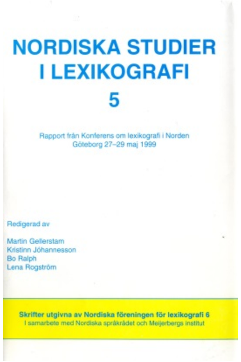


NORDISKA STUDIER
I LEXIKOGRAFI
5

Rapport från Konferens om lexikografi i Norden
Göteborg 27–29 maj 1999

Redigerad av

Martin Gellerstam
Kristinn Jóhannesson
Bo Ralph
Lena Rogström

Skrifter utgivna av Nordiska föreningen för lexikografi 6
I samarbete med Nordiska språkrådet och Meijerbergs institut

Dimitrios Kokkinakis
Sofie Johansson Kokkinakis

Sense-Tagging at the Cycle-Level Using GLDB

This report describes a large-scale attempt to identify automatically the appropriate sense for content words taken from Swedish open-source texts. Sense-tagging, 'the process of assigning the appropriate sense from some kind of lexicon to the (content) words in a text', is a difficult and demanding task in Natural Language Processing and researchers have been engaged in finding a suitable solution to the problem for a very long time. The usefulness of automatically assigning each word in unrestricted text with its most likely sense is necessary for a great spectrum of applications. The sense-tagger described here has been tested both on a random sample of content words, as well as on a large population of a single ambiguous entry. In the first case, the achieved precision was 84,21%, and in the second 82,75% respectively. Evaluation was made against manually sense-annotated texts.

1. Introduction

One of the many problems encountered in Natural Language Processing is that of semantic lexical ambiguity. This means that deciding which meaning of a word is intended in a given utterance or discourse is a very difficult task, that humans usually perform without even conciously noticing that the ambiguity exists. While a native speaker of Swedish can almost immediately recognize that in the following four examples the verb *handla* refers to four different meanings, i.e. 'to take action', 'to trade', 'to buy', 'to deal', the same task in computer processing is a major headache:

(1) *Polisen* ***handlade*** *snabbt denna gång.*
'The police acted quickly this time.'
(2) *EU har åter börjat* ***handla*** *med Kina.*
'EU has started trading with China again.'
(3) *John* ***handlade*** *mat för 1000 dollar.*
'John bought food for 1000 dollars.'
(4) *Filmen* ***handlade*** *om en ung mans väg till framgång.*
'The film was about a young man's way to success.'

The idea of performing sense disambiguation is a controversial matter in many respects, and it has been discussed and sometimes criticized with respect to whether *how* and *if* it can possibly be done. Criticism has been directed against several attempts at automatic sense disambiguation, and there are no simple answers to the otherwise well-justified questions associated with the issue whether it is feasible or not to make clear sense distinctions. Some of the criticism is based on the term 'sense' itself, which is not a well-defined concept; problems referring to the fact that humans cannot agree on what sense is appropriate for the words in a given sentence; sense distinctions are interpreted differently by different researchers, following different approaches to the disambiguation problem; and finally, that dictionaries differ substantially regarding the different sets of senses for the same word. Despite the criticism, we regard sense-tagging as a very important process and component within a wider and deeper text-processing architecture.

2. Sense and Semantic Tagging

Many words in the dictionaries have multiple senses or meanings, while, when a word is actually used in a context, just one of these meanings generally applies. By the term *sense*, we here mean *dictionary sense*. For instance, *development* may be a highly ambiguous word in English, but for photographers it refers unambiguously to *processing a film*, and for an architect it refers to *building*. Sense-tagging should not be confused with *semantic tagging*, which is a more general case, in which the labels assigned to the words in a text are broad semantic categories, or clusters of semantically related concepts. Semantic categories may be labels of the form ANIMATE, ARTIFACT, LOCATION or HUMAN; or WordNet synsets (Fellbaum 1998), such as *Life and Living Things* and *Food, Drink and Farming*. Despite their subtle differences, both sense and semantic tagging aim at the resolution of lexical ambiguity, either on a small or large scale. Furthermore, the more general term of

Word Sense Disambiguation or WSD, is used in the context of both. Both terms, *sense-tagging* and *WSD*, will be used interchangeably.

WSD tries to solve lexical ambiguity, which in turn is closely related to two lexical semantics concepts, that of *polysemy* (related word senses) and *homonymy* (unrelated word senses). There is no clearcut border between these two concepts. From the point of view of WSD the difference between these two concepts is not a controversial issue, lexical ambiguity, in the context of automatic means of WSD, refers to both.

