A long-standing problem in Corpus-Based Lexicography and a proposal for a viable solution

Dimitrios Kokkinakis
Språkdata, Göteborg University
Box 200, SE-405 30,
Sweden
Dimitrios.Kokkinakis@svenska.gu.se

Abstract

This paper describes the application of a framework for text analysis to the problem of distinguishing *unusual* or *non-standard* usage of words in large corpora. The need to identify such novel uses, and augment machine-readable dictionaries is a constant battle for professional lexicographers that need to update their resources in order to keep up with the development of the dynamic and evolving aspects of human language. Of equal importance is the need to devise automatic means upon which we can evaluate to what extent a (defining) dictionary accounts for what we find in corpus data. A combination of both semi-, and automatic means have been explored, and it seems that Machine Learning might be a plausible solution towards the stated goals.

1. Introduction

Lexicography is concerned with the study of how words are used in a natural language by abstracting away surface differences between them. This paper deals with the question of creating a methodology that can aid professional lexicographers to distinguish *novel* or *non-standard* usage of (lexicalized) words. The process is automatized to a large extent, using state-of-the-art implemented Natural Language Processing (NLP) software for written Swedish, such as different kinds of annotators and analyses tools, coupled with large repositories of static and dynamic lexical knowledge. The NLP tools are geared towards the goal of distinguishing non-standard usage applied on very large bodies of texts. Approximated by these tools, surface differences between words are reduced while at the same time linguistic information is successively added in the data in order to create a suitable representation that can be matched and measured by a Machine Learning software. The obtained results, that do not conform to the *norm* according to the lexical resources and the method explored, need manual inspection in order to make a decision regarding the actual, or not, identification of a new or extended usage. However, the inspection process can be automatized, and accelerated, as soon as large portions of texts have been appropriately marked and analyzed, since Machine Learning techniques can be applied to calculate the difference or distance between previously analyzed and newly processed material. A distance to the nearest marked instance for each test item over a certain threshold, might indicate that an unusual or novel usage is identified, while under a threshold might indicate a prototypical usage.

The presentation that follows is not about word sense disambiguation (WSD), sense or semantic tagging *per se*, it is rather about how to use WSD, and other supporting NLP technologies, in a practical situation. Moreover, the discussion that follows is not a polemic point of view against the defining lexicon we chose, or any similar lexicons of that kind. We are fully aware of the limitations in terms of space, time, resources etc. that are prohibitive for producing better coverage and richer in content dictionaries. Dictionaries *are* and will *always be* incomplete and it is simply impractical to give descriptions of all possible sense nuances of a word, regardless of the theoretical model incorporated in the dictionary. Still, however, a typical scenario for many scholars, such as professional lexicographers, is the need to have sophisticated software in their every-day work in order to cope and organize the rapidly growing bulk of large corpora in electronic form, to automatically aid the evaluation of (defining) dictionaries against textual data, and to automate the process of discovering non-standard usage of words. The operational definition of non-standard usage we adopt is tightly connected to the use of an existing dictionary. Accordingly, corpus instances can either fit into the dictionary descriptions provided (standard usage) or not (non-standard usage).

This paper is organized as follows. We start by describing what we mean by “non-standard” use and give some background information regarding other efforts towards the same or similar goals, chapter (2); a brief presentation of various static and dynamic lexical resources as well as annotation and analyses tools for written Swedish will be presented, chapter (3); chapter (4) describes the method explored for the discovery of non-standard usage; while chapter (5) presents a worked example and discusses some preliminary results obtained; finally chapter (6) end the presentation by giving some general conclusions and directions for future research.

2. Background

2.1 What is non-standard use?

There is a justified need to distinguish new senses and novel usage of words, and accordingly augment defining dictionaries with "fresh" material. This is regarded as a constant battle for professional lexicographers that need to update their artifacts reliably and rapidly, in order to keep up with the development of the dynamic and evolving aspects of human language, and its constant and rapid change. In this paper we will use the terms: *new*, *novel* and *non-standard usage* or *sense* rather interchangeably referring to, roughly, the same thing. Moreove, when we talk about these terms we should always try to keep in mind that they are closely related to given *dictionary* senses. That is, "any numbered section of a defining dictionary entry which supports its own definition or requires separate treatment from the surrounding material"; (Atkins 1987).

Thus, the definition we adopt for the work presented in this paper, regarding what might constitute a non-standard use of a particular word, is based on the lexical resources in disposal, in our case the Gothenburg Lexical Database (GLDB). Consequently, if a word in a specific context does not match any readings of the descriptions, provided by the lexicographers, in this specific dictionary, then it will be worth examining it manually, in order to establish if a non-standard usage has been detected. Manual inspection is justified by the fact that deviating readings might simply be productive extensions[1], (easily) covered by applying some kind of mapping rule or process to the available readings, or, they might be errors produced by the complex interaction of the different tools applied on the test data.

For similar experiments, but with manually inspecting large text instances for the purpose of evaluating the Generative Lexicon see Kilgarriff (2000). Kilgarriff acknowledges that close reading of definitions from a published dictionary does not provide an ideal method for distinguishing standard from non-standard uses of words. However, the method has no fundamental flaws, he says, and there is no better method available. A disadvantage of the work presented by Kilgarriff is that he starts by first identifying a set of words to test whether the Generative Lexicon accounts for a novel word use. However, this cannot always be practical due to the fact that working with multi-million corpora and the totality of a natural language, at least as described by a dictionary, makes the suggested approach infeasible. Therefore, what we need is better, more general means that can be applied to more than a handfull of words. Nevertheless, I agree that manual judgement is still necessary regardless the qualitative or quantitative view of the approach adopted.

