# Linguistic features in data-driven dependency parsing

**Lilja Øvrelid**
NLP-unit, Dept. of Swedish
University of Gothenburg
Sweden
lilja.ovrelid@svenska.gu.se

## Abstract

This article investigates the effect of a set of linguistically motivated features on argument disambiguation in data-driven dependency parsing of Swedish. We present results from experiments with gold standard features, such as animacy, definiteness and finiteness, as well as corresponding experiments where these features have been acquired automatically and show significant improvements both in overall parse results and in the analysis of specific argument relations, such as subjects, objects and predicatives.

## 1 Introduction

Data-driven dependency parsing has recently received extensive attention in the parsing community and impressive results have been obtained for a range of languages (Nivre et al., 2007). Even with high overall parsing accuracy, however, data-driven parsers often make errors in the assignment of argument relations such as subject and object and the exact influence of data-derived features on the parsing accuracy for specific linguistic constructions is still relatively poorly understood. There are a number of studies that investigate the influence of different features or representational choices on overall parsing accuracy, (Bod, 1998; Klein and Manning, 2003). There are also attempts at a more fine-grained analysis of accuracy, targeting specific linguistic constructions or grammatical functions (Carroll and Briscoe, 2002; Kübler and Prokić, 2006; McDonald and Nivre, 2007). But there are few studies that combine the two perspectives and try to tease apart the influence of different features on the analysis of specific constructions, let alone motivated by a thorough linguistic analysis.

In this paper, we investigate the influence of a set of linguistically motivated features on parse results for Swedish, and in particular on the analysis of argument relations such as subjects, objects and subject predicatives. Motivated by an error analysis of the best performing parser for Swedish in the CoNLL-X shared task, we extend the feature model employed by the parser with a set of linguistically motivated features and go on to show how these features may be acquired automatically. We then present results from corresponding parse experiments with automatic features.

The rest of the paper is structured as follows. In section 2 we present relevant properties of Swedish morphosyntax, as well as the treebank and parser employed in the experiments. Section 3 presents an error analysis of the baseline parser and we go on to motivate a set of linguistic features in section 4, which are employed in a set of experiments with gold standard features, discussed in section 5. Section 6 presents the automatic acquisition of these features, with a particular focus on animacy classification and in section 7 we report parse experiments with automatic features.

## 2 Parsing Swedish

Before we turn to a description of the treebank and the parser used in the experiments, we want to point to a few grammatical properties of Swedish that will be important in the following:

**Verb second (V2)** Swedish is, like the majority of Germanic languages a V2-language; the finite verb always resides in second position in





*CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning,* pages 25–32
Manchester, August 2008

declarative main clauses.

**Word order variation** Pretty much any constituent may occupy the sentence-initial position, but subjects are most common.

**Limited case marking** Nouns are only inflected for genitive case. Personal pronouns distinguish nominative and accusative case, but demonstratives and quantifying pronouns are case ambiguous (like nouns).

### 2.1 Treebank: Talbanken05

Talbanken05 is a Swedish treebank converted to dependency format, containing both written and spoken language (Nivre et al., 2006a).[1] For each token, Talbanken05 contains information on word form, part of speech, head and dependency relation, as well as various morphosyntactic and/or lexical semantic features. The nature of this additional information varies depending on part of speech:

| | |
|---|---|
| NOUN: | *definiteness, animacy, case* (Ø/GEN) |
| PRO: | *animacy, case* (Ø/ACC) |
| VERB: | *tense, voice* (Ø/PA) |

### 2.2 Parser: MaltParser

We use the freely available MaltParser,[2] which is a language-independent system for data-driven dependency parsing. MaltParser is based on a deterministic parsing strategy, first proposed by Nivre (2003), in combination with treebank-induced classifiers for predicting the next parsing action. Classifiers can be trained using any machine learning approach, but the best results have so far been obtained with support vector machines, using LIBSVM (Chang and Lin, 2001). MaltParser has a wide range of parameters that need to be optimized when parsing a new language. As our baseline, we use the settings optimized for Swedish in the CoNLL-X shared task (Nivre et al., 2006b), where this parser was the best performing parser for Swedish. The only parameter that will be varied in the later experiments is the feature model used for the prediction of the next parsing action. Hence, we need to describe the feature model in a little more detail.

