Supervised PP-Attachment Disambiguation for Swedish; *(Combining Unsupervised & Supervised Training Data)*

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

Abstract

This paper is about the application of *Machine Learning* techniques to the prepositional-phrase attachment ambiguity problem. Since Machine Learning requires large amounts of training instances, the mixture of unsupervised and restricted supervised acquisition of such data will be also reported. Training was performed both on a subset of the content of the *Gothenburg Lexical Database* (GLDB), and a combination of instances from large corpora. Testing was performed using a range of different algorithms and metrics. The application language is written Swedish.

1. Introduction

Learning techniques for Natural Language (NL) ambiguity problems, based on statistics or Machine Learning, has been an active field of research in recent years, Charniak (1993), Yarowsky (1994), Brill (1995), Zavrel *et al.* (1997). This report deals with the application of such a learning technique to a particular instance of NL ambiguity, namely structural ambiguity. Structural ambiguity is a very serious type of global ambiguity in NL. One of the most typical manifestation of such ambiguity is the prepositional phrase attachment, henceforth PP-attachment, where syntactic information is insufficient to make the right assignment decision, and a prepositional phrase can attach to a constituent of almost any syntactic category. This report deals with the disambiguation of such prepositional phrases appearing in post-modifier positions. The solution follows an empirically-based Natural Language Processing (NLP) perspective, and it is applied on written Swedish corpora.

Prepositional phrase ambiguity resolution is necessary for the accurate acquisition of functional relations, semantic preferences and subcategorization information from natural language texts. Several methods have been discussed and tested by researchers and there is a general consensus, independent of the chosen approach, that simple means seem adequate for solving the problem, at least to a fairly good and useful for further processing level. The reported figures in the relevant literature vary between 75-95% correct disambiguation.

Section (2) illustrates the problem by providing some typical examples oftenly discussed in connection to the PP-attachment problem; section (3) gives a survey of the different directions applied for solving the problem; section (4) gives a brief overview of the Memory-Based Learning method that is used with the Swedish data; section (5) presents methodological issues concerning the acquisition of training and testing instances required by the above technique; section (6) discusses evaluation issues; finally, conclusions end the presentation of this work.

2. The Problem

The problem of PP-attachment disambiguation needs special attention, and careful treatment during automatic processing by the computer, since such prepositional phrases give rise to high degree of syntactic ambiguity. Hence, during automatic syntactic analysis, a parser must have a mechanism that will aid it to make a right decision among *at least* two equally grammatical parse-trees for the same sentence. Particularly for partial parsers, such mechanism is most probably applied separately, after parsing, *cf.* Abney (1990).

Consider a few of the classical examples oftenly quoted in the related to parsing and PP-attachment literature. These are typical, slightly simplified examples investigated by NLP researchers, and should be sufficient for giving a first flavour of what the problem that will be discussed in the subsequent sections is about.

(1) [[I][saw$_{V}$[the man$_{N1\,[}$with a telescope$_{N2]PP}$]]]
(1´) [[I][saw$_{V}$ the man$_{N1\,[}$with a telescope$_{N2]PP}$]]
(2) [Buy$_{V}$[books$_{N1\,[}$for children$_{N2]PP}$]]
(2´) [Buy$_{V}$ books$_{N1\,[}$for money$_{N2]PP}$]
(3) [[He][eats$_{V}$ pizza$_{N1\,[}$with a fork$_{N2]PP}$]]
(3´) [[He][eats$_{V}$[pizza$_{N1\,[}$with anchovies$_{N2]PP}$]]]

For the interpretation of the first example the PP ”with a telescope” attached to the object NP ”the man” would give a meaning to the sentence in the lines of ”the man who had a telescope”, while when the same PP gets a higher attachment, example (1´), that is to the verb ”saw”, the interpretation of the same sentence becomes rather ”by means of a telescope”. Similarly, in the second example the PP can be either attach to the object noun ”books” or to the verb ”buy”. It is fairly straightforward that without the help of contextual information, extra-linguistic knowledge or other means, no correct interpretation can be accomplished. However in similar examples to (2), such as (2´), the PP can almost certainly attach to the verb. Examples (3) and (3´), are slightly easier for a human to decide the correct attachment. In example (3) the interpretation should be in the lines of that ”one eats pizza which does not contain forks as an ingredient”, so the PP ”with a fork” should be attached to the verb ”eats”, while in example (3´) the interpretation will be ”one eats pizza in which one of the ingredients is anchovies”, and thus the PP ”with anchovies” should be attached to the N1, namely the word “pizza”.

For a human the resolution of the above examples, particularly (2) and (3), is based primarily on life time experience, people draw conclusions based on their knowledge of the world, and it is common knowledge that forks are not eatable while anchovies are, and what people know resolves the potential ambiguities so rapidly, that it is not even noticed. Moreover, it is often the extended context or certain key-words that can guide a computer process or human to which syntactic structure, and hence interpretation, is the correct one. The question that will be discussed in this report deals with what kind of information a computer system should have to its disposal in order to solve the above ambiguities; a complex discourse model or one that relies on superficial knowledge?

3. Background

3.1 Overview

The methodologies usually tried for giving an adequate solution to the previously described ambiguity are primarily based on the observation of the co-occurrences of verbs and nouns with prepositions in large bodies of text, observations which seem to be reliable indicators of lexical preference. It is also quite usual that these words, particularly nouns, are augmented with semantic class information of different fine- or coarse-grained type, such as ‘LOCATION’ or ‘ARTIFACT’.

Some of the different[1] methods and techniques that have been explored by researchers for the attachment disambiguation of the PP-phrases will be given in this section. These methods range from statistic, symbolic, lexicon and corpus-based. Furthermore, a method that has been gained a lot of attention during the last couple of years is borrowing ideas from the field of *Machine Learning* (ML); see Cardie & Mooney (1999) for an introduction to ML and Natural Language, and Mitchel (1997), particularly chapter (8) on ”Instance-Based Learning”. Using such approach, the attachment problem is considered as a *classification task*, and is thus suitable for solving it using ML algorithms. Disambiguation tasks are viewed as a classification of a test case, an ambiguous parse, in the class of correct and wrong parses, according to a set of suitable observations, the training examples, *cf.* Basili *et al.* (1997). Given an input sentence, a single property out of a set of potential properties is assigned to the input.