Comparisons between established machine-readable dictionaries (MRDs) and text collections have shown that there is a gap between the two in different dimensions. Since accurate, robust analysis tools for large corpora and machine-readable dictionaries were rare, until recently, even for over-explored languages, such as English, only a very coarse estimate of the coverage of the dictionaries has been studied. This is usually taking the form of identifying 'new' words, words not in a 'master wordlist' of 'existing' words, or counting how many of the words found in various text collections were in various machine-readable resources, such as the Longman Dictionary of Contemporary English (LDOCE) and the COLLINS dictionary; (Renouf (1993), Krovetz (1994)). Renouf (1993), for instance, discusses the use of several analytical tools or 'filters' to identify 'new' words in a flow of data on the basis that they are not in a list of 'existing words'. It is easy to confuse the finding of 'new' words and the identification of 'new' senses, since there are some similarities between the two. However, the identification of the later requires more sophisticated and precise software tools, and is a much more difficult task. Therefore, the lack of greater sophistication in the software tools that were available for working with large corpora has been prohibitive for researchers to go a step further and actually try not only to discover new words but new senses of existing, lexicalized words. Clear (1994) comments that

[1] For instance, the compound word *brandvägg* 'fire-wall' is not found as an entry in GLDB. The head *vägg*, however, is, with a sub-sense (1c) intended to cover metaphoric or extended usage for that specific word. Consequently, the (metaphoric) sub-sense for *brandvägg* namely, "a way to protect unauthorised access to a computer" is covered this way.

if one had software that could reliably categorise citations (i.e. concordance lines[2]) into semantic subsets, one could find answers quickly and easily to questions such as which citations out of the set do not appear to match any of a pre-defined set of word sense categories. Accordingly, potentially *new* uses of this word could be identified. After all, in the very relevant research area of WSD the interest has been persisently focused, until very recently, on quality WSD of a handful of target words, rather than quantity (Yarowsky (1995); Leacock *et al.* (1996)). Some large-scale WSD efforts on all content words are described in Kilgarriff & Palmer (2000) in the framework of the SENSe EVALuation (SENSEVAL) exercise.

2.2 Previous research

The idea of employing automatic techniques dealing with the topic of discovering new or novel usage is not new, on the contrary. In early work by Wilks (1980), within the developed "Preference Semantic" system, methods for dealing with extensions of word-senses are discussed. These are based on the incorporation of richer semantic structures, called *pseudo-texts*, and the observation of unexpected contexts. The shortcoming of the approach presented there, however, was the inadequacy to deal with many forms of lexical ambiguity. The elaborated mechanism of templates for all part-of-speech that was developed had to be both too general, for the creation of semantic representations, and too specific, to aid disambiguation. In his experiments, however, Wilks (1980) observes that extended or new usage is actually the *norm* in ordinary language use (at least for English).

Syntactic cues are used by Dorr & Jones (1996) for the derivation of semantic information and for augmenting on-line dictionaries for novel verbal senses. The syntactic cues are divided into distinct groupings that correlate with different word senses. For a very large number of verbs, the syntactic signatures, or syntactic patterns, are used. These are compared with Levin's, Levin (1993), verb classes and information from LDOCE. According to the algorithm presented by the two authors: if a verb is in Levin's lists it is classified accordingly; if not, WordNet synsets are used (lists of synonym semantic concepts, Miller *et al.* (1990)), and if the synonym is in Levin's classes they select the class that has the closest match with canonical LDOCE codes; if there are no synonyms, or LDOCE codes, a new verb class is created. Syntactic signatures are of the form: 'X broke the vase to pieces' which, according to Dorr & Jones, becomes '[np, v, np, pp(to)]'.

Similarly, Wilks *et al.* (1996) describe a method, referring to an unpublished manuscript by Jim Cowie at CRL-NMSU, on an effort to piggyback a dictionary from a corpus and a *seed* MRD. By applying the described method, all the occurrences of a word in a corpus are classified as belonging to sets of senses defined by a lexicographer who has examined a subset of the occurrences of the word using concordances. The authors argue that after a sufficient number of example senses have been marked it should be possible to classify the remaining instances of a word using different techniques. More interestingly, it may be possible to highlight unusual usages (or different *unclassified* senses) by identifying instances where the overlap occurrence is low, and subsequently it is necessary to examine these instances manually. Cases where the overlap is high may indicate archetypical example usages.

Hanks (1996) argues that the semantics of verbs are determined by their complementation patterns, discussing an empirical, semi-automatic approach, where it is necessary to identify typical subjects, objects and adverbials, and then group individual lexical items into sets. In creating the behavioural profiles of verb lemmas, such as 'urge' in large corpora, Hanks showed that 10% of the uses of 'urge' are metaphors and figuratives, while the most common patterns account for 61% of the occurrences "a person urging another person to do something". Moreover, Hanks proposed, that for unusual uses of words it is advisable to statistically sort their collocates into relevenat sets, to give a name and note possible correlations among different sets in particular roles (subject, object), and explain the relation by appealing to criteria of ellipsis, rhetoric, etc.

Finally, Tapanainen & Järvinen (1998) describe a tool that utilizes syntactic information and produces dependency syntax-based concordances between lemmatised words. The tool can be used for detecting relatively invariable phenomena, which, according to the authors, are collocations with fixed order and strict precedence. The tool can be used for studying, for instance, long distance dependencies by clustering sentences according to different syntactic structures. Although the authors do not directly claim that their tool can be used for discovering new usage of words, we think that it can certainly be an important step towards that direction as well.