MaltParser uses two main data structures, a stack (S) and an input queue (I), and builds a dependency graph (G) incrementally in a single left-to-right pass over the input. The decision that needs to be made at any point during this derivation is (a) whether to add a dependency arc (with some label) between the token on top of the stack (*top*) and the next token in the input queue (*next*), and (b) whether to pop *top* from the stack or push *next* onto the stack. The features fed to the classifier for making these decisions naturally focus on attributes of *top*, *next* and neighbouring tokens in S, I or G. In the baseline feature model, these attributes are limited to the word form (FORM), part of speech (POS), and dependency relation (DEP) of a given token, but in later experiments we will add other linguistic features (FEATS). The baseline feature model is depicted as a matrix in Table 1, where rows denote tokens in the parser configuration (defined relative to S, I and G) and columns denote attributes. Each cell containing a $+$ corresponds to a feature of the model.

| | FORM | POS | DEP | FEATS |
|---|---|---|---|---|
| S:*top* | + | + | + | + |
| S:*top*+1 | | + | | |
| I:*next* | + | + | | + |
| I:*next*−1 | + | | | + |
| I:*next*+1 | + | + | | + |
| I:*next*+2 | | + | | |
| G: head of *top* | + | | | + |
| G: left dep of *top* | | | + | |
| G: right dep of *top* | | | + | |
| G: left dep of *next* | + | | + | + |
| G: left dep of head of *top* | | | + | |
| G: left sibling of right dep of *top* | | | + | |
| G: right sibling of left dep of *top* | + | | | + |
| G: right sibling of left dep of *next* | | + | + | |

Table 1: Baseline and extended (FEATS) feature model for Swedish; S: stack, I: input, G: graph; $\pm n$ = $n$ positions to the left($-$) or right ($+$)

## 3 Baseline and Error Analysis

The written part of Talbanken05 was parsed employing the baseline feature model detailed above, using 10-fold cross validation for training and testing. The overall result for unlabeled and labeled dependency accuracy is 89.87 and 84.92 respectively.[3]

Error analysis shows that the overall most frequent errors in terms of dependency relations involve either various adverbial relations, due to PP-attachment ambiguities and a large number of ad-

[1] The written sections of the treebank consist of professional prose and student essays and amount to 197,123 running tokens, spread over 11,431 sentences.

[2] http://w3.msi.vxu.se/users/nivre/research/MaltParser.html

[3] Note that these results are slightly better than the official CoNLL-X shared task scores (89.50/84.58), which were obtained using a single training-test split, not cross-validation. Note also that, in both cases, the parser input contained gold standard part-of-speech tags.

| Gold | Sys | before | after | Total |
|---|---|---|---|---|
| SS | OO | 103 (23.1%) | 343 (76.9%) | 446 (100%) |
| OO | SS | 103 (33.3%) | 206 (66.7%) | 309 (100%) |

Table 2: Position relative to verb for confused subjects and objects

verbial labels, or the argument relations, such as subjects, direct objects, formal subjects and subject predicatives. In particular, confusion of argument relations are among the most frequent error types with respect to dependency assignment.[4]

Swedish exhibits some ambiguities in word order and morphology which follow from the properties discussed above. We will exemplify these factors through an analysis of the errors where subjects are assigned object status (SS_OO) and vice versa (OO_SS). The confusion of subjects and objects follows from lack of sufficient formal disambiguation, i.e., simple clues such as word order, part-of-speech and word form do not clearly indicate syntactic function.

With respect to word order, subjects and objects may both precede or follow their verbal head. Subjects, however, are more likely to occur preverbally (77%), whereas objects typically occupy a postverbal position (94%). We would therefore expect postverbal subjects and preverbal objects to be more dominant among the errors than in the treebank as a whole (23% and 6% respectively). Table 2 shows a breakdown of the errors for confused subjects and objects and their position with respect to the verbal head. We find that postverbal subjects (after) are in clear majority among the subjects erroneously assigned the object relation. Due to the V2 property of Swedish, the subject must reside in the position directly following the finite verb whenever another constituent occupies the preverbal position, as in (1) where a direct object resides sentence-initially:

| | | | | |
|---|---|---|---|---|
| (1) | *Samma* | *erfarenhet* | *gjorde* | **engelsmännen** |
| | same | experience | made | englishmen-DEF |

‘The same experience, the Englishmen had’

For the confused objects we find a larger proportion of preverbal elements than for subjects, which is the mirror image of the normal distribution of syntactic functions among preverbal elements. As Table 2 shows, the proportion of preverbal elements among the subject-assigned objects (33.3%) is notably higher than in the corpus as a whole, where preverbal objects account for a miniscule 6% of all objects.