3. Background

The different approaches to lexical disambiguation that will be discussed here are classified according to the major source of information that researchers have used for the WSD task. Different classification schemes can be found in Wilson & Thomas (1997), in which they distinguish between manual, computer-assisted and fully automatic methods to WSD; Fujii (1998), in which he distinguishes between qualitative and quantitative approaches; and Sanfilippo *et al.* (1998 §5.3.2), in which they distinguish between knowledge-based, corpus-based and hybrid approaches.

3.1. WSD Using an Explicit Dictionary

Kelly & Stone (1975) manually developed a dictionary for approximately 2,000 words by studying senses from a word corpus of half a million words and writing disambiguation rules for each multi-sense word. For their help they had key-word-in-context (KWIC) concordances and the 1966 *Random House Unabridged Dictionary*. Their work was labour-intensive and manual, and the local context was their main source of information used. The manually-written disambiguation rules, each corresponding to a sense in the dictionary, were used by an algorithm achieving 90% accuracy.

The use of standard dictionaries is contributed to Lesk (1986). He was the first to suggest the use of dictionary definition overlap for WSD. Lesk proposed that a sentence could be disambiguated relative to a dictionary by choosing the configuration of senses that maximizes the number of words which are common to the textual definitions and the context of the word to be disambiguated. The quality of the results on his experiments lie within the 50–70% correct sense distinction.

Cowie *et al.* (1992), applied the simulated annealing technique to WSD. They were the first research team that used this method for WSD, in conjunction with the LDOCE, reporting 72% correct assignment of senses.

Wilks & Stevenson (1998) also used LDOCE for disambiguation. Their approach was based on combining different knowledge sources for achieving qualitatively better results, than merely using the definitions. The algorithmic details behind their high figures on precision rely on the use of an optimized version of the simulated annealing technique. Wilks & Stevenson are one of the very few research teams that have attempted sense disambiguation on all content words in a text, achieving 94% correct sense assignment.

3.2. WSD Using Thesauri and Ontologies

A few knowledge bases often discussed in connection with semantic disambiguation, such as the *WordNet* and the *Roget's Thesaurus*, The Princeton WordNet, Miller *et al.* (1990), and the *EuroWordNet*, Peters *et al.* (1998), are the most commonly used networks for disambiguation considering the relevant bibliography.

WordNet has a predominent position since it is publicly available, and has been extensively studied for quite some time by a number of different research teams. Miller *et al.* (1993) used WordNet for linking content words from a text to their appropriate sense in the lexicon. This was viewed by the Miller group either as a corpus, in which words have been tagged semantically, or as a lexicon in which example sentences can be found for many definitions.

3.3. WSD Using Information from Corpus

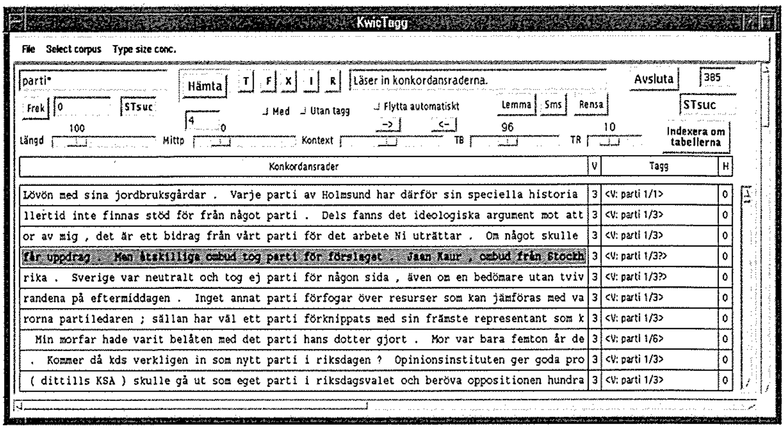
Some of the reasons in favour of using corpora for the WSD, and not using a dictionary or a semantic net, are the following: that copyright constraints are usually associated with the dictionaries; that dictionary descriptions are just a static view of a language in a particular time frame, while language is a dynamic system in constant change that can be better and more easily captured by monitoring text corpora; and finally, that imperfections and incomplete coverage are usually tied to the lexical resources.

The approaches can be divided into three different methods. (i) Supervised methods use as their primary source of information a *disambiguated corpus*. The annotated corpus is then used for the supervision and induction of rules, which are fed into stochastic models, and which can predict the correct sense of words in new contexts (*cf.* Yarowsky 1994). (ii) Restricted Supervised Methods based on bilingual texts, usually aligned; see Brown *et al.* (1991). (iii) Unsupervised Methods rely on raw, *unannotated corpus* and, in few cases, on the content of a machine-readable dictionary. One of the motivations behind the use of these methods is the fact that it is very difficult to find domain-dependent lexical knowledge sources. On the other hand, the major drawback of using unsupervised methods is that no fine-grained distinctions between senses can be made.