[2] A concordance line is a formatted version or display of all the occurrences or tokens of a particular type in a corpus, the type is usually called the keyword or target or search item. Concordances form the main source of information in computer assisted lexicography.

3. Lexical and algorithmic resources

3.1 Lexical resources

The lexical resources that are used in this work consist of the GLDB, which is the largest, most comprehensive lexical resource for modern Swedish, upon which a number of defining dictionaries have been produced, Malmgren (1992), and the extended content of the Swedish SIMPLE semantic lexicon (Lenci *et al.* (1998)), over 25,000 entries, Kokkinakis *et al.* (2000). For the classification of proper names into semantic classes, a named-entity recognizer (NE) is used, Kokkinakis (1998). The semantic classes in the NE recognizer fall into the categories LOCATION, HUMAN, TIME and ORGANIZATION. Proper names are both frequent and have a serious impact for the disambiguation of the surrounding context. The NE module is also classifying personal pronouns referring to humans as well as appositive nouns (e.g. he, she, professor) to the class HUMAN.

3.2 Algorithmic resources & machine learning

The most important NLP tools that comprise the algorithmic resources are a rule-based part-of-speech tagger; a semantic tagger, Kokkinakis *et al.* (2000); a sense tagger (for content words), Kokkinakis & Johansson Kokkinakis (1999a); and a cascaded finite-state parser, Kokkinakis & Johansson Kokkinakis (1999b). Moreover, various finite-state based software that identify and mark idioms, multiword expressions, phrasal verbs, and perform heuristic compound segmentation and lemmatisation are used.

The idioms consist of approx. 4,500 different ones, according to the GLDB. Compound segmentation is based on the distributional properties of graphemes, trying to identify grapheme combinations that are non-allowable when considering non-compound forms in the Swedish language, and which carry information of potential token boundaries. The heuristic technique 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. After manual adjustments and iterative refinement a list of such graphemes has been produced and used for segmentation. Ambiguities are unavoidable, although the heuristic segmentation has been evaluated for high precision. Finally, lemmatisation is based on the output from the part-of-speech tagger and the rich feature representation that can be found in the part-of-speech tags. Examples of such grapheme sequences are ‘ngs|s’ and ‘iv|b’, e.g. forskningsskola ‘research school’, skrivbord ‘writing desk’; ‘|’ denotes where the segmentation will take place.

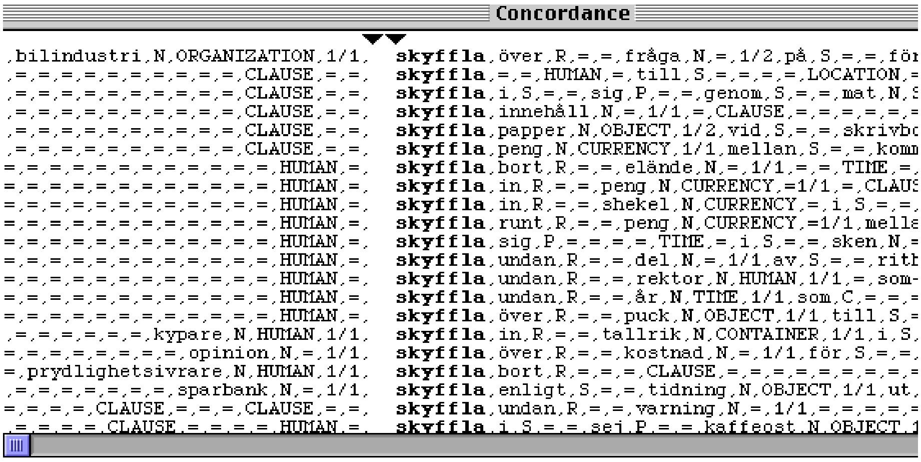
Machine learning techniques, Mitchell (1997) are used in order to automate the calculation of the overlap of the contexts between word that are candidates of defining a novel sense. More specifically, we adopt a supervised, inductive, classification-based variant of Machine Learning called Memory-Based Learning (MBL), and a specific implementation by Daelemans *et al.* (1999) called TiMBL. Using such techniques, the contexts, or instances or analyzed, modified concordance lines, can be *sorted* by calculating the distance of a new processed context with the distance to the nearest instance (or neighbour) of each test instance already processed. Here, the distance of two contexts is defined as the difference between the features within the instances.

Training and test instances consist of fixed-length vectors of symbolic *n* feature-value pairs (in the study presented in this paper n=37 for verbs and for nouns), and a field containing the classification of that particular feature-value vector. During classification an unseen example *X*, a test instance, is presented to the system and a distance metric *D* between the instances of the memory *Y* and *X* is calculated, *D*(X,Y). The algorithm tries to find the *nearest neighbour* and outputs its class, as prediction for the class of the test instance, as well as the distance from the nearest neighbour in the training instances.

4. Design and methodology

The methodology and design proposed in this report consists of an integrated approach to unify the results of the software outlined previously, that manipulate the content of lexical and textual resources, in order to aid the recognition of potential non-standard use of lexicalized words. Different tools are adding various types of annotations on the data, such as grammatical and semantic. Such annotations are means of giving to a text *added value*, in the sense that the added information can be used for a multitude of purposes; *cf.* Leech (1997). The model for the interpretation of non-standrd use that is presented is triggered by the use of a mixture of collocations and colligations, i.e. a collocation patterns based on syntactic information rather than individual words. This type of syntactic patterning demands extra information about the words in the corpus, information provided by the different software that will be described later.