3.2 Approaches

Whittemore *et al*. (1990) were one of the first teams to prove, using large samples of texts, that superficial knowledge, namely *lexical preference*, was the key to resolving the attachment ambiguity in a cheap and clear way. Structured-based, attachment predictors originating from the field of psycholinguistics, such as *right association*, RA, the tendency for constituents to associate with adjacent items to their right, and *minimal attachment*, MA, the tendency to attach in a manner in which the least number of syntactic rules are employed, proved to be poor estimators for attachment disambiguation. Their study indicated that the method with the best results was based on the notion of lexical preference via verbs, nouns and prepositions. This means that there is a tendency for PPs to attach to *verbs* that have a preference for them, that there is a tendency for PPs to attach to *nouns* that have a preference for them, and finally prepositions themselves have a tendency to seek out certain kinds of constructions. The judgment of the attachment preference was made by hand.

In statistical methods, probabilities of words from large corpora are estimated and are usually of the form: `probability(V attach|V N1 P N2)`. Due to data sparseness the above 4-tuple (`V N1 P N2`) is sometimes reduced to `V P` and `N1 P`, and ignoring `N2`, a method investigated by Hindle & Rooth (1993), which produced an accuracy of 78-80% correct disambiguation. Their motivation to use statistics was generated by the fact that in the Whittemore *et al.*, study, it was rather unclear where the necessary information about LP was to be found. Instead, Hindle & Rooth estimated the strength of association of the preposition with verbal and nominal heads, from a parsed corpus, the 13 million AP corpus, was the basis for resolving the ambiguities. They also pointed out that the strategies discussed in literature until that time, such as RA and MA were proven to be inadequate in practice, as Whittemore *et al.* also discussed. At the same time, they overcome the problem faced by Whittemore *et al.* by automatically creating a list of lexical preferences. Superficial knowledge, such as lexical preference, seemed sufficient enough for the disambiguation of the vast majority of the PP-attachment cases.

[1] This report will deal with the disambiguation of the first PP in a `[V NP PP]` configuration. For the problem of attaching multiple PPs, namely in `[V NP PP1 .. PP`*n*`]` configurations, we refer to Merlo *et al.* (1997).

Ratnaparkhi *et al.* (1994) use both word and class information for the nouns N1 and N2, obtained by the use of *Mutual Information Clustering* of words and classes from two corpora, one with computer manuals and the Penn Treebank Wall Street Journal (WSJ) Marcus *et al.* (1993). The performance of their *maximum entropy* model scored 77,7% using only words, 79,1% using only classes and 81,6% using both, tested on the WSJ texts. Similarly, the testing on the computer manuals gave the scores 82,2%, 84,5% and 84,1% respectively. For comparison reasons, note that human experts on the WSJ performed 95,7% correct disambiguation.

Brill & Resnik (1994) presented a symbolic, rule-based approach to the disambiguation problem using transformation-based error-driven learning, in which unannotated text is passed through an initial-state annotator and then compared to the *truth*, indicated by a manually annotated corpus. Consequently, transformations are learned that can be applied to the output of the initial state annotator to make it better resemble the *truth*. The tuples were the form:

(4) `VERB HEAD-of-OBJECT-NP(N1)` *`SEM-CLASS`*
`PREPOSITION HEAD-of-PREP-GOVERNED-NP(N2)` *`SEM-CLASS`*

They were taken from the Penn Treebank, a number of 12,266 such tuples were extracted, and used as truth, while 500 examples were tested by their system, resulting 80,8% accuracy with a baseline of 64%, meaning that 64% of the cases in the truth (the prepositional phrases) were attached to the noun. With the addition of semantic class information, selectional restrictions taken fom the WordNet, Miller *et al.* (1990), such as ”year” and ”month” the performance of their method was raised to 81,8%. Note, that due to efficiency problems classes of `N1` and `N2` were not considered simultaneously.

Zavrel *et al.* (1997) used a ML approach as means for finding a solution to the problem. The different algorithms investigated were given a set of examples from annotated corpora, each consisting of an input vector of the context of the attachment ambiguity in term of features, and the possible attachment position representing the correct one for the input text. The common feature behind all of the algorithms tested is that they store some representation of the training set explicitly in memory. During testing, new cases were classified by extrapolation from the most similar stored examples. Using different similarity metrics and the way the instances were stored and searched in memory they achieved 84,1% correct attachment tested on the same sample as Ratnaparkhi *et al.* (1994). A brief description of their approach and algorithms, which are also explored in the present study, is given in section (4).

Stetina and Nagao (1997) used also, as the previously describe experiments, machine learning techniques for the induction of decision trees from a large set of training examples. These examples contained 4-tuples of the form `V N1 P N2` with disambiguated senses. The WordNet hierarchy was used, and the concepts in the hierarchy were used as attribute values. The decision trees were then used for classification of unseen examples. A 88.1%-90.8% correct disambiguation was repported.

Sopena *et al.* (1998) used a neural network architecture for the PP-attachment task. They scored a higher score than previous approaches using the Wall Street Journal corpus, namely 86,8%, and class information for not only the `N1` and `N2` but for verbs as well, taken from WordNet. They defend their very good results by pinpointing that the previous approaches did not use classes over `N1`, `N2` and `V`, and if they did, they did not consider them simultaneously.

Finally, de Lima (1997) discussed a practical application of the PP-attachment disambiguation task, a task implicit in the problem of subcategorization acquisition. In her work, and particularly for the disambiguation part of her study, she used the *Expectation Maximization* or EM algorithm, an iterative method to obtain maximum likelihood estimators in cases of sparse data problems.

4. Memory-Based Learning

In this report the Memory-Based Learning (MBL), a supervised, inductive, classification-based approach is adopted. MBL has several practical advantages that will only be briefly mentioned here.

- MBL has produced the best results so far in PP-attachment experiments with English data;
- the MBL method is not sensitive to sparse or low-frequency data, a serious problem encountered with some of the previously described approaches. Low-frequency cases are not discarded and are kept in memory, hence, useful information can also be extrapolated from them;
- Due to its *explanation capabilities*, using different types of verbosity mechanisms implemented in the software used (see later this section), the nearest neighbour(s) from which the decision was extrapolated can be studied;
- Fast learning and incremental learning, new instances can be added in the memory, improving the performance of the system.

For these reasons, a short introduction to learning using the Memory-Based approach will be given in this section. The software used for the experiments with the Swedish data has been developed at the University of Tilburg, by Daelemans *et al.* (1999). The software package, version 2.0, will be referred to as TiMBL, for short.