3.4. WSD and Swedish

The only known attempt to word-sense disambiguate Swedish on a large scale is a project undertaken at Språkdata (financed by '*The Swedish Council for Research in the Humanities and*

Social Sciences' (HSFR)). The project is entitled *Lexikalisk betydelse och användningsbetydelse (SemTag)*, i.e. 'Lexical Sense and Sense in Context'. The sense annotation is carried out interactively through a concordance-based interface, interacting with GLDB, figure 1. All words are sense-tagged. The corpus used in SemTag is the SUC corpus (Ejerhed *et al.* 1992).



Läser in konkordansraderna.

Konkordansrader	V	Tagg	H
Lövön med sina jordbruksgårdar . Varje parti av Holmsund har därför sin speciella historia	3	<V: parti 1/1>	0
llertid inte finnas stöd för från något parti . Dels fanns det ideologiska argument mot att	3	<V: parti 1/3>	0
or av mig , det är ett bidrag från vårt parti för det arbete Ni uträttar . Om något skulle	3	<V: parti 1/3>	0
får uppdrag . Men åtskilliga ombud tog parti för förslaget . Jaan Kaur , ombud från Stockh	3	<V: parti 1/3?>	0
rika . Sverige var neutralt och tog ej parti för någon sida , även om en bedömare utan tviv	3	<V: parti 1/3?>	0
randena på eftermiddagen . Inget annat parti förfogar över resurser som kan jämföras med va	3	<V: parti 1/3>	0
rorna partiledaren ; sällan har väl ett parti förknippats med sin främste representant som k	3	<V: parti 1/3>	0
Min morfar hade varit belåten med det parti hans dotter gjort . Mor var bara femton år de	3	<V: parti 1/6>	0
. Kommer då kds verkligen in som nytt parti i riksdagen ? Opinionsinstituten ger goda pro	3	<V: parti 1/3>	0
(dittills KSA) skulle gå ut som eget parti i riksdagsvalet och beröva oppositionen hundra	3	<V: parti 1/3>	0

FIGURE 1. *The KwicTagg Interface.*

4. The Usefulness of WSD and Some Potential Applications

The lack of high quality as well as the slow progress within MT has been blamed on word-sense ambiguity. It is wellknown that a single non-ambiguous word in a source language might be translated by a number of different words or expressions in the target language (translational or transfer ambiguity), and a source word can have more than one sense (monolingual ambiguity) (Hutchins & Somers 1992).

In IR it is necessary to disambiguate content words in the queries sent to knowledge bases or free-text search; it is also useful for the purposes of text categorization or indexing, and thus for deciding whether a document is relevant for a particular application or not, by reducing the noise produced due to polysemy; *cf.* Schütze & Pedersen (1995). Sense-tagging can improve the performance of Information Extraction (IE). Despite the fact that in IE domain-specific ontologies are already employed, methods for large-scale WSD might improve the IE system's performance even more; *cf.* Kilgarriff (1997a), Chai & Bierman (1997). This can be accomplished by triggering patterns to perform extraction only of relevant senses.

Corpus-based lexicography would benefit from automatic means of identifying the appropriate senses of the words in large corpora for the sake of facilitating and qualitatively improving the information already present in dictionaries. This could be accomplished, by sorting out thousands of concordance lines of irrelevant text with senses not valuable for a specific lexicographic assigment, or by arranging the definitions in the lexicon according to frequency of use, in such a way that the most common senses preceed the least common. This is of course a matter dependent on the size and the representativity of the corpus we use, but it is not totally unfeasible.

5. The Critics

The identification of the right meaning of a word, regardless if it is taken from a dictionary or a semantic net is controversial in many respects. Some of the criticism of WSD is concentrated onto three points. First there is the criticism related to the fact that humans cannot agree on which sense is appropriate for the words in a given sentence; *cf.* Kilgarriff (1997b) for a survey. Then there is the fact, that sense distinctions are interpreted differently by different researchers, following different approaches to the disambiguation problem. In this respect, researchers are divided into those who use coarse-grained sense distinctions, or 'lumpers', and those who use fine-grained sense distinctions, or 'splitters'; for an interesting discussion on the matter see Kilgarriff (1997b) and Wilks (1997) published in the same volume. Finally, the fact that different dictionaries differ substantially regarding the different sets of senses that they associate with the same word complicates the problem even more.

