Annotating, Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon

Kokkinakis D., Toporowska Gronostaj M., Warmenius K.

Språkdata, Göteborg University
Box 200, SE-405 30,
Sweden
{svedk, svemt, svekws}@svenska.gu.se

Abstract

During recent years the development of high-quality lexical resources for real-world Natural Language Processing (NLP) applications has gained a lot of attention by many research groups around the world, and the European Union, through the promotion of the language engineering projects dealing directly or indirectly with this topic. In this paper, we focus on ways to extend and enrich such a resource, namely the Swedish version of the SIMPLE lexicon in an automatic manner. The SIMPLE project (*Semantic Information for Multifunctional Plurilingual Lexica*) aims at developing wide-coverage semantic lexicons for 12 European languages, though on a rather small scale for practical NLP, namely less than 10,000 entries. Consequently, our intention is to explore and exploit various (inexpensive) methods to progressively enrich the resources and, subsequently, to annotate texts with the semantic information encoded within the framework of SIMPLE, and enhanced with the semantic data from the *Gothenburg Lexical DataBase* (GLDB) and from large corpora.

Introduction

During recent years there has been an increased interest to acquire, on a large-scale, high-quality semantic lexicons, McKeown & Hatzivassiloglou (1993), Dorr & Jones (1996), Hearst & Schütze (1996), Takunaga *et al.* (1997), Roventini *et al.* (1998), Viegas *et al*. (1998). The methodology behind these approaches is usually corpus-driven. It is based on the (re-)use of machine readable resources of various types, and the application of cost effective ways to eliminate the acquistion bottleneck, such as derivational morphology, customization of off-the-shelf resources and statistical techniques. The approach adopted here for the extension task is in line with the methodologies mentioned.

In this paper, we focus on ways to extend and enrich, as far as possible, automatically the coverage of the Swedish semantic lexicon by taking into consideration compounding, a distinctive feature of the Swedish language, and semantic similarity in noun phrases of enumerative type. With the support of semantic data from the Swedish SIMPLE[1] lexicon (*Semantic Information for Multifunctional Plurilingual Lexica*, LE4-8346), Gothenburg Lexical DataBase (GLDB) and large corpora both raw and exposed to shallow parsing, we enhance the incorporation of new semantic entries into the SIMPLE lexicon. We expect to be able to extend the 6,000 entries in the Swedish SIMPLE lexicon to over 120,000 entries. Our assumption is based on the results obtained from the tests carried out so far on input data of 1,000 entries, which became 25,000 (22,000 through compounding and 3,000 through semantic similarity).

Furthermore, we semantically annotate texts with all the available material, and we apply Machine Learning techniques for the disambiguation of ambiguous readings. The annotation task provides an excellent opportunity to evaluate the usability of the semantic information encoded in SIMPLE.

This paper is organized as follows: first we give a brief presentation of the SIMPLE project and particularly of the Swedish lexicon; then we present how compounding and semantic similarity in enumerative phrases (under certain conditions) can contribute to the augmenting and enrichment of the lexicon, when subjected to compound segmentation and shallow parsing; we continue by describing a practical application of the semantic lexicon, namely semantic annotation and disambiguation; we then give some general remarks on the usability of the SIMPLE model, while conclusions end the presentation.

The SIMPLE Project

The EU-financed SIMPLE project aims at developing wide-coverage semantic lexicons for 12 European languages. The Swedish SIMPLE lexicon (hereafter Swe-S) is one of these. All lexicons share a common semantic model and a common encoding formalism in SGML. The semantic data in the SIMPLE lexicons is being linked to the morphological and syntactic data in their respective PAROLE lexicons, developed within the EU project PAROLE, (*Preparatory Action for Linguistic Resources Organisation for Language Engineering*). Out of the 20,000 words in the PAROLE lexicons, a subset of about 6,000 words, or approximately 10,000 senses, has been enriched with semantic descriptions in the SIMPLE counterpart. The content and the design of the SIMPLE model are documented in Lenci *et al.* (1998).