Learning approaches are usually categorized as statistical and symbolic. However, all learning methods are statistical in the sense that they attempt to make inductive generalizations from observed data and use it to make inferences with respect to previously unseen data. The difference may be that symbolic methods do not explicitly use probabilities in the hypothesis, Roth (1998). MBL are statistical methods originating from the field of ML. MBL is 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. The idea of storing all the training instances in memory results in the so-called instance base.

Training and test instances consist of fixed-length vectors of symbolic *n* feature-value pairs (in the study presented in this report n=13), 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 Δ between the instances of the memory *Y* and *X* is calculated, Δ (X,Y). The algorithm tries to find the *nearest neighbour* and outputs its class as prediction for the class of the test instance. The metrics used during classification can be one of the following *Overlap*, *Modified Value Difference, Gain Ratio* and *Information Gain*. These metrics are explored in three different algorithms by TiMBL, namely *Nearest Neighbour Search*, called *IB1* and *IB1-IG, IGTree* and a hybrid generalization of IGTree called *TRIBL*.

4.1 Metrics

The different metrics that can be tested and evaluated within TiMBL will be briefly described in the section, a more elaborated description can be found in a series of papers by Daelemans *et al.* (1996, 1999) and Zavrel *et al.* (1997).

Unweighted Overlap metric is the most basic one, described by the following two equations:

(5) $\Delta(X,Y) = \sum_{i=1}^{n} \delta(x_i,y_i)$ where:

(6) $\delta(x_i,y_i) = 1$ if $x_i \neq y_i$, 0 if $x_i = y_i$ or $x_i - y_i/\max_i - \min_I$

In these equations δ is the distance (or similarity) per feature and *n* is the number of features. The distance Δ between two patterns is the sum of the differences between the features. The Overlap Metric counts the number of (mis)matching feature values in both patterns *X* and *Y.*

Weighted Overlap is using *Information Gain (IG)* and *Gain Ratio*. Information gain of a classification task is defined in information theory as the average reduction in number of bits necessary to describe the correct classification or disambiguation. IG weighting looks at each feature in isolation, and measures how much information it contributes to our knowledge of a correct class label. Since IG tends to overestimate the relevance of features with large numbers of values it is used normalized, in this version the IG is divided by the entropy of the feature-values and called Gain Ratio.

Modified Value Difference Metric (MVDM) is a method to determine the similarity of the values of a feature by looking at co-occurrence of values with target classes. MVDM is fundamentally different than the previous, in the sense that the previous metrics are limited to exact match between feature-values. MVDM's problem is connected to sparse data, i.e. limited number of examples. In this case MVDM regards values in the same class as identical, and if the occur in different classes as completely different, the distance will be maximal.

4.2 Algorithms

The algorithm using the overlap metric is called *IB1*, the algorithm with IG metric is called IB1-IG. These algorithms are variants, or rather *naive* implementations, of the so called *nearest neighbour search* or *k-NN* classifier algorithm. IB1 and IB1-IG use a flat array of instances which is searched from the beginning to the end while computing the similarity of the test instances with each training instance.

IGTree is a structure which contains the same information as in the previous algorithm but restructured as a compressed decision tree structure. When the IG points to one feature (the most important), search can be restricted to matching a test instance at that feature. Instead of indexing all memory instances only once on this feature, the IGTree structure allows for the examination of the second most important feature, followed by the third most important feature, etc. IGTree is suitable for indexing and searching huge case bases.

When the differences in IG are very small IB1 and IB1-IG perform better than IGTree, for that reason Daelemans *et al.* (1999) designed *TRIBL* a hybrid generalization of IGTree. TRIBL exploits the trade-off between search speed and maximal generalization accuracy.

5. PP-Attachment and Swedish Data, Methodological Issues

Methodologically, the exploration of this study is centered around the great predictability of attachments and the strong preference between prepositions with nouns, verbs and adjectives. Furthermore, MBL will be the algorithmic approach behind this study. To a great extend the process is lexicon-driven, borrowing ideas from among others Whittemore *et al*. (1990), Hindle & Rooth (1993), and Jensen & Binot (1987), the latter for different reasons. Namely, that they acknowledge the usefulness of accesing the wealth in machine-readable dictionaries for solving complex ambiguity problems, such as the PP-attachment.

Although the work with disambiguating prepositional phrases in the literature is dealt with English, we can speculate that the results and ideas reported should be comparable to Swedish data as well, since both languages have structural and syntactic similarities.

Choosing the MBL approach leaves us with the problem of creating training data. Note, that a disadvantage with MBL and similar machine learning techniques is that the MBL's learning component requires a large number of instances, i.e. training data, for the good performance of the algorithms. For a language such as English, and for the particular task of disambiguating PP-attachment the problem of acquiring training data is eliminated by using existing parsed treebanks, such as the extensively used WSJ corpus. Using such treebanks the extraction of 4-tuples such as the ones required for this task, the heads from `VP NP` and `PP` constituents, namely `VERB NOUN1 PREPOSITION NOUN2` is a rather trivial extraction task from annotated corpora. Furthermore, heads of phrases are used since noun phrases can be arbitrarily complex.

In languages such as Swedish, in which the application of the MBL method will be tested, such parsed corpora do not exist. How can we then create the required instances, in the most inexpensive fashion? The problem is solved here using two different types of machine-readable material. The first is the content of the *Gothenburg Lexical Database* or GLDB for short, see Malmgren (1992), and partially parsed texts using a cascaded finite-state parser called Cass-SWE, Kokkinakis & Johansson-Kokkinakis (1999a).

5.1 GLDB & Corpora

GLDB is a rich lexical resource for modern Swedish, structured as a relational database. A number of printed Swedish monolingual, defining dictionaries have been generated from the GLDB, for instance the three-volume *Dictionary of the National Encyclopedia,* NEO (1996). GLDB was compiled on the basis of a large multi-genre corpora, and thus may be less subject to idiosyncracy. Using GLDB the acquisition of training data can be performed in an unsupervised manner. The second material is based on partially parsed texts. Using such parsed texts the training (and testing) data can be acquired in a restricted supervised way, using manual annotation of the test instances automatically extracted from texts. Since the material taken from GLDB contains sense information, a sense tagger for Swedish can be used for providing sense information associated with the words in the parsed texts as well, Kokkinakis &

Johansson-Kokkinakis (1999b). Furthermore, Named-Entity recognition can attach coarse-grained semantic information with the words in the training and testing material, *cf.* Kokkinakis (1998). The semantic information can be one of the following: *time sequence, location, person, organization, communication* and *transportation means, money expression* and *body-part.* The motivation for using the semantic information is based on the fact that a corpus might fail to provide a sufficiently extensive amount of word-word relationships, due to sparseness, and thus abstracting into such type of semantic information may improve the performance of the disambiguation task.