The method is inspired by the previously description of the work by Jim Cowie discussed in Wilks *et al.* (1996:240) and the work by Kilgarriff (2000). In particular, I am interested in producing different modified views of typical concordance lines (see figure 1) and then use a combination of manual and automatic techniques for inspecting the obtained results. Moreover, when enough modified concordance lines have been produced, inspected and marked, given a particular sense from the available lexical resources, the overlap between old and new material can be measured, and aid a human towards the identification of novel senses, according to the lexical resources used, here the GLDB.



```
Concordance
,bilindustri,N,ORGANIZATION,1/1,   skyffla,över,R,=,=,fråga,N,=,1/2,på,S,=,=,för
,=,=,=,=,=,=,=,=,=,=,CLAUSE,=,=,   skyffla,=,=,HUMAN,=,till,S,=,=,=,=,LOCATION,=
,=,=,=,=,=,=,=,=,=,=,CLAUSE,=,=,   skyffla,i,S,=,=,sig,P,=,=,genom,S,=,=,mat,N,S
,=,=,=,=,=,=,=,=,=,=,CLAUSE,=,=,   skyffla,innehåll,N,=,1/1,=,CLAUSE,=,=,=,=,=,=
,=,=,=,=,=,=,=,=,=,=,CLAUSE,=,=,   skyffla,papper,N,OBJECT,1/2,vid,S,=,=,skrivbo
,=,=,=,=,=,=,=,=,=,=,CLAUSE,=,=,   skyffla,peng,N,CURRENCY,1/1,mellan,S,=,=,komm
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,bort,R,=,=,elände,N,=,1/1,=,=,TIME,=,
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,in,R,=,=,peng,N,CURRENCY,=1/1,=,CLAUS
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,in,R,=,=,shekel,N,CURRENCY,=,i,S,=,=,
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,runt,R,=,=,peng,N,CURRENCY,=1/1,mella
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,sig,P,=,=,=,=,TIME,=,i,S,=,=,sken,N,=
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,undan,R,=,=,del,N,=,1/1,av,S,=,=,ritb
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,undan,R,=,=,rektor,N,HUMAN,1/1,=,som-
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,undan,R,=,=,år,N,TIME,1/1,som,C,=,=,=
=,=,=,=,=,=,=,=,=,=,=,=,HUMAN,=,   skyffla,över,R,=,=,puck,N,OBJECT,1/1,till,S,=
,=,=,=,=,=,=,kypare,N,HUMAN,1/1,   skyffla,in,R,=,=,tallrik,N,CONTAINER,1/1,i,S,
=,=,=,=,=,=,=,=,opinion,N,=,1/1,   skyffla,över,R,=,=,kostnad,N,=,1/1,för,S,=,=,
=,prydlighetsivrare,N,HUMAN,1/1,   skyffla,bort,R,=,=,=,CLAUSE,=,=,=,=,=,=,=,=,=
,=,=,=,=,=,=,=,sparbank,N,=,1/1,   skyffla,enligt,S,=,=,tidning,N,OBJECT,1/1,ut,
=,=,=,=,CLAUSE,=,=,=,CLAUSE,=,=,   skyffla,undan,R,=,=,varning,N,=,1/1,=,=,=,=,=
,=,=,=,=,CLAUSE,=,=,=,=,HUMAN,=,   skyffla,i,S,=,=,sej,P,=,=,kaffeost,N,OBJECT,1
```

Figure 1. A more abstract representation of a concordance, using combination of lemmata, features and labels of various types

Essentially, the following processing steps are considered:

Gather a large number of sentences (or concordance lines or contexts) for every word to be examined from a (newspaper) corpus

Annotate these sentences with any possible type of information available (such as part-of-speech, sense and semantic information)

Parse with a syntactic analyzer

Normalize the information obtained by the different tools (lemmatisation, uppercase to lowercase conversion, keeping the head of long chunks recognized during parsing, etc.)

Decide the format for the MBL vectors

Create the fixed-format vectors, by gathering the information provided by the steps (2), (3), and (4) above, and use them as training data

Perform the same steps again, this time on sentences taken from another text genre and use the result as test data

Run MBL with the training and test data, and calculate the distance between the test and training instances

Inspect the results: zero distance? (identical instances), small distance? (possibly prototypical sense), large distance? (possibly, non-standard sense or processing errors)

By applying these steps, concordance lines can be transformed from raw text to a more abstract, annotated representation, upon which MBL can calculate the overlap between the material, see the worked example given in chapter (5).

4.1 Thresholds

Using MBL different threshold values are tested depending the provided algorithms. For instance, one is using the normalized Information Gain, (i.e. Gain ratio), which is measured by computing the difference in uncertainty (i.e. entropy) between the situations without and with knowledge of the value of a feature. Another threshold is used when the algorithm tested is the nearest neighbor search (*k-NN* or IB1), using weighted overlap and information gain weighting. Accordingly, the instances with the highest threshold values produced by the TiMBL software are manually examined, (identical instances, complete overlap, produce '0' zero distance). Of course, the returned values are dependent on the amount of training data, therefore, for each test performed the thresholds are adjusted in accordance with the results.