This last claim is also strengthened by the fact that dictionaries tend to be incomplete, both with respect to coverage and content; *cf.* Boguraev (1995), for a discussion of the use of dictionaries in computational linguistic research. Finally, Wilks (1995) discusses in detail the question whether it is possible to sense-tag on a large-scale and systematically, or not. He examines and attacks two extreme views. According to Wilks, these two views are both misleading claims and are widely believed, though not simultaneously: 'sense-tagging has been solved' or 'it cannot be done at all'. His conclusion is that the field of sense-tagging is still open to further development and that dictionary-based and (unannotated) corpora-based efforts are equally useful for practical applications.

6. The Chosen Approach

The method chosen here is dictionary-driven and relies on an existing lexical resource for modern Swedish, structured as a relational database, i.e. the *Gothenburg Lexical Database* or GLDB. The content in GLDB has been used for the production of standard contemporary Swedish lexica, for instance the three-volume *Dictionary of the National Encyclopedia*, NEO (1996). The method is based on the *simulated annealing* technique (SA), which has been used for quite some time for sense disambiguation of English words, using a standard machine-readable dictionary, (namely LDOCE).

The idea behind SA is to perform enough exploration of the whole search space early on, so that the final solution is relatively intensive to the starting state. SA is often used to solve problems in which the number of moves from a given state is very large, and it has been applied to the *travelling salesman problem*, in which space is the different paths through the cities that the salesman must visit in an optimal way without visiting the same city twice.

7. Knowledge Sources

7.1. The Structure of GLDB

The work on the GLDB was started 25 years ago by Professor Sture Allén and his research group at Språkdata. The underlying linguistic, theoretical model of GLDB is the *lemma-lexeme* model, Allén (1981). The lemma comprises formal data such as part of speech and inflection(s). The lexemes (or numbered senses) are in turn divided into two categories, a compulsory kernel sense and a non-compulsory set of one or more sub-senses, called the *cycles*. GLDB contains a description of 61,050 lemmas, and 67,785 senses, while 19,082 lemmas contain valency information. GLDB has the advantage of covering the 'whole' language and not just a small subset. A particularly interesting feature of GLDB is the fact that metaphors, though not *dead* ones, are encoded as separate sub-senses of a lemma, usually preceded by the key-word *överfört*, i.e. 'transferred'. A number of printed Swedish monolingual, defining dictionaries have been generated from the GLDB; see Malmgren (1992).

7.2. The Information Used

For enhancing the performance of the sense-tagger we must be able to use as much as possible of the available information in GLDB. The following information seemed an adequate, necessary starting point for the sense-tagging: Definitions and definition extensions; Morphological examples; Syntactic examples; Deverbal Nouns and Valencies. Compared to LDOCE, the GLDB's definitions are much shorter, and the head-word entries usually contain fewer example samples.

8. Data Preparation

For enhancing the lexical disambiguation result using the available resources, it is necessary to perform pre-processing both in the resources and the text to be sense-tagged. This is motivated by the fact that by making certain normalizations and simplifications in the resources, such as lemmatization, we contribute to the production of qualitatively better results.

There have been several reasons that motivated the use of pre-processing. Some of these have been: (i) the fact that not all entries in the GLDB are relevant during the sense annotation of a "normal-length" newspaper text. This means that we extract only a subset of the GLDB depending on the unique occurrences of word forms in the text to be processed; and (ii) not all entries in GLDB consist of a single entry form (lemma). This is the case with 914 phrasal verbs consisting of two or three units, 115 multi-word nouns, and 5 multi-word adjectives. Moreover, it is absolutely necessary to reduce the complexity of the matching process by operating onto base forms by conducting lemmatization, both in the text and the information in the lexical entries, especially the GLDB definitions. Using base forms reduces the complexity and time required for the calculation of the overlap between the resources. Part-of-speech tagging is also an important aspect, since it eliminates *accidental* homography (*länkar* 'links, chains, guides' as verb, *länkar* 'links, chains' as noun). Another aspect in favour of the pre-processing required has to do with the productivity of the Swedish language in creating new compound words, especially nouns. New content words, not present in the lexicon, must be identified and possibly assigned a sense, based on entries with similar defining criteria, i.e. in this study, by using the definition of the last part of the compound.

8.1. Multi-Word Units (in Text and GLDB)

Multi-word expressions cannot be properly understood if they are not recognized as unique units. There are a number of different types of units recognized: *phrasal verbs*, *idioms*, *lexical* and *grammatical collocations*.

There are 914 phrasal verbs explicitly given a separate entry in GLDB, such as *brinna av:1/1* 'to go off, to explode', *dela ut:1/1* 'to distribute' and *ställa ut:1/1,1/2* 'to exhibit'; in 223 of these, the Swedish third-person reflexive pronoun *sig* 'himself/herself/ themselves' is the last part of the unit; i.e. *bekanta sig:1/1* 'to acquaint oneself', *dra på sig:1/1* 'to put (pull) on' and *ställa in sig:1/1* 'to intend to'.