5.2 Required Format for Indata

Since different types of software are available for Swedish and can produce different types of feature-values, such as sense information, semantics labels, etc., see section (5.1), all this available information is taken under consideration in processing texts with the TiMBL software and used in the constructed vectors. Hence, the vectors used are of the following 13-tuple format, while the 14th element is the class assigned to the tuple, and can be either noun (N), verb (V) or adjective (A):

(7) `VERB` *`byte-offs`* `ADVERB SENSE`
NOUN1|ADJECTIVE *`byte-offs`* `SENSE SEMANTICS`
PREP NOUN2 *`byte-offs`* `SENSE SEMANTICS [CLASS {N,V,A}]`

The 13-tuple for training and test instances consists of a verb (`VERB`) its position in the discourse, using the byte-offsets, an adverb, (particle[1]), and sense number (`SENSE`) taken from the GLDB. The noun head of the object noun phrase (`NOUN1`) or the head of an adjective phrase (`ADJECTIVE`) with their byte-offsets, a sense label, as returned by the sense-tagger, and a semantic label, as it is returned by the named-entity recognition software. The preposition (`PREP`); and the head noun of noun phrase within the prepositional phrase (`NOUN1`) also with its byte-offsets, sense and semantic label.

Note that the question mark '?' can be used in the vectors. This should be interpreted as a particular value for a feature, which is either non-applicable or missing. Note, that ML algorithms require all instances to be of equal length, using '?' helps to fill such unspecified values. For instance, a representation for the minimal type of information that can be provided without the use of sense/semantic information might take the form of:

(8) ”`VERB ? ? ? NOUN1 ? ? ? PREP NOUN2 ? ? ?`”

In the TiMBL implementation of MBL, certain features in the instances can be skipped during processing, thus making the approach suitable for integrating it in visualization environments such as the General Architecture for Text Engineering, GATE, Cunningham *et al.* (1995). This is because in tools such as GATE, byte-offsets are used for component communication, components which produce information about texts

[1] In GLDB not all phrasal verbs are coded as separated entries. This can be explained by the fact that phrasal verbs are very productive and it is impractical for a dictionary of Swedish to contain information and definitions for all possible phrasal verbs. There is, unfortunately, a large number of verbs, that in their description include the information: '`(ofta med partikel)`' i.e. '(often with a particle)'; '`vanl. med partikel`' i.e. 'usually with particle' or '`ibl. med partikel`', i.e. 'sometimes with particle'. Furthermore the valency slots in the database may contain a mixture of particles and typical prepositions for the verbal entries, for instance: `blöta 1/1 (ned/upp) NP (i NP)`, i.e 'to make wet', here `ned` i.e. 'down' and `upp` i.e. 'up' are adverbs and `i` i.e. 'in' is a preposition. Moreover, `NP` stands for noun phrase, parenthetic information is optional.

which are stored separately with references back to the original text using the byte-offset information. Byte-offsets are also extracted automatically from the parsed texts for all the heads of the constituents but are skipped during processing by TiMBL, in (8) the fields 2, 6 and 11.

5.3 Unsupervised Extraction of Instances from GLDB

The unsupervised extraction of instances from GLDB can be divided into two main types. The first type is extracted from the 'valency' slot of the database and the second of the 'syntactic examples' slot of the lemma entries. GLDB gives information for over 61,000 lemmata, while 20,000 of those contain valency information.

The first type is of the form: `LEMMA SENSE VALENCY`, and the seond is of the form: `LEMMA SENSE SYNTACTIC-EXAMPLE`. The number of extracted instances of the first type were 4,500 for verbs, 7,000 for nouns and 1,500 for adjectives, while the instances acquired from the syntactic examples were approximately 1,000.

5.3.1 Valencies

The way the valencies are used as training instances in this study will be illustrated in this section, by the use of authentic examples. Consider for instance the valency slot for the verb `avstänga`, i.e. 'to shut off' which is given as:

(9) **avstänga** 1/3 NP **från** NP

this is automatically transformed to the 13-tuple format described previously which in this case is of the form:

(9´) VERB? ? SENSE? ? ? ? PREP ? ? ? ?
(9´´) **avstänga**? ? 1/3? ? ? ? **från** ? ? ? ?

In case the valency contains a typical adverb/particle for a particular verb, such as with the verb `haka`, i.e. 'to unhook', the information is encoded in the following way:

(10) **haka** 2/1 **av** NP **från** NP
(10´) VERB? ADVERBSENSE? ? ? ? PREP ? ? ? ?
(10´´) **haka**? **av**2/1? ? ? ? **från** ? ? ? ?

The noun entries, for instance for `förstöring` 'destruction', and the instance produced out of them take the following form:

(11) **förstöring** 1/1 **av** NP
(11´)? ? ? ? NOUN1? SENSE? PREP ? ? ? ?
(11´´)? ? ? ? **förstöring**? 1/1? **av** ? ? ? ?

Similarly, the adjectival entries, for instance for `arg` 'angry', and the instance produced out of them, gets the following format:

(12) **arg** 1/1 **på** NP
(12´)? ? ? ? ADJECTIVE? SENSE? PREP ? ? ? ?
(12´´)? ? ? ? **arg**? 1/1? **på** ? ? ? ?

5.3.2 Syntactic Examples

The second type of information from the GLDB is extracted from the syntactic examples associated in (almost) every lemma entry in the database. All the syntactic examples were parsed by Cass-SWE and then those that were parsed by the clause-patterns:

(13) <s>ADVERB* **NP** ADVERB* **VERBAL-GROUP** ADVERB* **(AP|NP)**
ADVERB* **PP** .*</s>
(14) <s>**VERBAL-GROUP** ADVERB* **(AP|NP)** ADVERB* **PP** .*</s>

were automatically extracted and used as training instances. Pattern (13) can be interpreted as any sentence (i.e. syntactic example) that contains a noun phrase before a verbal group, in order to secure straight word order, followed by the (infinite or finite) verbal group, followed by a noun phrase or an adjectival phrase and followed by a prepositional phrase. Pattern (14) is similar to (13) but with the requirement that the initial constituent is a verbal group, a phenomenon common in the database's syntactic examples. Bold face marks the obligatory constituents that will be extracted, while adverbials can freely intervene between these constituents and are ignored during the subsequent process. The annotation <s> and </s> mark the start and end of an input string. The notation '.*' means whatever follows until the end of the input string, this portion of the string is also ignored.