The notion of semantic type is central for the SIMPLE model and its ontology. It corresponds to a word sense assigned to a lexical item. There are 139 semantic types distinguished in the SIMPLE ontology. Each semantic type is defined as a cluster of structured semantic information significant for a given word sense. Information on semantic class, domain, argument structure of predicative expressions and selectional restrictions on arguments as well as qualia roles constitute

[1] The following sites provide more information about SIMPLE: http://spraakdata.gu.se/simple/swedish.simple.lexicon.html & www.ub.es/gilcub/SIMPLE/reports/simple/Site_simple.htm.

a relevant part of the semantic type specification; (Calzolari (1999), Pedersen & Keson (1999)). The SIMPLE ontology is multidimensional as it is based on the principle of orthogonal inheritance (Pustejovsky 1995), and in this respect, it contrasts with the LexiQuest's[2] semantic class ontology which is based on a standard, monodimensional approach. The latter ontology includes 95 semantic classes. Both ontologies are hierarchically structured.

The Swedish SIMPLE Lexicon

The theoretical and formal design of the Swe-S lexicon is conformant to the SIMPLE's linguistic guidelines presented by the specification group, Lenci *et al.* (1999). In the Swe-S lexicon, there are about 10,000 semantic units (hereafter Usems) encoded, comprising 7,000 noun, 2,000 verb and 1,000 adjective Usems. These 10,000 units are mapped onto 6,000 entries. Usems are described with respect to the following information:

- ***semantic type***, whose value is an element in the SIMPLE ontology list (e.g. Usem <katt> 'cat': EARTH_ANIMAL).
- ***domain***, whose value is an element in the LexiQuest's domain list (Usem <katt>: ZOOLOGY).
- ***semantic class,*** whose value is an element in the LexiQuest's semantic class list. (Usem <katt>: MAMMAL).
- ***glossa,*** a definition taken from GLDB.
- ***semantic argument structure***, list of arguments assigned by the predicative expression.
- ***selectional restrictions/preferences on arguments***, whose values are either semantic types or representations of Usems. Usems 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>.
- ***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).
- ***link to the syntactic unit*** (Usyn). The Usyns in the Swedish PAROLE lexicon are linked to the Usems in the Swe-S 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 either 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.

In the course of building the Swedish SIMPLE (and PAROLE) lexicons we have, to a large extent, reused lexical data from GLDB which is the most comprehensive source of lexical information on contemporary Swedish, and information from the SO (1992) and NEO (1996).

[2] LexiQuest is the French partner in the SIMPLE project.

Extending the Coverage of the Swe-S

The Swe-S resources are not quantitatively sufficient for realistic, large-scale Natural Language Processing (NLP) tasks, such as semantic annotation, and need to be extended. For this particular task, we take advantage of the productive compounding characteristic of Swedish and the use of raw and partially parsed corpora.

We assume that a considerable number of casual, or on the fly created compounds in Swedish can inherit relevant parts of semantic information provided on their heads by the Swe-S lexicon and thus, can be incorporated into the lexicon. By relevant parts, we mean in the first place the information concerning semantic type, domain and semantic class. To avoid errors, we exclude the information on argument structure from the inheritance, as the argument structure can undergo alternations in the process of compounding. This is the case when verbs and verbal nouns build compounds with either an obligatory or optional argument in the non-head position. The occurrence of an adjunct in the non-head position does not usually alter the predicative structure.

Compounding

The fact that over 70%, or approximately 80,000, of all the entries in the SAOL (1998) are compound forms casts light not only onto an immense lexical repository, which is available for this particular extension task, but also on the need to design effective tools and routines for compound segmentation, as new, casual compounds are created constantly in Swedish. Most of these casual compounds are relatively transparent, which implies that their meaning is a function of the meaning of its components being related to each other by an implied predicative functor. For instance, *brödkniv* $brödX knivY$ 'bread knife' implies '*Y* for (cutting) *X*' and *bärsaft* $bärX saftY$ 'juice from berries' implies *Y* which contains *X*. In Swedish, compounds are written as single orthographic units and nouns are the most frequent modifiers occurring in non-head positions[3].