There are two issues which need special attention with respect to phrasal verbs in Swedish. One is that they can be discontinuous in the text, and the other that in some cases it is only the intonation and/or extended context that can decide whether a verb is a phrasal or not. Consider for instance the verb *köra* 'to drive, to run, to force, to convey' combined with the token *på* 'on/at/into', which in the first example below is a particle, while in the second a preposition: (i) *marknader kör på som om inget hänt* 'markets keep on as if nothing has happened'; (ii) *att köra bil på nätter*, 'to drive a car at night' .

8.2. Known Compounds, Morphological Examples (in GLDB)

The morphological examples in GLDB are simply compound nouns in which the lemma-entry participates as the main information carrier of the compound. For example, the first sub-sense of the second sense of the noun entry *avgång* 'wastage, retirement, resignation' contains two morphological examples (compounds) :

> avgång 1/2/a: **avgång**sbetyg 'leaving certificate', **avgång**sklass 'final class'

All the compounds have been automatically split into their respective parts, by identifying the lemma or part of it in the compound. The morphological examples of the noun *avgång* are actually used by the sense-tagger as:

avgång 1/2/a: avgångsbetyg, betyg, avgångsklass, klass

After the automatic segmentation, the split compounds were automatically post-validated for erroneous splitting, or for completing the produced segments. We anticipate that a small number of segmentation errors might be present in the morphological examples.

8.3. Unknown Compounds (in Text)

For the cases where no entries in GLDB cover these compounds, the application of heuristic compound segmentation is performed.

Previous attempts to segment compounds without the help of a lexicon are described in Brodda (1979), and Klenk & Langer (1989). The segmentation algorithm we use proceeds by scanning unidentified word forms from left to right, trying to identify grapheme combinations which feel *unnatural* or simply unallowed as non-compound forms in the Swedish language, and which carry information of potential token boundaries. The heuristic method behind the segmentation of compounds in our method is based on producing 3-gram and 4-gram character sequences of several hundreds of non-compound lemmas, and then generating 3-grams and 4-grams that are not part of the lists produced, some manual adjustments being also imposed. Furthermore, 4-grams with 4, 3 or 2 vowels, and 3-grams with 3 or 2 vowels were not used, except in the cases with two similar consecutive vowels, such as *ii* and *ee*. Ambiguities are unavoidable, although the heuristic segmentation has been evaluated for precision, and over 90% accuracy was measured, a more thorough evaluation is beyond the scope of this report, and that is why we concentrated on precision alone, which is easier to estimate than coverage.

Consider for instance the following examples:

TABLE 1. *Splitting of* Unnatural *Grapheme Combinations.*

”Unnatural” *Grapheme* *Combinations*	*Splitting* Point *i.e. (’|’)*	*Examples*
ivb	iv|b	skr**iv|b**ord (writing desk), kollekt**iv|b**oende (collective housing)
ktm	kt|m	konta**kt|m**an (contact person), ma**kt|m**issbruk (power abuse)
ksf	ks|f	olyc**ks|f**all (accident), Danmar**ks|f**ärjan (Denmark ferry)
tss	ts|s	rät**ts|s**alen (court room), arbetslöshe**ts|s**ifrorna (unemployment figures)
gsk	gs|k	växlin**gs|k**ontor (exchange office), tillverknin**gs|k**apacitet (manufacturing capacity)
ngss	ngs|s	forskni**ngs|s**kola (research school), bantni**ngs|s**tudie (slimming study)

8.4. Idiomatic Expressions (in Text and GLDB)

Idiomatic expressions are recognized and are not treated for WSD. A list of over 4,900 idiomatic expressions has been extracted from GLDB and implemented as a finite-state recognition machine, used during the pre-processing stage. The original list of all the idioms in GLDB has been expanded to about 6,500, since we had to cope with the expansion of parenthetic and shorthand information for different variations of an idiom. For instance:

> GLDB: *ben: (han är) bara skinn och &*, i.e. ’(he is) nothing but skin and bone’, has been encoded as:
> (i’) *han är bara skinn och ben*
> (i’’) *skinn och ben*

8.5. Deverbal Nouns (in Text and GLDB)

Nominalized verbs are yet another problematic set of items that have to be processed in a specific manner, since these are not treated as separate lemmas in the database, but are (usually) encoded under the verb entry. In Swedish, deverbal nouns are usually constructed by means of the morphemes: *~(n)ing*, *~ande*, *~ende* and *~nde*. Some of these nouns are very productive, while some are only theoretically possible or less frequent. The method we chose to deal with these cases is first to identify them in the text, and then mark them accordingly, so that they will be analyzed during the WSD procedure, depending on the corresponding verbal entries. Notice, however, that *there are* separate lemmas in GLDB ending in: *~(n)ing*, *~ande*, *~ende* and *~nde*, that originate from verbs. The criterion for having such entries in GLDB is based on the fact that these nouns have very specific, and concrete meanings, (e.g. *kaffeservering:1/1* 'café, coffee-room', *pristävling:1/1* 'prize competition'), and no longer denote a verbal action of some kind.