Moreover, it is worth mentioning that Cass-SWE operates on part-of-speech annotated texts. For this purpose we use Brill's rule-based tagger, Brill (1994), trained on Swedish material, Johansson-Kokkinakis & Kokkinakis (1996).

Since all rules (patterns) in the grammar are indexed with a unique identifier, it is fairly simple to extract the portion of the parsed examples that satisfied the above two rules. Note that one of the many output formats that can be produced using Cass-SWE is a string with the only annotation being the name of the rule matched. These strings, the syntactic examples, were then automatically transformed to the 13-tuple format required by TiMBL.

Consider the following two examples. The first under the lemma entry of the noun album, i.e. 'album' and the syntactic example associated to it, i.e. 'to put the post-cards in the album'. The second under the verb lemma entry sterilisera, 'to sterilize' and the syntactic example associated to it, i.e. 'to sterilize the instruments before the surgery'. The analysis in a slightly simplified manner for these examples is given below.

(15) **album** 1/1: sätta vykorten i **album**
(15´) **album**_1/1 $_{\text{/INDX}}$ sätta$_{\text{/INF-VERB}}$ vykorten$_{\text{/NOUN}}$ i$_{\text{/PREP}}$ **album**$_{\text{/NOUN}}$
(15´´) clause_00 -> $_{\text{INDX}}$[**album**_1/1] $_{\text{INF-VERBAL-GROUP}}$[sätta] $_{\text{NP}}$[vykort]
$_{\text{PP}}$[i $_{\text{NP}}$**[album]**]

(16)**sterilisera** 1/2: **sterilisera** instrumenten före operationen
(16´) **sterilisera**_1/2$_{\text{/INDX}}$ **sterilisera**$_{\text{/INF-VERB}}$ instrumenten$_{\text{/NOUN}}$
före$_{\text{/PREP}}$ operationen$_{\text{/NOUN}}$
(16´´) clause_00 -> $_{\text{INDX}}$[**sterilisera**_1/2] $_{\text{INF-VERBAL-GROUP}}$[**sterilisera**]
$_{\text{NP}}$[instrumenten x] $_{\text{PP}}$[före $_{\text{NP}}$[operationen]]

INDX is simply a dummy invented tag for the lemma information that appears in the beginning of every entry. Byte offsets are irrelevant, and not used within the training instances.

Consequently, the parsed format is automatically converted to a 13-tuple:

(17) VERB ? ? ? NOUN1 ? ? ? PREP NOUN2 ? SENSE ?
(17´) sätta ? ? ? vykort? ? ? i **album** ? 1/1 ?
(18) VERB ? ? SENSE NOUN1 ? ? ? PREP NOUN2 ? ? ?
(18´) **sterilisera** ? ? 1/2 instrument ? ? ? före operation ? ? ?

5.4 Unsupervised and Supervised Extraction of Instances from Corpus

Since the number of automatically extracted instances from GLDB, discussed previously, were not large enough, the training list was completed with a number of instances taken from a large corpus, parsed by Cass-SWE and similarly as before, strings that satisfied the patterns (13) and (14) were used as training material. The total number of instances extracted from the corpus is 3,000, making the total number of all available instances to a number of approximately 17,000. The training material that is used in this work might seem relatively low; however, it is comparable, and even have larger coverage, than the material used within the English experiments. For comparison reasons, Brill & Resnik’s (1994) training data consisted of 20,810 non-lemmatized instances, (see http://www.cs.jhu.edu/~brill/home.html, particularly document: pp-attach-english-train), out of these, the number of unique verbs is 3,347 verbs (e.g. 800:is, 395:was, 266:be); 4,405 NOUN1 (e.g. 800:’%’, 423:million, 183:it), and 5,695 NOUN2 (541:million, 239:’%’, 189:billion), while there is a large number of duplicate instances, for example “rose % to million” occurs 31 times, and “fell % to million”, also 31 times.

5.4.1 Unsupervised Instances from Corpus

Out of the 3,000 instances from corpora, a thousand of these were extracted and classified automatically using a 100% unambiguous heuristic, namely that *a preposition is attached to the verb if the noun phrase head is a (personal) pronoun*; Hindle & Rooth (1993) used also this heuristic, as well as few other similar cases in which “sure” verb attachment could be estimated from texts. Conisder the examples (19) and (20):

(19) … ta henne/PRONOUN på en promenad …
‘take her for a walk’
(20) … skilde honom/PRONOUN från de andra …
‘separated him from the rest’

5.4.2 Supervised Extraction of Instances from Corpus

In order to use more information from corpus, the texts were previously automatically pre-processed, as in the previously discussed cases. They were first annotated with part-of-speech, some with sense information, as well as semantic labels, if applicable, (21´, 22´ and 23´), and then partially parsed (21´´, 22´´ and 23´´). After the automatic creation of the 13-tuple format (21´´´, 22´´´ and 23´´´) all the instances were manually classified according to whether the prepositional phrase would be attached to the main verb (CLASS=V) of a verbal group, the head noun of the object noun phrase (CLASS=N), or the head adjective of an adjectival group (CLASS=A), examples (21´´´´, 22´´´´ and 23´´´´).

The following three simplified examples illustrate this approach:

(21) Den gripne lockade sällskapet med narkotika.
‘The arrested tempted the crowd with drugs.’
(21´) Den/DETERMINER gripne/PARTICIP/PERSON lockade/VERB/1/2
sällskapet/NOUN/PERSON/1/2 med/PREP narkotika/NOUN/1/1 ./F
(21´´) clause_01 -> [<IGNORED> $_{\text{FIN-VERBAL-GROUP}}$[lockade] $_{\text{NP}}$[sällskapet]

$_{\text{PP}}$[med $_{\text{NP}}$[narkotika]] <IGNORED>]

(21´´´) Extracted string: `locka sällskap med narkotika`

(21´´´´) Instance: `locka 11-17 ? 1/2 sällskap 21-28 1/2 PERSON med narkotika 34-42 1/1 ?` **CLASS=V**

(22) `Se enskild rapport om kollegornas förfarande med Alexander Lukas.`
‘See separate report about the colleagues conduct with Alexander Lukas.’