A combination of various heuristic methods is used for the extension. Compound segmentation[4] is applied to compound noun tokens on large corpora and lists of new nouns are produced. To maintain quality assurance and compatibility with the rest of the data in the lexicon, new heuristics are applied to the content of the noun lists produced. To avoid generation of incorrect data, these heuristics inspect the modifying component of a

[3] According to Blåberg (1988) the most frequent modifying part of Swedish compounds are, in order, nouns, followed by adjectives, proper names, adverbs and prepositions, then verbs and finally numerals.

[4] Compound segmentation is based on the distributional properties of graphemes. It involves identifying grapheme combinations that are not-permitted when considering non-compound forms in the Swedish language, which carry information of potential token boundaries. The heuristic principle behind the segmentation is based on producing 3-*gram* and 4-*gram* character sequences from several hundreds of non-compound lemmas, and then generating 3-*gram* and 4-*grams* that are not part of the lists produced. Some manual adjustments have also been imposed. Ambiguities are unavoidable, although the heuristic segmentation has been evaluated for high precision, (over 95%).

compound in order to distinguish its characteristics, such as its part-of-speech and semantic category (if any). These characteristics of the modifier, when enriched with the corresponding characteristics of the compound head, provide data for a preliminary estimation of the correctness of the heuristics. Few examples will illustrate this point.

If the part-of-speech of the compound modifier is an adjective, a new Swe-S entry, which will not cause semantic anomalies in the derived lexical set, can be created with great confidence. The inheritance criterion applies here and the compounds are hyponyms to the head. For example, the lemma *klocka* 'bell/watch' can be extended with compounds of type [ADJ-MODIFIER]+HEAD: [*digital*]*klocka*, [*guld*]*klocka*, [*lill*]*klocka*, [*silver*]*klocka*, [*stor*]*klocka*, where the adjectival modifying part in these examples are 'digital, gold, little, silver' and 'big'. Similar results are obtained if the modifying part is a proper noun. For instance, *anhängare* 'supporter' with modifiers such as: *Berisha, Hammarby, Hitler, Likud* and *Mobutu*, signal unambiguous compounds.

It is well known that the heuristics have a variable degree of performance on different types of compounds, and that some simple constraints are needed to exclude segmentation and interpretation errors. Particularly in the case where the part-of-speech of the modifying part of a compound is a noun (e.g. NOUN-MODIFIER[*kultur*]*fråga* 'cultural question') or verb (e.g. VERB-MODIFIER[*betal*]*teve* 'pay-TV'). These constraints are formed by means of subroutines which impose checking of derived compounds against different lists to eliminate incorrect data. The lists with bound morphemes or lexicalized compounds, extracted from the GLDB allow exclusion of such compounds from the derived sets. Such constraints have proven to be a cheap way to automatically constrain the overgeneration of new entries in the lexicons.

For instance, when using large corpora, over 40 compounds with *feber* 'fever', as head, could be extracted. However, it became evident that not all of them belong to the semantic class of ILLNESS, e.g. *resfeber* 'excitement before a journey'. Thus, in some cases, additional inspection seems unavoidable, if we want to restrain automatic incorporation of lexicalised compounds with idiomatic, metaphoric or metonymic meanings. This inspection can be performed automatically by simply checking whether a given compound is included as a separate entry in GLDB. If this is the case, it means that the compound is lexicalised and should not be subjected to automatic inheritance. The manual inspection is needed, only if the derived compound shows diverging semantic and/or morphological patterns and the word is neither in a bound morpheme list, nor in the lexicalised compound list.

Moreover, the content of the Swe-S has been used as a means of bootstrapping the process. For instance, *glas* 'glass', can be extended with compounds having SUBSTANCE as a modifier in the compound form. Consequently the [NOUN-MODIFIER{SUBSTANCE }]+HEAD compounds [*vatten*]*glas*, [*vin*]*glas*, [*öl*]*glas*, [*likör*]*glas* all have SUBSTANCE as the modifier part, namely 'water, wine, beer' and 'liqueur'.