In addition to the word order variation discussed above, Swedish also has limited morphological marking of syntactic function. Nouns are marked only for genitive case and only pronouns are marked for accusative case. There is also syncretism in the pronominal paradigm where the pronoun is invariant for case, e.g. *det, den* ‘it’, *ingen/inga* ‘no’, and may, in fact, also function as a determiner. This means that, with respect to word form, only the set of unambiguous pronouns clearly indicate syntactic function. In the errors, we find that nouns and functionally ambiguous pronouns dominate the errors where subjects and objects are confused, accounting for 84.5% of the SS_OO and 93.5% of the OO_SS errors.

The initial error analysis shows that the confusion of argument relations constitutes a frequent and consistent error during parsing. Ambiguities in word order and morphological marking constitute a complicating factor and we find cases that deviate from the most frequent word order patterns and are not formally disambiguated by part-of-speech information. It is clear that we in order to resolve these ambiguities have to examine features beyond syntactic category and linear word order.

## 4 Linguistic features for argument disambiguation

Argument relations tend to differ along several linguistic dimensions. These differences are found as statistical tendencies, rather than absolute requirements on syntactic structure. The property of *animacy*, a referential property of nominal elements, has been argued to play a role in argument realization in a range of languages see de Swart et.al. (2008) for an overview. It is closely correlated with the semantic property of agentivity, hence subjects will tend to be referentially animate more often than objects. Another property which may differentiate between the argument functions is the property of *definiteness*, which can be linked with a notion of givenness, (Weber and Müller, 2004). This is reflected in the choice of referring expression for the various argument types in

[4] We define argument relations as dependency relations which obtain between a verb and a dependent which is subcategorized for and/or thematically entailed by the verb. Note that arguments are not distinguished structurally from non-arguments, like adverbials, in dependency grammar, but through dependency label.

Talbanken05 – subjects are more often pronominal (49.2%), whereas objects and subject predicatives are typically realized by an indefinite noun (67.6% and 89.6%, respectively). As mentioned in section 2, there are categorical constraints which are characteristic for Swedish morphosyntax. Even if the morphological marking of arguments in Scandinavian is not extensive or unambiguous, *case* may distinguish arguments. Only subjects may follow a finite verb and precede a non-finite verb and only complements may follow a non-finite verb. Information on *tense* or the related *finiteness* is therefore something that one might assume to be beneficial for argument analysis. Another property of the verb which clearly influences the assignment of core argument functions is the *voice* of the verb, i.e., whether it is passive or active.[5]

# 5 Experiments with gold standard features

We perform a set of experiments with an extended feature model and added, gold standard information on animacy, definiteness, case, finiteness and voice, where the features were employed individually as well as in combination.

## 5.1 Experimental methodology

All parsing experiments are performed using 10-fold cross-validation for training and testing on the entire written part of Talbanken05. The feature model used throughout is the extended feature model depicted in Table 1, including all four columns.[6] Hence, what is varied in the experiments is only the information contained in the FEATS features (animacy, definiteness, etc.), while the tokens for which these features are defined remains constant. Overall parsing accuracy will be reported using the standard metrics of *labeled attachment score* (LAS) and *unlabeled attachment score* (UAS).[7] Statistical significance is checked using Dan Bikel's randomized parsing evaluation comparator.[8] Since the main focus of this article is on the disambiguation of grammatical functions, we report accuracy for specific dependency relations, measured as a balanced F-score.