5. A case study and results

The purpose of the experiment I conducted and describe in this chapter is twofold. First, to investigate whether the proposed method is sufficient or not for the discovery of non-standard usage of words, and second, to devise automatic methods for evaluating to what extent a dictionary accounts for whatever we can find in corpora. The last point can be seen as a side-effect of the first, since both are closely interrelated. In order to test the applicability of the proposed architecture, and methodology, I chose to use as training material, data taken from the Swedish language bank (http://spraakbanken.gu.se/) which consists, to a large extent, of newspaper material. As testing material I chose to use data downloaded from the Internet, this time by searching on Swedish sites for sentences containing the words chosen for my experiments. The reason I chose the test material from the Internet is because the corpora that has been used for developing training data, and some of the tools that I will describe later, have also been used by the lexicographers for the production of the static resources, particularly the GLDB.

However, there is nothing that excludes that there might be cases in the training material alone where one can find non-standard sense or usage not used or ignored in the production and description of the material in the lexicon. Similarly, we can speculate, and certainly anticipate, that the test data may contain prototypical use of the words for the majority of the cases downloaded.

5.1 A worked example

In order to make the methodology described in chapter (4) clearer I will provide some examples showing how different tools process few sample instances. The key word under investigation in the small sample given, is the verb *skyffla*, which according to the GLDB has two senses, the first is similar to 'to shovel' while the second is most similar to 'to shove (away)'. All instances are taken from the Swedish language bank, while the annotations provided are, in some cases, simplified. Four such lines, with a very approximate interpretation, are:

(1) Socialdemokraterna försöker skyffla diskussionen om EMU under mattan.
'The social democrats are trying to shovel the discussion about EMU under the rug.'
(2) Han försökte också skyffla ansvaret för utvecklingen rörande Cypern på EU.
'He also tried to shove the responsibility regarding the development in Cyprus on EU.'
(3) Sakic skyfflade över pucken till Ozolinsh.
'Sakic shovelled the puck to Ozolinsh.'
(4) Morfar skyfflade kol i fabriken i femtio år.
'Grand father has shovelled coal in the factory for fifty years.'

5.1.1 Tokenization, part-of-speech annotation & lemmatisation

Tokenization not only identifies graphic words (tokens) but also recognizes idioms, phrasal verbs and multi-word expressions. The part-of-speech tags shown below are edited and simplified for the sake of simplicity and readability. The original tagset is using a slightly modified version of the Swedish PAROLE morphosyntactic description, (http://spraakdata.gu.se/parole/lexikon/swedish.parole.lexikon.html).

Lemmatisation is based on the output from the previous tool, and it is implemented as a finite state mechanism with three distinct states, one for verbs, one for nouns and one for adjectives. The suffix of the content words, along with their rich morphosyntactic features returned by the tagger are sufficient for establishing the base-form of each content word with very high accuracy.

(1a) Socialdemokraterna/NOUN försöker/AUX-VERB skyffla/VERB diskussionen/NOUN om/PREP EMU/ABBREVIATION under_mattan/IDIOM ./PUNC

Socialdemokraterna[socialdemokrat] försöker[försöka] skyffla [skyffla] diskussionen[diskussion] om EMU under_mattan .

(2a) Han/PRONOUN försökte/AUX-VERB också/ADVERB skyffla/VERB ansvaret/NOUN för/PREP utvecklingen/NOUN rörande/PREP Cypern/PROPER-NOUN på/PREP EU/ABBREVIATION ./PUNC

Han försökte[försöka] också skyffla[skyffla] ansvaret[ansvar] för utvecklingen[utveckling] rörande Cypern på EU.

(3a) Sakic/PROPER-NOUN skyfflade/VERB över/PARTICLE pucken/NOUN till/PREP Ozolinsh/PROPER-NOUN ./PUNC

Sakic skyfflade[skyffla] över pucken[puck] till Ozolinsh .

(4a) Morfar/NOUN skyfflade/VERB kol/NOUN i/PREP fabriken/NOUN i/PREP femtio/NUMERAL år/NOUN ./PUNC

Morfar[morfar] skyfflade[skyffla] kol[kol] i fabriken[fabrik] i femtio år[år] .

5.1.2 NE-recognition and semantic annotation

The named-entity labels are obtained by the NE-recognition, these are underlined in the sample below. The semantic annotation is following the SIMPLE model, colon ':' designates the hyper-hyponym relation in the semantic hierarchy, e.g. IDEO:HUMAN:CONCRETE means that the class IDEO "humans identified according to an ideological criterion" is hyponym of the semantic class HUMAN, which in turn is hyponym of the class CONCRETE. Note, that not all nouns get a semantic class annotation since the entries in the SIMPLE lexicon (at the time this work was completed) did not account for more than 25,000 noun senses.

(1b) Socialdemokraterna/IDEO:HUMAN:CONCRETE försöker skyffla diskussionen/ABSTRACT om EMU under_mattan .

(2b) Han/HUMAN försökte också skyffla ansvaret/COGNITIVE-FACT:ENTITY: ABSTRACT för utvecklingen rörande Cypern/LOCATION på EU/AGENCY: ENTITY:ABSTRACT .

(3b) Sakic/HUMAN skyfflade över pucken/ARTIFACT:OBJECT:NON-LIVING:CONCRETE till Ozolinsh/HUMAN .

(4b) Morfar/BIO:HUMAN:CONCRETE skyfflade kol/MATTER:NON-LIVING: CONCRETE I fabriken/FUNCTIONAL-SPACE:LOCATION:NON-LIVING:CONCRETE i femtio år/TIME .

5.1.3 Shallow parsing

Shallow parsing or chunking is based on the output from the part-of-speech tagging. During analysis, only the lemmatised head of each chunk (noun, adverbial, and adjective phrases and verbal groups) is preserved, as well as particles and the preposition heading a prepositional phrase. Here 'NP' is a noun phrase, 'VG' a verbal group, 'PP' a prepositional phrase 'RP' temporal adverbial phrase. Chunking is used for a couple of reasons; to obtain the head of the chunks, since phrases can be arbitrarily long and complex, and to give a shorthand name to constituents lying at a certain distance on the left and right of the word under investigation, e.g. different types of clauses.