A large number of already disambiguated compounds has been also extracted from GLDB, since the Swe-S entries are linked to the various senses and sub-senses in GLDB, and subsequently to the morphological examples of every entry (alias compounds). For instance, Swe-S encodes the non-compound lemma *ämne* (as having four senses, marked with 1/1-1/4), which are disambiguated here by means of their assignment to the following semantic types and semantic classes:

Material: MATTER 'material'
Substance: SUBSTANCE 'stoff'
Part: ABSTRACT 'topic'
Domain: NOTION 'subject, discipline'

Each of these senses is exemplified in GLDB with a number of compounds, comprising totally 26 compounds with *ämne* as the head. Some of these are listed in the right column of table (1). Since there is only one compound with that head in the Swe-S lexicon (*grundämne* 'element'), incorporating new, disambiguated compounds was straightforward.

Swe-S	GLDB
ämne:1/1:MATTER	*färgämne*:1/1
ämne:1/2:SUBSTANCE	*hornämne*:1/1
ämne:1/3:ABSTRACT	…
ämne:1/4:NOTION	*yxämne*:1/2
	fruktämne:1/2
grundämne:1/1:MATTER	…
	predikoämne:1/3
	uppsatsämne:1/3
	…
	läroämne:1/4
	skolämne:1/4

Table 1: *ämne* in Swe-S, and GLDB compounds with *ämne* as head.

Heuristic Incorporation of New Entries through Shallow Parsing

So far we have addressed the problem of the acquisition of compound nouns based on the content of the Swe-S lexicon, by applying heuristics, filters, and manual inspection, in some cases, in order to guarantee consistency. But how can we cope with the rest of the vocabulary?

Wilson and Thomas (1997:55-57) argue that one of the conditions that a semantic system should satisfy is that is should be able to account exhaustively for the whole vocabulary in the corpus, not just for a part of it. We have experimented with a corpus-based approach, using a cascaded finite-state syntactic parser (CASS-SWE), based on work done by Kokkinakis & Johansson Kokkinakis (1999), which seems a plausible way of progressively enriching the Swedish semantic resources.

An advantage of CASS-SWE is its ability to identify with high accuracy noun phrases, a property that we consider here as crucial for aiding the "discovery" of new semantic entries. Essentially the approach, which has similarities to naive clustering, is as follows. Gather large corpora (here 13 million tokens[5]), part-of-speech tag, and

[5] The parsed corpus is newspaper articles from 1997 (the so called *press97*) taken from the Swedish Language Bank: http://spraakdata.gu.se/lb/.

then parse with CASS-SWE (the parser uses part-of-speech annotated input); from the resulted analyzed forest of chunks we filter out long noun phrases, namely those containing three or more common nouns. Finally, the overlap between the nouns in the NPs produced and the entries in Swe-S is measured. If at least two of the nouns (a figure arbitrarily taken) are also entries in the Swe-S, with the same semantic class, then there is a strong indication that the rest of the nouns are co-hyponyms, and thus semantically similar with the two already encoded in Swe-S. Accordingly, we take advantage of the transitivity aspect of hyponymy, and of the fact that two lexical items *X* and *Y* are co-hyponyms if: (i) they are disjuncts and therefore complementary; and (ii) have a common superordinate, e.g. *animal* is superordinate of *cat, dog, horse* and *camel*, cf. Sanfilippo *et al.* (1999).

Similarity plays an important role in word acquistion, and preliminary results have shown that the simple overlap works fairly well for the majority of the cases examined. However, the noise which is produced can be eliminated, if the semantic tags of all the words in a phrase are compared. Caution should be taken for cases where different semantic classes[6] are involved in an enumerative NP, e.g.:

kvinnor:BIO, *barn*:BIO, *husdjur*:? *och möbler*:FURNITURE
'women, children, pets and furniture'

immiga flaskor:CONTAINER#ARTIFACT, *feta cigarrer*:?, *och tangodansande kvinnor*:BIO#SITU
'steamy bottles, fat cigars and tango-dancing women'

The unclassified *husdjur* in the first example, should not be assigned to a class BIO since there is another class involved in the same NP, namely FURNITURE. Similarly, no action should be taken in the second example, since two semantically ambiguous words with distinct classes are involved.