(22´) `Se/VERB/1/2 enskild/ADJ rapport/NOUN/1/1 om/PREP kollegornas/NOUN/PERSON förfarande/NOUN/1/1 med/PREP Alexander/PROP-NOUN/PERSON Lukas/PROP-NOUN/PERSON ./F`

(22´´) clause_01 -> [$_{\text{INF-VERBAL-GROUP}}$[Se] $_{\text{NP}}$[enskild rapport] $_{\text{PP}}$[om $_{\text{NP}}$[kollegornas förfarande]] <IGNORED>]

(22´´´) Extracted string: `se rapport om förfarande`

(22´´´´) Instance: `se 0-1 ? 1/2 rapport 11-17 1/1 ? om` förfarande 34-43 1/1 ? **CLASS=N**

(23) `Bert är den bästa i branschen.`
‘Bert is the best in the business’

(23´) `Bert/PROP-NOUN/PERSON är/VERB den/DETERMINER bästa/ADJECTIVE i/PREP branschen/NOUN ./F`

(23´´) clause_01 -> [<IGNORED> $_{\text{FIN-VERBAL-GROUP}}$[är] $_{\text{AP}}$[bästa] $_{\text{PP}}$[i $_{\text{NP}}$[branschen]] <IGNORED>]

(23´´´) Extracted string: `vara bra i bransch`

(23´´´´) Instance: `vara 5-6 ? 1/2 bra 12-16 1/1 ? i bransch 22-28 1/1 ?` **CLASS=A**

6. Evaluation

6.1 Testing Data

Testing was performed on a subset of the *press97*, using the methodology described in section (5.4) a sample of 250 instances was randomly extracted from that corpus. The manual classification was made by three human annotators, native speakers of Swedish. After a thorough examination of the manually classified instances, three were discarded due to parsing errors, and two were discarded due to unresolved ambiguity, ambiguity that could only be resolved if extended context was available; furthermore, 38 instances were not used for testing, since the results provided by the annotators were different on these cases. The results proved that the task was far from trivial for the human annotators. The remaining “unambiguous” 207 instances were tested in TiMBL, using many different combinations of the provided metrics and algorithms. The manually classified and lemmatized sample of the 207 instances used for the testing is available from (~~, and in the appendix. We consider as baseline the 52,65%, which is the most frequent attachment observed in the 207 test examples, in this sample it was the 109 occurrences of the noun attachments that were most frequent.

6.2 Results

The testing was performed using two sets of training data. The first set consisted only of the valency information for nouns, adjectives and verbs, extracted from GLDB, 13,000 instances, as described in section (5.3.1), see table (1). The second set consisted of all the available training material, that is valencies, syntactic examples and instances from corpora, see table (2). All three available algorithms in TiMBL were tested given all the different metrics and weighting combinations, furthermore all the tests were conducted using the first best and the three best nearest neighbours.

The idea of testing the TiMBL on these two different training sets was generated by the need to investigate the coverage of the valency information in GLDB. The results given in the first table show what results one should expect using a hypothetical parser that only uses information from the GLDB, during the PP-attachment disambiguation. Only the eight highest scores produced are given in table (1).

Base-line	-	-	-	52,65 %
IB1	MVD	NW	k3	**70.4%**
TRIBL	MVD	NW	k3	**70,4%**
TRIBL	MVD	NW	k1	69,8%
IB	WO	NW	k3	68,1%
IB	MVD	NW	k1	68,1%
TRIBL	WO	GR	k1	67,6%
TRIBL	WO	NW	k1	67,1%
IB	WO	GR	k1	67,1%

Table 1. The score for the PP-attachment test sample using the various metrics and algorithms based on training of the GLDB valencies, (top 8 classifiers).

The highest score, (70,4%), was produced by both the IB1 algorithm, using modified value difference, no weighting and the three best neighbours (k=3), as well as by the TRIBL algorithm with the same parameters.

Using all available material, the highest score, (86,47%), was produced by the TRIBL algorithm, using weighted overlap, information gain and the first best neighbour (k=1).

Base-line	-	-	-	52,65 %
TRIBL	WO	IG	k1	**86,47%**
TRIBL	WO	NW	k1	85,99%
TRIBL	WO	GR	k1	85,02%
TRIBL	MVD	IG	k1	79,22%
TRIBL	MVD	GR	k1	78,26%

Table 2. The score for the PP-attachment test sample using the various metrics and algorithms based on all the available training material, (top 5 classifiers).

In tables (1) and (2) the abbreviations stand for:
WO: Weighted Overlap, MVD: Modified Value Difference, GR: Gain Ratio, NW: No Weighting, k1: best neighbour, k3: three best neighbours

7. Conclusions and Further Work

Obviously, it is fairly inappropriate to make comparisons with the English experiments. Nevertheless, as one might have speculated, the obtained results of the application of MBL techniques to English data are comparable, and even slightly better in some cases, with the Swedish data, based on the all the available training material. This can partly depend on the fact that the attachment decision is calculated on multiple sources of information, and lemmatized data. Some might be coarse-grained, such as the semantic information, but they seem to work pretty well in practice. It is of equal importance to stress that the training material has been to a great extend acquired from a machine readable dictionary. By prefering this methodology the data-acquisition-phase bottleneck was considerably eliminated, a serious drawback for the MBL approach which requires large samples of training material.

As an equally important side-effect of the presented work is the investigation of the adequacy of the valency content of the lexical database as a valuable source for practical NLP experiments. Using the valencies alone, the results in table (1), show that (70,4%) disambiguation accuracy could be obtained by using them, an 18 point increase from the baseline. While the combination of the content of the GLDB and instances from corpora gave an improvement of almost 34 points from the baseline.

The obtained results will be used in the context of producing better automatic, syntactic analysis of Swedish texts, the PP-attachment disambiguation will be combined with the output produced by a large-coverage, partial parser for Swedish, already developed, thus enhancing it qualitatively. The results will be also used for the acquisition of subcategorization information for verbs and nouns.

Acknowledgements

We are grateful to the three annotators for manually annotating the samples tested; as well as to Sven-Göran Malmgren for enabling the use of the GLDB in this work. This report was presented at the seminar series: *Empirisk Språkteknologi*, (Empirical Language Technology), Department of Swedish language, Språkdata, 18th november, 1999; and at the ”24e Sammankomst för Svenskans Beskrivning”, conference, Linköping, 22th october, 1999, with the Swedish title “*Bestämning till Verb eller Nomen? Disambiguering av svenska prepositionsfraser m.h.a. automatisk betydelse-identifierng och innehållet i GLDB*”.