(1c) NP[Socialdemokraterna/NOUN] VG[skyffla/VERB] NP[diskussionen/NOUN] PP[om/PREP NP[EMU/ABBREVIATION]] IDIOM[under_mattan/IDIOM] ./PUNC

(2c) NP[Han/PRONOUN] VG[skyffla/VERB] NP[ansvaret/NOUN] PP[för/PREP NP[utvecklingen/NOUN]] PP[rörande/PREP NP[Cypern/PROPER-NOUN]] PP[på/PREP NP[EU/ABBREVIATION]] ./PUNC

(3c) NP[Sakic/PROPER-NOUN] VG[skyfflade/VERB över/PARTICLE] NP[pucken /NOUN] PP[till/PREP NP[Ozolinsh/PROPER-NOUN]] ./PUNC

(4c) NP[Morfar/NOUN] VG[skyfflade/VERB] NP[kol/NOUN] PP[i/PREP NP[fabriken/NOUN]] RP[i/PREP NP[år/NOUN]] ./PUNC

5.1.4 Sense annotation

Sense annotation is given for content words, nouns, main verbs and adjectives. The notation provided, according to the GLDB, gives the lemma number followed by the lexeme or sense number. The underlying model adopted in GLDB is the so called *lemma-lexeme* model described in Allén (1981). The *lemmas* are grammatical paradigms, comprising formal data, e.g. technical stem, spelling variations, part of speech, inflection(s), pronunciation(s), stress, morpheme division, compound boundary, abbreviated form(s) and much more. The *lexemes*, are the numbered senses of a lemma, and are divided into two main categories, a compulsory kernel sense and a non-compulsory set of one or more sub-senses, called the *cycles*. A large number of metonymic uses of words are encoded under separate lexemes or cycles for a lemma.

(1d) Socialdemokraterna:1/1 försöker skyffla:1/2 diskussionen:1/1 om EMU under_mattan.

(2d) Han försökte också skyffla:1/2 ansvaret:1/1 för utvecklingen:1/1 rörande Cypern på EU.

(3d) Sakic skyfflade:1/2 över pucken:1/1 till Ozolinsh.
(4d) Morfar:1/1 skyfflade:1/1 kol:1/2 i fabriken:1/1 i femtio år:1/1.

5.1.5 Creating vectors

The vectors used by ML are of a fixed-format, and we have experimented with vectors of different size and content. We think that for optimal results and for different part-of-speech categories we need to model the content of the vectors differently. Therefore, the vectors for verbs consist of four 'contexts' to the left and four 'contexts' to the right of a verb under investigation (v1); while for nouns we model less context, namely two chunks to the left and two to the right (v2). If the noun under investigation is found in a prepositional phrase the preposition is also a part of the vector, finally the nearest to the noun modifier (if any) is also used in the vector.

(v1) Key-Word Part-of-Speech Byte-Offs {Left-Context} Key-Word {Right-Context} Class
(v2) Key-Word Part-of-Speech Byte-Offs {Left-Chunk} Prep Modifier Key-Word {Right-Chunk} Class

The left and right 'contexts' in (v1) and the chunks in (v2) are defined as clusters of four features of the form:

TOKEN:MORPHOSYNTAX:SEMANTIC-TAG-or-NE:SENSE-TAG

With 'MORPHOSYNTAX' we mean a part of speech (if the context concerns a single token, which is the head of a chunk, within the clause where the keyword appears) or a larger syntactic label (if the context concern a syntactic unit outside the keyword's own clause, e.g. 'CLAUSE'). With 'SEMANTIC-TAG-or-NE' is meant the result obtained by the semantic annotation and the NE-recognition, while with 'SENSE-TAG' is meant the result from the sense annotation (a GLDB lemma and lexeme number). Any of the features, or even all, can be absent, in this case a missing feature is marked with a 'dummy' character, an equal sign '='. The reason to it is that MBL requires that the vectors are of equal size, one of the few disadvantages of the method in general.

Given the sample of the worked examples, the results are gathered in the format below, and then converted to a fixed format of equal size for all vectors. Truncation is performed when a lot of context is available. 'BYTE-OFFS' is simply the position of the key word in the discourse, a mechanism that is inhereted by the tokenizer and helps linking the results with the original text from where it was taken.

(1e) SKYFFLA VERB BYTE-OFFS = = = socialdemokrat:NOUN:HUMAN:1/1 skyffla:VERB:=:1/2 diskussion:NOUN:ABSTRACT:1/1 om:PREP:=:= EMU:ABBR:=:= under_mattan:IDIOM:=:=

(2e) SKYFFLA VERB BYTE-OFFS = = = Han:PRONOUN:HUMAN:= skyffla:VERB:=:1/2 ansvar:NOUN: ABSTRACT:1/1 för:PREP:=:= utveckling:NOUN:=:1/1 rörande:PREP:=:= Cypern:PROPER-NOUN: LOCATION:= på:PREP:=:= EU:PROPER-NOUN:AGENCY:=

(3e) SKYFFLA VERB BYTE-OFFS = = = Sakic:PROPER-NOUN:HUMAN:= skyffla: VERB:=:1/2 över:PREP:=:= puck:NOUN:ARTIFACT:1/1 till:PREP:=:= Ozolinsh:PROPER-NOUN:HUMAN:=