The best results were achieved for the semantic classes: PHENOMENA (ILLNESS and PSYCHOLOGICAL-FEATURE), OCCUPATION, ANIMAL and HUMAN (BIO, ETHNOS and OCCUPATION-AGENT). Some examples of the last mentioned class are given below, these are NPs taken from the parsed corpus. In these examples, (*) marks an original Swe-S entry, ($^+$) marks an entry incorporated through the compound analysis, (^{N}) marks a completely new entry and ($^?$) marks errors:

italienare**, finländare***, jugoslaver*^{N}*, greker**
'Italians, Finnish, Jugoslavians and Greeks'
amerikaner^N*, japaner*^{N}*, tyskar** *och italienare**
'Americans, Japanese, Germans and Italians'

jurister**, läkare***, optiker***, psykologer*^{N}*, sjukgymnaster*^{N}
'lawyers, doctors, opticians, psychologists, physiotherapists'
läkare**, psykologer** *och andra brotssutredare*^{N}
'doctors, psychologists and other crime investigators'
några läkare**, präster** *och socialarbetare*$^+$
'some doctors, priests and social workers'
samtliga politiska partier$^?$*, läkare***, jurister**
'all political parties, doctors, lawyers'
advokater**, psykiater*^{N}*, specialistläkare*$^+$ *m.m.*
'lawyers, psychiatrists, specialist doctors etc.'

Quantitative Results

Using the previously described heuristics and observations, the relatively limited inventory of semantic information in Swe-S, has been extended to a large semantic resource, appropriate for a large number of *intermediate* NLP tasks, i.e. simpler processes which are carried out to help final tasks.

Regarding the use of the compounds for extending the entries, an estimated average of 20-25 compounds per Swe-S entry has been extracted by combining information from large corpora and the GLDB. Thus, by using only 1,000 nouns we could increase the total vocabulary size to over 22,000 semantic entries. For some entries, having both concrete and abstract senses, the number of compounds extracted from large corpora could be measured into several hundreds. Table (2) shows the top-10 non-compound entries, most rich in compound variants.

Swe-S Entry	Occ.
program 'programme, program'	469
arbete 'work, employment'	402
chef 'chief'	390
bok 'book'	357
verksamhet 'activity, operation'	299
skola 'school'	275
man 'man'	273
rum 'room, space'	244
kort 'card, photo'	231
bolag 'company'	217

Table 2: Swe-S entries richest in compound variants

Regarding now the shallow parsing approach of a l3 million corpus, over 15,600 NPs could be extracted, having the content we were interested in, namely over three common nouns. Approximately 3,000 new noun entries to the Swe-S could be identified without any further processing (bootstrapping the compound analysis). However, as mentioned in the previous section, some noise was produced and for this reason we do not use these new nouns for the semantic annotation discussed in the next section, until we find more reliable ways to eliminate the limited number of errors produced.

Annotating with Swe-S (Semantic Tagging)

Semantic tagging is appealing since it is believed to contribute to the improvement of the performances and robustness of NLP systems, *cf.* Resnik & Yarowsky

[6] Explications of some less obvious semantic classes, used here and in following examples: BIO refers to "any classification of human beings (groups or individuals) according to biological characterstics like age, sex, etc"; SITU refers to "individuals or groups of humans identified according to an accidental behavioural or punctual criterion"; ETHNOS refers to "designation of humans according to ethnological criteria", while OCCUPATION-AGENT refers to "individuals or groups of humans identified according to a role in professional, social or religious disciplines".

(1997). The appropriate content from the core Swe-S, i.e. "semantic class", "domain" and "template type" information, has been extracted and implemented as finite-state machines suitable for semantic tagging, the case of assigning semantic categories or clusters of semantically related concepts to words. These machines are then applied sequentially to lemmatized textual data resulting in all possible annotation for the tokens matched.

Testing was performed using 1,800 nouns from the Swe-S, while approximately 150 of those could be ambiguous, in the sense that more than one semantic label, class, domain and template, could be asssociated with a single token. For instance, the Swedish noun *administration* 'administration' is semantically classified for four different semantic classes: AGENCY, FUNCTIONAL-SPACE, HUMAN and OPERATION, while the noun *affär* 'shop, business, affair' is classified for: FUNCTIONAL-SPACE, OPERATION, STATE and EVENT.

Supervised Learning

We adopted Machine Learning (ML), particularly Memory Based Learning, for the disambiguation of the semantic annotation of text samples.