8.6. Part-of-Speech Tagging and Lemmatization (in Text and GLDB)

The application of part-of-speech tagging is carried out for sorting out non-relevant definitions of homograph tokens during the extraction of a relevant subset of the GLDB with respect to the text that is analyzed, and for making easier the lemmatization, which is applied to tagged tokens. Brill's (1994) rule-based tagger, is used for part-of-speech tagging, trained on Swedish texts; Johansson Kokkinakis & Kokkinakis (1996).

From a computational and performance point of view, it is attractive and desirable to operate on the roots (or stems) of words, both in the text and the lexical resources. For that reason we hâve lemmatized the definitions in GLDB and we use the lemmatizer prior to processing a text by the sense-tagger. The lemmatizer is applied to part-of-speech annotated texts, which enhances the quality of the stemmed results.

9. Computer Processing and GLDB – some Problems

The experience gained through working with GLDB and the sense-tagging task has proved that GLDB *is* an adequate resource for the WSD process, although the evaluation needs to be extended over a larger sample of the language than the one we have used so far (see next section). Nevertheless, there were a few occasions where the structure of GLDB lacked consistency. Of course, we do not disregard the fact that the GLDB's organizational structure is made *by* humans *for* humans and not for any particular computer processing, and that explicit encoding of all occasional word forms (especially compound nouns), or phrasal verbs would have led to an unmanageable explosion of the entries in GLDB. A final point regards the way the valencies and the deverbal nouns are described within the different sub-senses (cycles) in the database. The problem associated with this issue is that this information is *not* explicitly denoted for the individual cycles. The valency and deverbal information descriptions are given only for the lexemes, and implicitly for all the different sub-senses of a sense. This encoding methodology makes the accurate identification of the sub-senses much more difficult.

10. Evaluation

For the evaluation part of this study we have manually sense-tagged different text samples. The evaluation is performed after the texts have been tokenized, the idioms, deverbal nouns and multi-word expressions identified, then part-of-speech tagged, lemmatized, and content words identified and marked appropriately. The manual annotation has proven to be a very labour-intensive but challenging and necessary process. Two experiments were conducted. In the first case, we randomly extracted short newspaper samples and sense-tagged the verbs, nouns and adjectives. The evaluation was carried out in two different ways: (i) WSD *only* by using the definition and definition extensions of the lexical entries in the GLDB; and (ii) WSD by using the definitions supplemented with definition extensions, morphological and syntactic examples, and even the typical prepositions, (valencies), for each entry. In the second case-study, we extracted 60 concordance lines in which a single ambiguous verb, the verb *handla* 'to deal, to trade, to take action, to buy', was the object of the investigation, and was then sense-tagged.

10.1. Results

Table (2) shows the results of the evaluation, using only the definitions and definition extensions, and the definitions and definition-extensions extended with other knowledge, such as the morphological and syntactic examples (All Material). Here the metric *precision* is defined as the percentage of the sense-tagged words that were found correct: (relevant hits/all hits). The manual annotation was far from straightforward.

TABLE 2. *Sense-Tagging Evaluation Results (Case Study 1).*

Part-of-Speech	*Occ.*	*Definitions & Extensions*	*All Material*
Adjectives (20),		*Precision*	*Precision*
Nouns (88),			
Verbs (44)	152	*37,5%*	*84,21%*
(32 sentences)			

As shown in table 2 the performance of the WSD with the use of only the definitions is, as one might have expected, very much lower than when considering the second case in which the definitions have been supplemented with a lot of other information. The qualitative improvement between the two cases is very high. If the definitions in GLDB had been richer as to information, the performance might have improved even more. Note, however, that Wilks & Stevenson (1998) observed that by using simulated annealing, the longer definitions in LDOCE tended to win over the shorter ones, since the length of the definitions varied considerably between different entries and thus influenced the software towards an erroneous solution. Accordingly, they elaborated a method to penalize the longer definitions, something that we did not consider in our study.

Table 3 shows the evaluation figures for a single entry, the verb *handla*, at the sense level, using example sentences randomly extracted from various sources.