(4e) SKYFFLA VERB BYTE-OFFS = = = morfar:NOUN:HUMAN:1/1 skyffla:VERB:=:1/1 kol:NOUN: MATTER:1/2 i:PREP:=:= fabrik:NOUN:FUNCTIONAL-SPACE:1/1 i:PREP:=:= år:NOUN:TIME:1/1

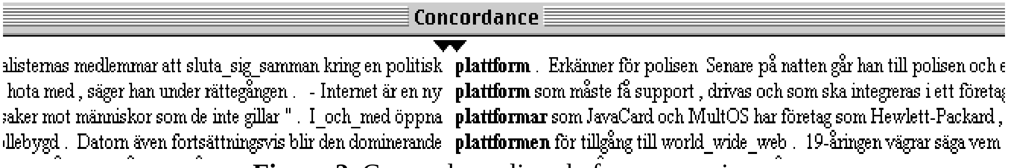
5.2 Small scale evaluation

We carried out a small evaluation of the presented approach by producing vectors such as the ones given in section 5.1.5, for a large sample of both training (newspaper articles) and testing material (gathered from the Internet) for some verbs and a number of common nouns. More specifically, some of the verbs we looked at were: *skyffla* 'to shovel, to shove away' (2 senses, 100 training contexts and 15 test contexts), *publicera* 'to publish' (1 sense, 300 training contexts and 20 test contexts) and the phrasal verb *hoppa in* 'to step in, to interfere' (2 senses, 200 training contexts and 10 test contexts). The results produced by MBL were sorted according the distance to the nearest instance in the training sample, a distance calculated on the metrics gain ratio and information gain, using the IB1 algorithm. The instances with the highest distance from the training material in every case were:

> (5) ColorFusion-kortet skyfflar hela 9 MB videodata per sekund.
> ‘The ColorFusion-card shovels 9 megabyte video-data per second.’
> (6) Framtidens taxibolag publicerar sina bilars positioner på Internet.
> ‘The future’s taxi companies publish their car’s positions on the Internet.’
> (7) Vid trafikavbrott kan den ena ringen hoppa in och ersätta den andra.
> ‘During interruption of traffic the one ring can interfere and replace the other.’

The characteristic for the test instances with the longest distance, in all three cases, has been the fact that a concrete object (e.g. ‘card’, ‘ring’) is initiating an action usually performed by a human or organization in the training material. While the longest distance for the instance of the verb ‘to publish’ has to do with publishing a ‘location’ or ‘position’ while the majority of the cases in the training sample has been to publish a concrete object (e.g. ‘article’, ‘report’). Over half of the test instances for the verb ‘to publish’ had short distance to the training material, explicitly refering to prototypical usage, while the opposite could be observed with the other two verbs.

Regarding the examined nouns, among others *plattform* ‘platform’, we provide here a more detailed picture of how the results look like. According to the GLDB, plattform has one main sense (lexem) and four sub-senses (cycles). Roughly, these sub-senses are: ‘a’ platform (concrete), ‘b’ tram, ‘c’ oil-rig and ‘d’ platform (abstract); ‘d’ was the commonest sense in the training material and used as default in all testing instances. Some of the results produced by the ML software, using 372 training and 18 testing examples, are given below (9-12). The character and number on the right designates the suggested sense for the key word and the distance of the example from the training instances. The underlied features simply helps to distinguish which features belong together (see 5.1.5). Small number (distance) designates that there are examples in the training material with many features in common with the test. Examples (9a-12a) show the nearest instance found in the training examples. Figure (2) shows the examples in concordance format.



alisternas medlemmar att sluta_sig_samman kring en politisk **plattform** . Erkänner för polisen Senare på natten går han till polisen och e
hota med , säger han under rättegången . - Internet är en ny **plattform** som måste få support , drivas och som ska integreras i ett företag
saker mot människor som de inte gillar " . I_och_med öppna **plattformar** som JavaCard och MultOS har företag som Hewlett-Packard ,
llebygd . Datorn även fortsättningsvis blir den dominerande **plattformen** för tillgång till world_wide_web . 19-åringen vägrar säga vem

Figure 2. Concordance lines before processing

(9) *plattform N 00* dator N APPARATUS 1 bli V = = = = = = dominerande A = 1 PLATTFORM för S = = tillgång N = 2 *till S* = = world_wide_web Y = = d **_d 2.306898_**

(9a) tåg N VEHICLE 1 bli V = = = = = = = = = = PLATTFORM för S = = samtal N ABSTRACT 1 om S = = utveckling N = =

(10) *plattform N 00* = = = = = = = = i_och_med S = = öppen A = = PLATTFORM som = = = JavaCard = = = ha V = = företag N AGENCY 1 d **_d 2.061991_**

(10a) = = = = = = = = = = = = = = = = PLATTFORM som = = = bära V = = = = = = system N = =

(11) *plattform N 00* medlem N SITU = sluta_sig_samman V = 1 kring S = = politisk A = 1 PLATTFORM = = = = = = = = = = = = = = = = d **_d 1.799718_**

(11a) folk N HUMAN 1 identifiera_sig V = 1 med S = = politisk A = 1 PLATTFORM = = = = = = = = = = = = = = = =

(12) *plattform N 00* Internet N = = vara V = = = = = = ny A = 2 PLATTFORM som = = = få V = = support N = = = = = = d **_d 1.074738_**

(12a) Internet N = = vara V = = = = = = ny A = 2 PLATTFORM = = = = = = = = = = = = = = = =