References

Abney S. (1990), *Rapid Incremental Parsing with Repair*, In Proceedings of the 6th New OEDC, pp. 1-9, Waterloo, Ont.

Basili R., Candito M.H., Pazienza M.T. and Velardi P. (1997), *Evaluating the Information Gain of Probability-Based PP-Disambiguation Methods*, In New Methods in Language Processing, Jones D. and Somers H. (eds.), pp. 241-255, UCL Press

Brill E. and Resnik P. (1994), *A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation*, In Proceedings of the COLING ’94, pp. XX-XX, Paper available from: http://www.cs.jhu.edu/~brill/acadpubs.html

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

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

Charniak E. (1993), Statistical Language Learning, MIT Press

Cunningham H., Gaizauskas R. and Wilks Y. (1995), *A General Architecture for Text Engineering (GATE) – A New Approach to Language Engineering R&D*, Technical report CS - 95 - 21, University of Sheffield, Department of Computer Science, http://www.dcs.shef.ac.uk/research/groups/nlp/gate/. Site visited 06/11/97

Daelemans W., Zavrel J., Berck P. and Gillis S. (1996), *MBT: A Memory-Based Part of Speech Tagger-Generator*, In Proceedings of the 4th Workshop on Very Large Corpora, pp. 14-27, Ejerhed E. & Dagan I. (eds), Copenhagen, Denmark

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

Hindle & Rooth (1993), *Structural Ambiguity and Lexical Relations*, In Journal of Computational

Linguistics, Vol.19:1, pp. 103-120

Jensen K. and Binot J-L. (1987), *Disambiguating Prepositional Phrase Attachment by Using On-Line Dictionary Definitions*, In Journal of Computational Linguistics, Vol. 13:3,4, pp. 251-260

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

Kokkinakis D. (1998), *AVENTINUS, GATE and Swedish Lingware*, In Proceedings of the 11th NODALIDA Conference, Nordiska Datalingvistikdagarna, pp. 22-33, Copenhagen, Denmark, Paper available from: http://www.nodali.sics.se/bibliotek/nodalida/1998_kph/, Site visited 06/11/1998

Kokkinakis D. and Johansson-Kokkinakis S. (1999a), *A Cascaded Finite-State Parser for Syntactic Analysis of Swedish*, In Proceedings of the 9th EACL, pp. 245-248, Bergen, Norway

Kokkinakis D. and Johansson-Kokkinakis S. (1999b), *Sense-Tagging at the Cycle-Level Using GLDB*, In Proceedings of the 'Nordisk Förening i Lexikografi' NFL Symposium, Göteborg, Sweden

de Lima E.F. (1997), *Acquiring German Prepositional Subcategorization Frames from Corpora*, In Proceedings of the 5th Workshop on Very Large Corpora, pp. 153-167, Zhou J. and Church K. (eds), China & Hong Kong

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

Marcus M., Santorini B. and Marcinkiewicz M. (1993), *Building a Large Annotated Corpus of English: the Penn Treebank*, In Journal of Computational Linguistics, 19(2)

Merlo P., Crocker K. and Berthouzoz, (1997), *Attaching Multiple Prepositional Phrases: Generalized Backed-off Estimation*, In Proceedings of the 2nd Conference on EMNLP, Cardie C. and Weischedel R. (eds), pp. 149-155, Rhode Isl., USA

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

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

NEO, (1996), *Natonalencyklopedinsordbok*, Volumes 1-3, Språkdata & Bra Böcker AB

Ratnaparkhi A., Reynar J. and Roukos S. (1994), *A Maximun Entropy Model for Prepositional Phrase Attachment*, In Proceedings of the ARPA Human Language Technology Workshop, pp. 250-255, Paper available from: http://www.cis.upenn.edu/~adwait/statnlp.html

Roth D. (1998), Learning to Resolve Language Ambiguities: A Unified Approach, In Proceedings of the AAAI-98

Stetina J. and Nagao M. (1997), *Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary*, In Proceedings of the 5th Workshop on Very Large Corpora, pp. 66-80, Zhou J. and Church K. (eds), China & Hong Kong

Sopena J.M., Lloberas A. and Moliner J.L. (1998), *A Connectionist Approach to Prepositional Phrase Attachment for Real World Texts*, In 17th COLING-36th ACL, Vol. 2, pp. 1233-1237, Montreal, Canada

Whittemore G., Ferrara K. and Brunner H. (1990), *Empirical Study of Predicative Powers of Simple Attachment Schemes for Post-Modifier Prepositional Phrases*, In Proceedings of the 28th ACL, pp. 23-30, Pittsburgh, Pen., USA

Zavrel J., Daelemans W. and Veenstra J. (1997), *Resolving PP attachment Ambiguities with Memory-Based Learning*, In Proceedings of the Computational Natural Language Learning Conference, Elison M. (ed.), pp. 136-144, Madrid

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* (pp. 19-32), Kyoto, Japan.