TABLE 3. *Sense-Tagging Evaluation Results (Case Study 2).*

Case Study (2)	*Occ.*	*Definitions & Extensions*	*All Material*
		Precision	*Precision*
<handla>	60*	*41,37%*	*82,75%*

(*Since two of the occurrences were phrasal verbs not covered by GLDB, the evaluation was calculated on 58 examples.)

Large annotated data for single ambiguous words would probably reveal interesting groups of patterns and clear differences between such groups and entries, this is left for future research.

11. Conclusions

We have ported a large-scale sense-tagger, originally developed for the sense-tagging of English, to Swedish. We used one of the most comprehensive lexical resource available, the *Gothenburg Lexical Database* and tested the performance on open-source (newspaper) texts. The methodology behind the approach implemented in the sense-tagger follows the simulated annealing technique, used recently for the sense-tagging of English texts with the *Longman Dictionary of Contemporary English* (LDOCE). The combination of simulated annealing with a machine-readable dictionary has outperformed all other approaches based on dictionaries, achieving very high accuracy on the sense-tagging of *all* the content words in a text and not only a very small set, as is usually the case described in WSD-related literature.

Evaluation of WSD for Swedish gave good evidence that the task is feasible and the precision figures were very encouraging. The tagger will be used in the context of a larger architecture, for the acquisition of lexical semantic knowledge from open-source texts. The sense-tagger will contribute very important information for the precise assignment of lexical semantic knowledge to polysemous and homonymous content words. In this respect, we regard sense-tagging as a very important process and component when it is seen in the context of a wider and deeper text-processing architecture. Of equal importance is the way that the results returned by the sense-tagger can be used for the resolution of the preposition phrase attachment problem. GLDB contains information (typical prepositions) associated with the different senses of verbs, nouns and even a large number of adjectives. Once the right sense for a token is identified, the information found in the valency slot can more efficiently guide an algorithm to make the right decision as to whether a prepositional phrase functions as an argument or adjunct.

This issue will be investigated in the near future, as well as the evaluation of the sense-tagger on a larger scale.

Acknowledgements

This research was *partially* supported by the EU Language Engineering programme LE1-2238/AVENTINUS. We would like to thank Yorick Wilks and Mark Stevenson (University of Sheffield) for giving us the opportunity to test the optimized version of the simulated annealing software with Swedish resources. We are also grateful to Jerker Järborg for manually annotating the samples tested and for comments on a previous draft, as well as to Sven-Göran Malmgren for enabling the use of the GLDB.

References

NEO 1996 = *Nationalencyklopedins ordbok*, volumes 1–3, Språkdata & Bra Böcker AB.

Roget's Thesaurus = http://humanities.uchicago.edu/forms_unrest/ROGET.html

Allén, S. 1981. "The Lemma-Lexeme Model of the Swedish Lexical Database." In: Rieger, B. (ed.): *Empirical Semantics*. Bochum. Pp. 376–387. (Reprinted in: Allén, S. 1999. *Modersmålet i fäderneslandet. Ett urval uppsatser under fyrtio år av Sture Allén.* (*Meijerbergs Arkiv för svensk ordforskning*. 25.) Pp. 268–278.)

Boguraev, B. 1995. "Machine-Readable Dictionaries and Computational Linguistic Research." In: Walker D., Zampolli A. and Calzolari N. (eds.): *Automating the Lexicon. Research and Practice in a Multilingual Environment.* Oxford University Press. Pp. 301–336.

Brill, E. 1994. "Some Advances In Rule-Based Part of Speech Tagging." In: *Proceedings of the 12th AAAI '94*, Seattle Wa.

Brodda, B. 1979. "Något om de svenska ordens fonotax och morfotax: Iakttagelse med utgångspunkt från experiment med automatisk morfologisk analys." (PILUS 38). Institutionen för lingvistik, Stockholms universitet.

Brown, P.F., *et al.* 1991. "Word-Sense Disambiguation Using Statistical Methods." In: *Proceedings 29th ACL.* Berkeley, California. Pp. 264–270.

Chai, J.Y. and Bierman, A.W. 1997. "The Use of Lexical Semantics in Information Extraction." In: Vossen, P. et al. (eds.): *Proceedings of Automatic Information Extraction and Building of Lexical Semantic Resources Workshop.* Madrid, Spain

Cowie, J., Guthrie, J. and Guthrie, L. 1992. "Lexical Disambiguation Using Simulated Annealing." In: *Proceedings of 15th COLING*, Vol. 1. Nantes. Pp. 359–365.

Fr. Dolan, W.B. 1994: "Word Sense Ambiguity: Clustering Related Senses." In: *Proceedings of the 15th COLING*, Vol. II. Kyoto, Japan. Pp. 712–716.

