konferens om lexikografi i Norden. Färöarna.
- Landes S., C. Leacock and R.I. Tengi, 1998. Building Semantic Concordances. WordNet: An Electronic Lexical Database, Chapter 8. Fellbaum C. (ed.). The MIT Press, 199-216.
- Lenci A., F. Busa, N. Ruimy, E. Gola, M. Monachini, N. Calzolari and A. Zampolli, 1999. *Linguistic Specifications, SIMPLE Work Package 2, Deliverable D2.1*. University of Pisa and ILC of CNR.
- Ng H. T. and H.B.Lee, 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, California, USA, 40-47.
- NODE, 1998. *The New Oxford Dictionary of English*. Oxford University Press.

Pustejovsky, J., 1995. *The Generative Lexicon*. Cambridge. MA: The MIT Press.

SIMPLE:

http://spraakdata.gu.se/simple/swedish.simple.lexicon.html

SO, 1986. *Svensk ordbok*. Språkdata and Esselte Studium AB.

8. References