## 5.2 Results

The overall results for these experiments are presented in table 3, along with p-scores. The experiments show that each feature individually causes a significant improvement in terms of overall labeled accuracy as well as performance for argument relations. Error analysis comparing the baseline parser (NoFeats) with new parsers trained with individual features reveal the influence of these features on argument disambiguation. We find that animacy influences the disambiguation of subjects from objects, objects from indirect objects as well as the general distinction of arguments from non-arguments. Definiteness has a notable effect on the disambiguation of subjects and subject predicatives. Information on morphological case shows a clear effect in distinguishing between arguments and non-arguments, and in particular, in distinguishing nominal modifiers with genitive case. The added verbal features, finiteness and voice, have a positive effect on the verbal dependency relations, as well as an overall effect on the assignment of the SS and OO argument relations. Information on voice also benefits the relation expressing the demoted agent (AG) in passive constructions, headed by the preposition *av* 'by', as in English.

The ADCV experiment which combines information on animacy, definiteness, case and verbal features shows a cumulative effect of the added features with results which differ significantly from the baseline, as well as from each of the individual experiments ($p<.0001$). We observe clear improvements for the analysis of all argument relations, as shown by the third column in table 4 which presents F-scores for the various argument relations.

# 6 Acquiring features

A possible objection to the general applicability of the results presented above is that the added information consists of gold standard annotation from a treebank. However, the morphosyntactic features examined here (definiteness, case, tense, voice) represent standard output from most part-of-speech taggers. In the following we will also

[5] We experimented with the use of tense as well as finiteness, a binary feature which was obtained by a mapping from tense to finite/non-finite. Finiteness gave significantly better results ($p<.03$) and was therefore employed in the following, see (Øvrelid, 2008b) for details.

[6] Preliminary experiments showed that it was better to tie FEATS features to the same tokens as FORM features (rather than POS or DEP features). Backward selection from this model was tried for several different instantiations of FEATS but with no significant improvement.

[7] LAS and UAS report the percentage of tokens that are assigned the correct head *with* (labeled) or *without* (unlabeled) the correct dependency label, calculated using eval.pl with default settings (http://nextens.uvt.nl/∼conll/software.html)

[8] http://www.cis.upenn.edu/∼dbikel/software.html

| | UAS | LAS | p-value |
|---|---|---|---|
| NoFeats | 89.87 | 84.92 | – |
| Anim | 89.93 | 85.10 | p<.0002 |
| Def | 89.87 | 85.02 | p<.02 |
| Case | 89.99 | 85.13 | p<.0001 |
| Verb | 90.24 | 85.38 | p<.0001 |
| ADC | 90.13 | 85.35 | p<.0001 |
| ADCV | 90.40 | 85.68 | p<.0001 |

Table 3: Overall results in gold standard experiments expressed as unlabeled and labeled attachment scores.

| | | NoFeats | Gold | Auto |
|---|---|---|---|---|
| SS | subject | 90.25 | 91.80 | 91.32 |
| OO | object | 84.53 | 86.27 | 86.10 |
| SP | subj.pred. | 84.82 | 85.87 | 85.80 |
| AG | pass. agent | 73.56 | 81.34 | 81.02 |
| ES | logical subj. | 71.82 | 73.44 | 72.60 |
| FO | formal obj. | 56.68 | 65.64 | 65.38 |
| VO | obj. small clause | 72.10 | 83.40 | 83.12 |
| VS | subj. small clause | 58.75 | 65.56 | 68.75 |
| FS | formal subj. | 71.31 | 72.10 | 71.31 |
| IO | indir. obj. | 76.14 | 77.76 | 76.29 |

Table 4: F-scores for argument relations with combined features (ADCV).

| Feature | Application |
|---|---|
| Definiteness | POS-tagger |
| Case | POS-tagger |
| Animacy - NN | Animacy classifier |
| Animacy - PN | Named Entity Tagger |
| Animacy - PO | Majority class |
| Tense (finiteness), voice | POS-tagger |

Table 5: Overview of applications employed for automatic feature acquisition.

| Class | Types | Tokens covered |
|---|---|---|
| Animate | 644 | 6010 |
| Inanimate | 6910 | 34822 |
| Total | 7554 | 40832 |

Table 6: The animacy data set from Talbanken05; number of noun lemmas (Types) and tokens in each class.

show that the property of animacy can be fairly robustly acquired for common nouns by means of distributional features from an automatically parsed corpus.

Table 5 shows an overview of the applications employed for the automatic acquisition of our linguistic features. For part-of-speech tagging, we employ MaltTagger – a HMM part-of-speech tagger for Swedish (Hall, 2003). The POS-tagger distinguishes tense and voice for verbs, nominative and accusative case for pronouns, as well as definiteness and genitive case for nouns.