Appendix

The 207 instances used in the presented experiments.

```
avslöja   brist   i budgeteringsarbete   N
berätta   detta   med ömsinthet   V
besöka   knarkare   i Hagsätra   N
blanda   soja   med majsenavatten   V
blanda   ned mjöl   i taget   V
bli   Benelux-derby   i VM-slutspel   V
bli   bank   efter Bank_of_Tokyo-Mitsubishi   V
bli   hjärtbytespatient   i Nord   V
bli   hyllning   till gud   N
bli   nation   i VM-omgång   V
bli   rubrik   i kvällstidning   N
bli   underhållning   till Nobelbankett   N
bli   valuta   utanför euroblock   N
blåsa   liv   i figur   V
bygga   tunnel   genom Hallandsåsen   N
bygga_upp   anseende   på nytt   V
bära   ulster   av slag   N
börja   liv   med pojkvän   N
existera   gemenskap   utanför plan   V
falla   dom   i Senna-rättegång   N
finnas   argument   mot tanke   N
finnas   fördom   om Persson   N
finnas   hjälp   för dataamatör   N
finnas   klåfingrighet   från sida   N
finnas   missbrukare   i Storstockholm   V
finnas   retrospektion   inom nyromantik   V
finnas   risk   för hjärtverksamhet   N
finnas   socialvård   i Ryssland   V
finnas   volym   i ämnen   V
fira   guldbröllop   i Huskvarna   V
fråga   Aschberg   i inledning   V
fråga   Julia   från Verona   N
få   brev   av honom   V
få   fängelse   för underrättelseverksamhet   V
få   försoningsstund   under permission   V
få   hedersutmärkelse   av stadsdelsnämnd   V
få   intryck   av klubb   N
få   krona   av honom   V
få   krona   för dom   V
få   stipendium   om krona   N
förbli   sanning   i   V
föreslå   ring   hos riksgälden   N
förlora   tro   på politiker   N
försätta   Pettersson   i vakuum   V
förutse   parti   när_det_gäller avgiftsväxling   V
ge   kontur   åt huvudroll   V
ge   stöd   åt handel   V
gilla   arbete   med barn   N
godta   al   för konserveringsarbete   V
gälla   Mexiko   i väster   N
gälla   öl   av styrka   N
gå   bra   på biograf   V
gå  emot beslut   om Stadsgårdskajen   N
gå  emot uttalande   från kommun   N
göda   förnöjsamhet   istället_för omprövning   V
göra   tidning   till magasin   V
göra   återkomst   på skandalkarta   V
ha   alkohol   i kroppen   V
ha   bankman   i land   V
ha   budplikt   i Finland   V
ha   dalbanehumör   med stubin   N
ha   dem   på recept   V
ha   familj   på plats   V
ha   förtroende   för henne   N
ha   handlag   i situation   V
ha   huvud   i eld   V
ha   katalysator   på hjälpmotor   V
ha   melodi   på platta   V
ha   målvakt   av klass   N
ha   namn   i tidning   V
ha   problem   med anorexi   N
ha   roll   i Cityakuten   N
ha   seger   i lägen   V
ha   sinne   för tradition   N
ha   son   i närhet   V
ha   säljkurva   framför sig   V
ha   tid   framför sig   V
ha   vara   i ficka   V
hota   vakt   till livet   V
hålla   föredrag   om IT-samhälle   N
hålla   valupptakt   i Uppsala   V
höra   ord   om honom   N
inleda   byteshandel   med verklighet   N
inleda   störningsaktion   mot gatulangning   N
innebära   intrång   i integritet   N
innebära   steg   mot valfrihet   N
klara   allt   utom ätstörning   N
klicka   smörklick   på kyckling   V
komma   han   från familj   V
konstatera   Narkotikakommission   i rapport   V
landa   handske   på sandsäck   V
ligga   laddning   i luften   V
locka   många   av unga   N
lägga   miljard   på Ferrari   V
lägga   räka   i äggvita   V
lägga_ut   krona   på honom   V
lämpa   hus   med avgångsvederlag   V
möta   Tyskland   i Super_Cup   V
notera   säljökning   på procent   N
nå   Söderhamn   efter lunch   V
nå   kultstatus   genom uppläsning   V
plocka_fram   namnlista   på narkoman   N
prata   tag   om beroendet   V
presentera   bild   av lighet   N
publicera   rön   i tidskrift   V
reda_ut   problem   före konsert   V
representera   Sverige   i Europafinalen   V
reta  upp medlem   i fackförbund   N
råda   kostnadsjakt   i kommun   V
se   behov   av konferens   N
se   ljuspunk   i statistik   V
se   orsak   till problem   N
se_ut   dyster   för järnvägsföretag   V
sjunka   något   vid öppning   V
skaffa  sig profil   å_la socialdemokrati   N
skildra   katastrof   med värme   V
skriva   Cicerobiografi   i form   N
skära   den   i skiva   V
skära   fisk   i bit   V
skära   kött   i strimla   V
släppa_in   främling   i hem   V
spela   musik   i_stället_för ishockey   V
sprida   information   om forskning   N
sträcka_ut   hand   efter kartong   V
stuva   dem   i rum   V
ställa   fyllning   åt sida   V
ställa   skyldiga   till svars   V
säga   inget   om oskuld   V
sätta   hopp   till Gore   V
söka   Engqvist   för kommentar   V
ta   avstånd   från terrordåd   V
ta   fixen   hos langare   V
ta   förstapris   bland niorna   N
ta   hand   om brev   V
ta   intryck   av kritik   V
ta   tag   i arm   V
ta_över   attityd   från håll   V
trolla  fram målsättning   ur snuttefilten   V
tänka  sig försök   med lördagsöppet   N
tänka  sig liknande   i Sverige   N
uppleva   maken   till tystnad   N
upptäcka   brist   i beräkning   N
utbyta   ord   efter lektionstid   V
utnyttja   möjlighet   till omvandling   N
vara   advokat   i Växjö   V
vara   artikel   i serien   N
vara   artikel   om Butler   N
vara   besök   på kockskola   N
vara   bild   av generationsväxling   N
vara   bra   i bransch   A
vara   början   på marknadssanering   N
vara   del   av rörelse   N
vara   docent   i arbete   N
vara   drag   från sida   N
vara   drömstart   för oss   N
vara   effekt   av samgåendet   N
vara   ensam   i villa   A
vara   expert   på posthistoria   N
vara   fjärdedel   av arbetsstyrka   N
vara   gåva   till människa   N
vara   hållning   i drag   N
vara   improvisation   i tecken   N
vara   kommunalråd   i Tyresö   N
vara   krona   över börskurs   N
vara   kränkning   av medmänniska   N
vara   landsman   med poäng   N
vara   lokal   utan musik   N
vara   löneskillnad   på befattning   N
vara   marknad   för tobaksförsäljning   N
vara   milstolpe   i liv   N
vara   musikkritiker   i SvD   N
vara   märke   för parfym   N
vara   anisation   med roll   N
vara   part   i mål   N
vara   sak   för oss   N
vara   serb   från Sarajevo   N
vara   sida   av hemlighet   N
vara   skärpning   av kvinnoprästbråk   N
vara   slag   efter Matthew   N
vara   storbildstavla   över Sverige   N
vara   symbol   för jämlikhet   N
vara   sångare   i KFUM-kören   N
vara   tillstånd   för mig   N
vara   tragisk   för minkar   A
vara   triumf   för kollektivet   N
vara   typ   av kontor   N
vara   undantag   i varukvintet   N
vara   undertecknare   av debattartikel   N
vara   upplevelse   för Hamlet   N
vara   valör   på ordet   N
vara   varv   med dammsugare   N
vara   verksamhet   i form   N
vara   överraskning   i laget   N
visa   OS-form   med assist   V
```