Ejerhed, E. *et al.* 1992. "The Linguistic Annotation of the Stockholm-Umeå Corpus project." Technical Report No. 33, Univ. of Umeå.

Fellbaum, C. (ed.) 1998. *WordNet, an Electronic Lexical Database.* MIT Press.

Fujii, A. 1998. *Corpus-Based Word Sense Disambiguation,* PhD thesis, Tokyo Inst. of Computer Science, Japan.

Hutchins, W.J. and Somers, H.L. 1992. *Introduction to Machine Translation.* Academic Press.

Johansson-Kokkinakis, S. and Kokkinakis, D. 1996. "Rule-Based Tagging in Språkbanken." Research Reports from the Department of Swedish, Göteborg University, GU-ISS-96-5.

Kelly E., and Stone, P. 1975. *Computer Recognition of English Word Senses.* North-Holland Linguistic Series.

Kilgarriff, A. 1997a. "Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction." In: *Proceedings of the Lexicon Driven Information Extraction Workshop,* Frascati, Italy.

Kilgarriff, A. 1997b. "I Don't Believe in Word Senses." In· *Computers and the Humanities,* Vol. 31:2.

Klenk, U. and Langer, H. 1989. "Morphological Segmentation Without a Lexicon." In: *Literary and Linguistic Computing,* Vol. 4:4. Oxford University Press. Pp. 247–253.

Krovetz, R. 1997: "Homonymy and Polysemy in Information Retrieval." In: *Proceedings of the joined 35th ACL and 8th EACL.* Madrid, Spain. Pp. 72–79.

Lesk, M.E. 1986. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice-cream Cone." In: *Proceedings of the ACM SIGDOC Conference.* Toronto, Ca. Pp. 24–26.

Malmgren, S.G. 1992. "From Svensk ordbok ('A dictionary of Swedish') to Nationalencyklopediens ordbok ('The Dictionary of the National Encyclopedia')." In: Tommola H. *et al.* (eds.): *Proceedings of the EURALEX '92.* Vol. 2. Pp. 485–491.

Miller, G.A. (ed.) 1990. ”WordNet: An on-line Lexical Database.” In: *International Journal of Lexicography*, 3(4), Special Issue.

Miller, G.A., Leacock, C., Tengi, R. and Bunker, R.T. 1993. ”A Semantic Concordance.” In: *ARPA Human Language technology Workshop*. New Jersey, USA. Pp. 303–308.

Nag, H. T. 1997. ”Exemplar-Based Word Sense Disambiguation: Some recent Improvements.” In: Cardie, C. & Weischedel, R. (eds.): *Proceedings of the 2nd Conference on Empirical Methods in NLP*. Rhode Isl., USA. Pp. 208–213.

Peters, W., Peters, I. and Vossen, P. 1998. ”Automatic Sense Clustering in EuroWordNet.” In: *Proceedings of the LREC*, Vol. 1. Granada, Spain. Pp. 409–416.

Riggs, F. 1993. ”Social Science Terminology: Basic Problems and Proposed Solutions.” In: Sonneveld, H. and Loening, K. (eds.): *Terminology, Applications in Interdisciplinary Communication*. J. Benjamins Publ. Co. Pp. 195–222.

Sanfilippo, A. (chair) 1998. ”EAGLES: Preliminary Recommendations on Semantic Encoding.” Interim Report, The EAGLES Lexicon Interest Group. (http://www.ilc.pi.cnr.it/EAGLES96/, site visited 10 Feb. 1999)

Schütze, H. and Pedersen, J.O. 1995. ”Information Retrieval Based on Word Senses.” In: *Proceedings of the 4th Annual Symposium on document Analysis and Information Retrieval.* Las Vegas, NV. Pp. 161–175.

Wilks, Y. 1995. ”Texts and Senses.” Memoranda in Computer and Cognitive Science, CS-95-23. Univ. of Sheffield.

Wilks, Y. 1997. ”Senses and Texts.” In: *Computers and the Humanities*, Vol. 31:2.

Wilks, Y. and Stevenson, M. 1998. ”Word Sense Disambiguation Using Optimised Combinations of Knowledge Sources.” In: *Proceedings of the COLING/ACL ’98*. Montréal, Canada.

Wilson, A. and Thomas, J. 1997. ”Semantic Annotation.” In: Garside, R. et al. (eds.): *Corpus Annotation. Linguistic Annotation from Computer Text Corpora.*
Longman. Pp. 54–65.

Yarowsky, D. 1994. "A Comparison of Corpus-Based Techniques for Restoring Accents in Spanish and French Text." In: *Proceedings of the 2nd Workshop on Very Large Corpora,* Kyoto, Japan. Pp. 19–32.