### 6.1 Animacy

The feature of animacy is clearly the most challenging feature to acquire automatically. Recall that Talbanken05 distinguishes animacy for all nominal constituents. In the following we describe the automatic acquisition of animacy information for common nouns, proper nouns and pronouns.

**Common nouns** Table 6 presents an overview of the animacy data for common nouns in Talbanken05. It is clear that the data is highly skewed towards the non-person class, which accounts for 91.5% of the data instances. Due to the small size of the treebank we classify common noun *lemmas* based on their morphosyntactic distribution in a considerably larger corpus. For the animacy classification of common nouns, we construct a general *feature space* for animacy classification, which makes use of distributional data regarding syntactic properties of the noun, as well as various morphological properties. The syntactic and morphological features in the general feature space are presented below:

**Syntactic features** A feature for each dependency relation with nominal potential: (transitive) subject (SUBJ), object (OBJ), prepositional complement (PA), root (ROOT)[9], apposition (APP), conjunct (CC), determiner (DET), predicative (PRD), complement of comparative subjunction (UK). We also include a feature for the complement of a genitive modifier, the so-called ‘possessee’, (GENHD).

**Morphological features** A feature for each mor-

[9] Nominal elements may be assigned the root relation in sentence fragments which do not include a finite verb.

phological distinction relevant for a noun: gender (NEU/UTR), number (SIN/PLU), definiteness (DEF/IND), case (NOM/GEN). Also, the part-of-speech tags distinguish dates (DAT) and quantifying nouns (SET), e.g. *del, rad* 'part, row', so these are also included as features.

For extraction of distributional data for the Talbanken05 nouns we make use of the Swedish Parole corpus of 21.5M tokens.[10] To facilitate feature extraction, we part-of-speech tag the corpus and parse it with MaltParser, which assigns a dependency analysis.[11] For classification, we make use of the Tilburg Memory-Based Learner (TiMBL) (Daelemans et al., 2004).[12] and optimize the TiMBL parameters on a subset of the full data set.[13]

We obtain results for animacy classification of noun lemmas, ranging from 97.3% accuracy to 94.0% depending on the sparsity of the data. With an absolute frequency threshold of 10, we obtain an accuracy of 95.4%, which constitutes a 50% reduction of error rate over a majority baseline. We find that classification of the inanimate class is quite stable throughout the experiments, whereas the classification of the minority class of animate nouns suffers from sparse data. We obtain a F-score of 71.8% F-score for the animate class and 97.5% for the inanimate class with a threshold of 10. The common nouns in Talbanken05 are classified for animacy following a leave-one-out training and testing scheme where each of the $n$ nouns in Talbanken05 are classified with a classifier trained on $n-1$ instances. This ensures that the training and test instances are disjoint at all times. Moreover, the fact that the distributional data is taken from a separate data set ensures non-circularity since we are not basing the classification on gold standard parses.

**Proper nouns** In the task of named entity recognition (NER), proper nouns are classified according to a set of semantic categories. For the annotation of proper nouns, we make use of a named entity tagger for Swedish (Kokkinakis, 2004), which is a rule-based tagger based on finite-state rules, supplied with name lists, so-called "gazetteers". The tagger distinguishes the category 'Person' for human referring proper nouns and we extract information on this category.

**Pronouns** A subset of the personal pronouns in Scandinavian, as in English, clearly distinguish their referent with regard to animacy, e.g. *han, det* 'he, it'. There is, however, a quite large group of third person plural pronouns which are ambiguous with regards to the animacy of their referent, e.g., *de, dem, deras* 'they, them, theirs'. Pronominal reference resolution is a complex task which we will not attempt to solve in the present context. The pronominal part-of-speech tags from the part-of-speech tagger distinguish number and gender and in the animacy classification of the personal pronouns we classify based on these tags only. We employ a simple heuristic where the pronominal tags which had more than 85% human instances in the gold standard are annotated as human.[14] The pronouns which are ambiguous with respect to animacy are not annotated as animate.

In table 7 we see an overview of the accuracy of the acquired features, i.e., the percentage of correct instances out of all instances. Note that we adhere to the general annotation strategy in Talbanken05, where each dimension (definiteness, case etc.) contains a null category ∅, which expresses the lack of a certain property. The acquisition of the morphological features (definiteness, case, finiteness and voice) are very reliable, with accuracies from 96.9% for voice to 98.5% for the case feature.