6. Conclusions and further research

This paper has outlined an approach to create a framework for aiding the identification of *novel* senses of words in large corpora. We think that important variation of word-usage is ‘hidden’ and hard to identify by merely looking at thousands of (non-processed) concordance lines. Therefore any means of organising the bulk of the data is absolutely necessary for future enhancements of the dictionary content with ‘new’ corpus-based material. Empirical preliminary results applied on a small sample of words have shown that although a lot of *noise* is produced by the different modules of the system in the form of errors in part-of-speech sense or semantic annotation sense differences between the annotated concordance instances can be observed under the threshold conditions briefly outlined. The noise caused by the reliability of some of the tools has been the main reason for the production of a number of false positive instances. However, while the error rates encountered are high, and the method immature for allowing fully-automatic integration of new data in the lexicon, our method can be used as a supporting tool that can aid the lexicographers identify new usage. Actually, we never intended to provide fully automatic means, since the manual refinement, inspection and judgement will always be the crucial factor left for the lexicographer; that is to make a final decision regarding what will be included and what will be left out from a dictionary.

The design presented in this paper is dictionary-dependent but the method is not specific to Swedish, as long as there are tools that can contribute with various types of morphological, syntactic, lexical semantic or other information to corpus data. Although our experiments are limited in magnitude, the results showed that the coverage of GLDB is pretty good. The future direction for the work presented will operate on a larger sample of the language and try to define more rigorous evaluation criteria.

References

Allén S. 1981 The Lemma-Lexeme Model of the Swedish Lexical Database. In Rieger B. (ed) *Empirical Semantics*. Bochum pp. 376–387

Atkins B.T. 1987 Semantic ID Tags: Corpus Evidence for Dictionary Senses. In *Proceedings of the 3rd OED*. Waterloo Canada

Clear J. 1994 I Can't See the Sense in a Large Corpus. In Kiefer F. Kiss G. and Pajzs J. (eds.). *Papers in Computational Lexicography COMPLEX '94*. Budapest pp. 33–45

Daelemans W. Zavrel J. van der Sloot K. 1999 TiMBL: Tilburg Memory Based Learner version 2. *ILK Technical Report 99-01*. Paper available from http://ilk.kub.nl/~ilk/papers/ilk9901.ps.gz

Dorr B. and Jones D. 1996 Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the 16th COLING*. Vol. 1. Copenhagen Denmark pp. 322-327

Hanks P. 1996 Contextual Dependency and Lexical Sets. *Journal of Corpus Linguistics*. Benjamins 1(1): 75–98

Kilgarrif A. 2000 Generative Lexicon Meets Corpus Data: the Case of Non-Standard Word Uses. In Bouillon P. Busa F. (eds) *Word Meaning and Creativity*. Cambridge UP

Kilgarriff A. and Palmer M. 2000 Introduction to the Special Issue on SENSEVAL. *International Journal of Computer and the Humanities. Special Issue on SENSEVAL*. 00:1-13 Kluwer Academic Publishers

Kokkinakis D. 1998 AVENTINUS GATE and Swedish Lingware. In *Proceedings of the 11th NODALIDA Conference (Nordisk Datalingvistik)*. Copenhagen Denmark pp. 22–33

Kokkinakis D. and Johansson-Kokkinakis S. 1999a Sense Tagging at the Cycle-Level Using GLDB. In *Proceedings of the NFL Symposium (Nordic Association of Lexicography)*. Gothenburg Sweden. Paper available from: http://svenska.gu.se/~svedk/publics/nfl.pdf

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

Kokkinakis D. Toporowska-Gronostaj M. and Warmenius K. 2000 Annotating Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon. In *Proceedings of the 2nd LREC.* Athens, Hellas (2000)

Krovetz R. 1994 Learning to Augment a Machine-Readable Dictionary. In *Proceedings of the EURALEX '94*. Amsterdam Holland pp. 107–116

Leacock C. Towell G. and Voorhees E.M. 1996 Towards Buidling Contextual Representations of Word Senses Using Statistical Models. Boguraev B. Pustejovsky J. (eds.): *Corpus Processing for Lexical Acquisition*. Bradford pp. 98–113

Leech G. (1997) Introducing Corpus Annotation In *Corpus Annotation. Linguistic Information from Computer Text Corpora* pp. 1-18 Longman

Lenci *et al.* 1998 *SIMPLE WP2 Linguistic Specifications* Deliverable 2.1 Pisa

Levin B. 1993 *English Verb Classes and Alternations: a Preliminary Investigation*. UCP

Malmgren S.G. 1992 From Svenska ordbok ('A dictionary of Swedish') to Nationalencyklopediensordbok ('The Dictionary of the National Encyclopedia'). In Tommola H. Varantola K. Salmi-Tolonen T. Schopp J. (eds). In *Proceedings of the EURALEX '92* Vol. 2. Tampere Finland pp. 485–491

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

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

Renouf A. 1993 A Word in Time: First Findings from the Investigation of Dynamic Text. Aarts J. de Haan P. Oostdijk N. (eds.) *English Language Corpora: Design Analysis and Exploitation*. Rodopi

Tapanainen P. and Järvinen T. (1998) Dependency Concordances. *Journal of Lexicography*. OUP 11(3): 187–203

Wilks Y. 1980 Frames Semantics and Novelty. In Metzing D. (ed) *Frame Conceptions and Text Understanding*. de Gruyter pp. 134–163

Wilks Y. Slator B. and Guthrie L. 1996 *Electric Words Dictionaries Computers and Meanings*. MIT

Yarowsky D. 1995 Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd ACL*. Cambridge MA pp. 189–196