Memory Based Learning

Memory-Based Learning (MBL) is a supervised, inductive, classification-based method originating from the field of machine learning (ML), Mitchel (1997). MBL has several practical advantages, such as: (i) it has produced state-of-the-art results in many natural ambiguity problems (*cf.* Cardie & Mooney (1999)); (ii) the MBL method is not sensitive to sparse or low-frequency data, as low-frequency cases are not discarded but are kept in memory, hence, useful information can also be extrapolated from them; and (iii) fast learning and incremental learning; new instances can be added to the memory, improving the performance of the system. The software used for the experiments with the Swedish data has been developed at the University of Tilburg, by Daelemans *et al.* (1999).

MBL is closely based on the assumption that "performance in cognitive tasks is based on reasoning on the basis of similarity of new situations to stored representations of earlier experiences", Daelemans *et al.* (1999). An MBL system consists of two components: a learning component, which is memory-based, adding training instances to memory, and a performance component, in which the product of the learning component is used for performing the classification of the input.

Training Material

It is rather difficult to give an exact number of examples required for an adequate description of noun senses. Intelligent example selection for supervised learning is an important issue in ML, an issue that we have not explored. However, from the (human) lexicographical point of view, an experienced scholar would need, roughly, a hundred arbitrarily chosen excerpts for each word in order to cover the majority of sense distinctions (Jerker Järborg personal communication). For a machine, that figure should be higher, although we have not empirically tested the validity of this statement.

We have automatically created large training non-lemmatized data, taken from concordances and then manually classified the training instances. The deliberate choice of non-lemmatized material should be emphasized here, as our experiments proved that noun morphology supports sense disambiguation, both for compound and non-compound forms in Swedish.

For instance, plural forms of *finska* or *tyska*, 'Finnish, German', refer almost exclusively to ETHNOS (denoting a person) while the base form is ambiguous between ETHNOS and ABSTRACT (denoting either the person or language). Likewise, plural forms of *begåvning* 'talented (person), talent' refer almost exclusively to SITU, while its base form refers to PSYCHOLOGICAL-FEATURE.

For the training and test instances we organized the near context of the ambiguous semantic entries into fixed-length vectors of symbolic *n* feature-value pairs (in the experiments in this paper n=12) which consist of the left and right context of the word under investigation, its part-of-speech and its byte-offset in the discourse, and a field containing the classification of that particular feature-value vector. Unknown features are marked with a question mark '?' while long context is truncated. Moreover, we took advantage of the syntactic examples in GLDB, given for almost every lemma in the database, and in this way we could complement the training material automatically with already classified training instances. This last point can be illustrated by the use of two syntactic examples provided by the GLDB for the noun *medicin* 'medicine'. Since these are already disambiguated, designated by their sense number, they can be directly mapped onto the respective Swe-S semantic classes for that particular word:

GLDB: *medicin:1:studera medicin*
MBL: byte-offs noun ? ? ? *studera medicin* ? ? ? ?
OCCUPATION
'medicine:sense1:study medicine'

GLDB: *medicin:2:skriva ut recept på en bra medicin*
MBL: byte-offs noun *recept på en bra medicin* ? ? ? ?
SUBSTANCE
'medicine:sense2:write a prescription on a good medicine'

During classification an unseen example *X*, a test instance, is presented to the system and a distance metric Δ between the instances in the memory *Y* and *X* is calculated, $\Delta(X,Y)$. Various implemented algorithms (variants of the k-nearest neighbour algorithm) try to find the nearest training instance for *X* and create a class as prediction for the class of the test instance.

Results

At present, the standard for calculation of sense disambiguation algorithms is the "exact match" (or accuracy) criterion. Specifically for ML, our goal is to perform significantly better than the most-frequent-semantic classifier to be worthy of serious consideration.

Table (3) summarized the results for few ambiguous cases examined. In every case we try to improve the baseline for every semantic entry we want to disambiguate. Here by *baseline* is meant the most frequent class attached to an ambiguous token in the test sample.