It is not surprising that we observe the largest discrepancies from the gold standard annotation in the automatic animacy annotation. In general, the annotation of animate nominals exhibits a decent precision (95.7) and a lower recall (61.3). The automatic classification of human common nouns

[10] Parole is available at http://spraakbanken.gu.se

[11] For part-of-speech tagging, we employ the MaltTagger – a HMM part-of-speech tagger for Swedish (Hall, 2003). For parsing, we employ MaltParser with a pretrained model for Swedish, which has been trained on the tags output by the tagger. It makes use of a smaller set of dependency relations than those found in Talbanken05.

[12] TiMBL is freely available at http://ilk.uvt.nl/software.html

[13] For parameter optimization we employ the paramsearch tool, supplied with TiMBL, see http://ilk.uvt.nl/software.html. Paramsearch implements a hill climbing search for the optimal settings on iteratively larger parts of the supplied data. We performed parameter optimization on 20% of the total $>0$ data set, where we balanced the data with respect to frequency. The resulting settings are $k = 11$, GainRatio feature weighting and Inverse Linear (IL) class voting weights.

[14] A manual classification of the individual pronoun lemmas was also considered. However, the treebank has a total of 324 different pronoun forms, hence we opted for a heuristic classification of the part-of-speech tags instead.

| Dimension | Features | Instances | Correct | Accuracy |
|---|---|---|---|---|
| Definiteness | DD, Ø | 40832 | 40010 | 98.0 |
| Case | GG, AA, Ø | 68313 | 67289 | 98.5 |
| Animacy$_{NNPNPO}$ | HH, Ø | 68313 | 61295 | 89.7 |
| Animacy$_{NN}$ | HH, Ø | 40832 | 37952 | 92.9 |
| Animacy$_{PN}$ | HH, Ø | 2078 | 1902 | 91.5 |
| Animacy$_{PO}$ | HH, Ø | 25403 | 21441 | 84.4 |
| Finiteness | FV, Ø | 30767 | 30035 | 97.6 |
| Voice | PA, Ø | 30767 | 29805 | 96.9 |

Table 7: Accuracy for automatically acquired linguistic features.

| | Gold | | Automatic | | |
|---|---|---|---|---|---|
| | UAS | LAS | UAS | LAS | p-value |
| NoFeats | 89.87 | 84.92 | 89.87 | 84.92 | – |
| Def | 89.87 | 85.02 | 89.88 | 85.03 | p<0.01 |
| Case | 89.99 | 85.13 | 89.95 | 85.11 | p<.0001 |
| Verb | 90.24 | 85.38 | 90.12 | 85.26 | p<.0001 |
| Anim | 89.93 | 85.10 | 89.86 | 85.01 | p<.03 |
| ADC | 90.13 | 85.35 | 90.01 | 85.21 | p<.0001 |
| ADCV | 90.40 | 85.68 | 90.27 | 85.54 | p<.0001 |

Table 8: Overall results in experiments with automatic features compared to gold standard features.

(Animacy$_{NN}$) also has a quite high precision (94.2) in combination with a lower recall (55.5). The named-entity recognizer (Animacy$_{PN}$) shows more balanced results with a precision of 97.8 and a recall of 85.2 and the heuristic classification of the pronominal part-of-speech tags (Animacy$_{PO}$) gives us high precision (96.3) combined with lower recall (62.0) for the animate class.

## 7 Experiments with acquired features

The experimental methodology is identical to the one described in 5.1 above, the only difference being that the linguistic features are acquired automatically, rather than being gold standard. In order to enable a direct comparison with the results from the earlier experiments, we employ the gold standard part-of-speech tags, as before. This means that the set for which the various linguistic features are defined is identical, whereas the feature values may differ.

Table 8 presents the overall results with automatic features, compared to the gold standard results and p-scores for the difference of the automatic results from the NoFeats baseline. As expected, we find that the effect of the automatic features is generally lower than their gold standard counterparts. However, all automatic features improve significantly on the NoFeats baseline. In the error analysis we find the same tendencies in terms of improvement for specific dependency relations.