Swe-S	(Swe-S) Class	Tr. Data	Base-line	Acc.
administration 'administration'	AGENCY-39 HUMAN-18 OPERATION-81 FUNCT-SPACE-5	143	56.6%	76%
affär 'shop, affair, business'	EVENT-102 OPERATION-175 STATE-2 FUNCT-SPACE-83	362	48.3%	92%
danska 'Danish'	ETHNOS-258 ABSTRACT-32	290	88.9%	88%
klyfta 'segment, cleft, rift'	PHENOMEN.-67 FORM-33 ALTERNATION-3	119	56.3%	88%
medicin 'medicine'	SUBSTANCE-380 OCCUPATION-103 SUBST/OCCUP.-6	489	77.7%	72%
område 'area, zone, field'	LOCATION-168 ABSTRACT-151	319	52.6%	84%
vatten 'water'	SUBSTANCE-396 SUBST/LOC.-110 LOCATION-61	567	69,8%	100%
teater 'theatre, play-acting'	ABSTRACT-105 AGENCY-103 FUNC-SPACE-102 HUMAN-12 ACTIVITY-7	329	31.9%	85%

Table 3: Data used by MBL for semantic disambiguation (*Tr. Data*: amount of training data, *Acc.*: accuracy based on the MBL approach)

Our experiments using the MBL approach returned 84.8% correct disambiguation, tested on 25 ambiguous entries (with 20-25 test instances in each case), with an average baseline of 69.4%.

Usability of the SIMPLE Model

In this section, we are going to reflect on the usability of the SIMPLE model for different NLP tasks, which require access to semantic information. Many NLP applications can be actively supported by the SIMPLE lexicon which offers multiple access points to the semantic data. 10,000 word senses can be accessed either directly, or by means of selective information searches starting with 139 ontological categories provided by the SIMPLE ontology, 95 semantic class categories and to 364 domain specifications. Since the two first capture somewhat different aspects of word meaning for a number of cases, the double ontological specifications not only provide more precise information, but also increase the granularity of semantic description.

The ontological information cluster can be extended with information on domains. The domain information, indispensable for text-recognition tasks can support disambiguation of senses with identical ontological clusters. For example, the word *grad* 'degree, grade' has nine senses assigned, and four of these denote different units of measurment representative for domains such as GEOMETRY, EARTH-SCIENCES, TYPOGRAPHY and METEOROLOGY. Since those four display identical ontological categorization, the domain information supports disambiguation in a relevant way. In consequence, a tripartite cluster including both ontological and domain information seems to be preferred. The explicit specification of domain information in the SIMPLE lexicon makes it possible to generate domain-based sublexicons, which are basic for text-recognition tasks.

The attempt to harmonize the encoding of data makes it possible to multilink the SIMPLE lexicons for different languages, which is substantial for building the lexicon modules for machine-aided-translation.

Since the content of the Swe-S lexicon is linked to the GLDB database, the information exchange can proceed in two directions, which promotes development of both resources. These two resources describe and formalize lexical information concerning a word's morphology, syntax and semantics, which is a prerequisite for advanced NLP tasks. As was already hinted, the SIMPLE project has aimed at harmonization of lexical resources by using a common lexicon model and formalism for 12 EU languages. This initiative has opened new prospects for further developments within the language engineering field.

Conclusions and Further Research

This paper has discussed means to automatically extend the lexical inventory of the Swe-S semantic lexicon, by profiting from the productive compounding characteristic for Swedish, the semantic similarity in the enumerative noun phrases, by accessing corpora both in raw and parsed form, and the morphological, syntactic and semantic content of GLDB. Using a combination of all the available data, a relatively limited inventory of semantic information, such as the Swe-S, can be extended to a large semantic resource appropriate for a large number of intermediate NLP tasks. Moreover, its compatibility with the manually developed Swe-S lexicon, can be guaranteed and its high quality maintained, as we applied heuristics that do not try to overproduce semantically anomalous entries. We have also used the Swe-S resource for semantic annotation of texts, while for the disambiguation, we employed Machine Learning techniques, supported by manually created large portions of training data for a small number of ambiguous semantic entries. Work within the SIMPLE project was still in progress when writing this paper, so a future task would be to extend the rest of the material using the same methodology, and even to devise better ways to eliminate the noise produced by the syntactic parsing. Reliable extraction of similar words from text corpora opens up many exciting opportunities for further linguistic analysis.

Acknowledgements

We thank three anonymous reviewers for some useful comments on a previous draft. The first author is also indebted to the "Birgit & Gad Rausings" foundation for providing financial support for the participation at the conference.

References

Blåberg, O. (1988). *A Study of Swedish Compounds*. Report 29, General Linguistics, Umeå university, Sweden

Calzolari, N. (1999). *SIMPLE: Harmonised Semantic Lexicons for the European Languages*. In Proceedings of the SIGLEX-99 Workshop: "Standardizing Lexical Resources", Maryland, USA

Cardie, C. and Mooney, R.J. (1999). Guest Editors' Introduction: Machine Learning and Natural Language. In *Journal of Machine Learning, Special Issue on Natural Language Learning*, Vol. 34, pp. 1-5, Kluwer AP

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A. (1999). *TiMBL: Tilburg Memory Based Learner, version 2.0, Reference Guide*. ILK Technical Report 99-01. Paper available from: http://ilk.kub.nl/~ilk/papers/ilk9901.ps.gz

Dorr, B. and Jones, D. (1996). *Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision*. In Proceedings of the SIGLEX Workshop "Breadth and Depth of Semantic Lexicons", pp. 42-50, Santa Cruz, California, USA

Hearst, M.A. and Schütze, H. (1996). Customizing a Lexicon to Better Suit a Computational Task. In *Corpus Processing for Lexical Acquistion*, pp. 77-94, Boguraev B. and Pustejovsky J. (eds.). MIT Press

Kokkinakis, D. and Johansson Kokkinakis, S. (1999). *A Cascaded Finite-State Parser for Syntactic Analysis of Swedish*. In Proceedings of the 9th EACL, pp. 245-248, Bergen, Norway. Paper available from: http://svenska.gu.se/~svedk/publics/eaclKokk.ps

Lenci, A. et al., (1998). *SIMPLE WP2, Linguistics Specifications*. Deliverable 2.1, Pisa

McKeown, K. and Hatzivassiloglou, V. (1993). *Augmenting Lexicons Automatically: Clustering Semantically Related Adjectives*. In Proceedings of the ARPA HLT Workshop, pp. 272-277, Princeton, NJ

Mitchell, T. M. (1997). *Machine Learning*. Series on Computer Science, McGraw-Hill

NEO, (1996). *Nationalencyklopedins ordbok*. Volumes 1-3, Språkdata & Bra Böcker AB

Pedersen, B.S. and Keson, B. (1999). *SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Some Examples of Danish Concrete Nouns*. In Proceedings of the SIGLEX-99 Workshop: "Standardizing Lexical Resources", Maryland, USA

Resnik P. and Yarowsky D. (1997). *A Perspective on Word Sense Disambiguation, Methods and their Evaluation*. In Proceedings of the Workshop: "Tagging Text with Lexical Semantics. Why, What and How?", pp. 79-86, Washington D.C., USA

Roventini, A., Peters, C., Calzolari, N. and Bertagna, F. (1998). *Building a Semantic Network for Italian Using Existing Lexical Resources*. In Proceedings of the 1st LREC, Vol. 1, pp. 377-383, Granada, Spain

Sanfilippo, A. *et al.* (1999). *Preliminary Recommend-ations on Lexical Semantic Encoding*. EAGLES LE3-4244, Draft version

SAOL, (1998). *Svenska Akademiens Ordlista över Svenska Språket* (The Swedish Academy Word-List). Norstedts Förlag & Svenska Akademien

SO, (1992). *Svenska Ord*. Statens Skolverk, Nordstedts Förlag

Takunaga, T., Fujii, A, Iwayama, M., Sakurai, N. and Tanaka, H. (1997). *Extending a Thesaurus by Classifying Words*. In Proceedings of the Workshop: "Automatic Information Extraction and Building of Lexical Semantic Resources", Vossen P., Adriaens G., Calzolari N., Sanfilippo A. and Wilks Y. (eds), pp. 16-21, Madrid, Spain

Viegas, E., Ruelas, A., Beale, S. and Nirenburg, S. (1998). *Extending a Core lexicon Using On-Line Language Resources with Savoir-Faire*. In Proceedings of the 1st LREC, Vol. 1, pp. 97-104, Granada, Spain