The morphological argument features from the POS-tagger are reliable, as we saw above, and we observe almost identical results to the gold standard results. The addition of information on definiteness causes a significant improvement ($p<.01$), and so does the addition of information on case ($p<.0001$). The addition of the automatically acquired animacy information results in a smaller, but significant improvement of overall results even though the annotation is less reliable ($p<.03$). An interesting result is that the automatically acquired information on animacy for common nouns actually has a significantly better effect than the gold standard counterparts due to capturing distributional tendencies (Øvrelid, 2008a). As in the gold standard experiments, we find that the features which have the most notable effect on performance are the verbal features ($p<.0001$).

In parallel with the results achieved with the combination of gold standard features, we observe improvement of overall results compared to the baseline ($p<.0001$) and each of the individual features when we combine the features of the arguments (ADC; $p<.01$) and the argument and verbal features (ADCV; $p<.0001$). Column 4 in Table 4 shows an overview of performance for the argument relations, compared to the gold standard experiments. We find overall somewhat lower results in the experiment with automatic features, but find the same tendencies with the automatically acquired features.

## 8 Conclusion

An error analysis of the best performing data-driven dependency parser for Swedish revealed consistent errors in dependency assignment, namely the confusion of argument functions. We established a set of features expressing distinguishing semantic and structural properties of arguments such as animacy, definiteness and finiteness and performed a set of experiments with gold standard features taken from a treebank of Swedish. The experiments showed that each feature individually caused an improvement in terms of overall labeled accuracy and performance for the argument relations. We furthermore found that the results may largely be replicated with automatic features and a generic part-of-speech tagger. The features were acquired automatically employing a part-of-speech tagger, a named-entity recognizer and an animacy classifier of common noun lemmas employing morphosyntactic distributional features. A set of corresponding experiments with automatic features gave significant improvement from the addition of individual features and a cumulative effect of the same features in combination. In particular, we show that the very same tendencies in improvement for specific argument relations such as subjects, objects and predicatives may be obtained using automatically acquired features.

Properties of the Scandinavian languages connected with errors in argument assignment are not isolated phenomena. A range of other languages exhibit similar properties, for instance, Italian exhibits word order variation, little case, syncretism in agreement morphology, as well as pro-drop; German exhibits a larger degree of word order variation in combination with quite a bit of syncretism in case morphology; Dutch has word order variation, little case and syncretism in agreement morphology. These are all examples of other languages for which the results described here are relevant.

## References

Bod, Rens. 1998. *Beyond Grammar: An experience-based theory of language*. CSLI Publications, Stanford, CA.

Carroll, John and Edward Briscoe. 2002. High precision extraction of grammatical relations. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 134–140.

Chang, Chih-Chung and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Daelemans, Walter, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. 2004. TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. Technical report, ILK Technical Report Series 04-02.

de Swart, Peter, Monique Lamers, and Sander Lestrade. 2008. Animacy, argument structure and argument encoding: Introduction to the special issue on animacy. *Lingua*, 118(2):131–140.

Hall, Johan. 2003. A probabilistic part-of-speech tagger with suffix probabilities. Master's thesis, Växjö University, Sweden.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 423–430.

Kokkinakis, Dimitrios. 2004. Reducing the effect of name explosion. In *Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks*.

Kübler, Sandra and Jelena Prokić. 2006. Why is German dependency parsing more reliable than constituent parsing? In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT)*, pages 7–18.

McDonald, Ryan and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing. In *Proceedings of the Eleventh Conference on Computational Natural Language Learning (CoNLL)*, pages 122–131.

Nivre, Joakim, Jens Nilsson, and Johan Hall. 2006a. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

Nivre, Joakim, Jens Nilsson, Johan Hall, Gülşen Eryiğit, and Svetoslav Marinov. 2006b. Labeled pseudo-projective dependency parsing with Support Vector Machines. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Øvrelid, Lilja. 2008a. *Argument Differentiation. Soft constraints and data-driven models*. Ph.D. thesis, University of Gothenburg.

Øvrelid, Lilja. 2008b. Finite matters: Verbal features in data-driven parsing of Swedish. In *Proceedings of the International Conference on NLP, GoTAL 2008*.

Weber, Andrea and Karin Müller. 2004. Word order variation in German main clauses: A corpus analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 71–77.