Finite matters
Verbal features in data-driven parsing of Swedish

Lilja Øvrelid
NLP-unit, Dept. of Swedish
University of Gothenburg

Abstract. This paper investigates the effect of a set of verbal features in data-driven dependency parsing of Swedish. Following an error analysis of a baseline parser, we show that the addition of information on verbal features such as tense and voice can give significant improvements over this baseline and, in particular, in the analysis of syntactic arguments. We furthermore show the importance of the binary property of finiteness for the parsing of Scandinavian and demonstrate that highly similar effects may be achieved with automatically acquired information.

1 Introduction

With the development of syntactic treebanks for a range of languages other than English, there is now considerable efforts in data-driven parsing and studies which highlight the effect of different linguistic properties of these languages are important for further improvements. In recent work on syntactic parsing of German, for instance, it has been debated whether certain structural properties of the language call for different parsing strategies or representational frameworks than the parsing of English (Dubey and Keller, 2003; Kübler et al., 2006). The Scandinavian languages share with German certain syntactic properties, such as a rigid verb placement in combination with word order variation, which make these languages interestingly different from English. In particular, the so-called V2-constraint requires that the finite verb be the second constituent of declarative main clauses and finiteness has been claimed to be a defining property of Scandinavian syntax in more theoretically oriented work (Holmberg and Platzack, 1995; Eide, 2008).

In strictly data-driven approaches to syntactic parsing, a grammar, whether hand-crafted or induced, does not figure at all. The parser is trained on a treebank containing the correct analyses with respect to some representational framework, e.g., constituent analysis or dependency analysis and without a formal grammar to guide parsing, data-driven models typically condition on a rich linguistic context in the search for the most probable analysis.

In this paper we address the effect of a set of verbal features on the data-driven dependency parsing of Swedish, and in particular on the parsing of core
grammatical functions such as subjects and objects. The parsing framework is deterministic classifier-based dependency parsing, more precisely the MaltParser system (Joakim Nivre and Nilsson, 2006), which achieved the highest parsing accuracy for Swedish in the CoNLL-X shared task on dependency parsing (Buchholz and Marsi, 2006).

The paper is organized as follows. In section 2, we start out by briefly outlining some relevant syntactic properties of Scandinavian and we present the treebank and parser employed in section 3. Section 4 presents an in-depth error analysis of the results from a baseline parser, focusing on errors for syntactic dependents of verbs, like arguments and adverbials. The experiments presented in section 5 investigate the effect of additional information on the morphosyntactic properties of verbs, employing gold standard annotation. We evaluate the results both in terms of overall parse performance, as well as more detailed evaluation for individual dependency relations. We go on to assess the scalability of these results by employing automatically acquired verbal features. Finally, section 6 concludes and provides some suggestions for future research.

2 Scandinavian morphosyntax

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. Like the majority of Germanic languages, but unlike English, the Scandinavian languages are verb second (V2); the finite verb is the second constituent in declarative main clauses. Pretty much any constituent may occupy the sentence-initial position, as illustrated by (1)-(3).

(1) Statsministern håller ett tal i morgon
    primeminister-DEF holds a speech in tomorrow
    ‘The primeminister gives a speech tomorrow’
(2) Ett tal håller statsministern i morgon
    a speech holds primeminister-DEF in tomorrow
    ‘A speech, the primeminister gives tomorrow’
(3) I morgon håller statsministern ett tal
    in tomorrow holds primeminister-DEF a speech
    ‘Tomorrow, the primeminister gives a speech’

In (1) sentence-initial position is occupied by the subject, in (2) by the direct object, whereas we in (3) find an adverbial in sentence-initially. Word order in subordinate clauses, however, are not restricted by this constraint:
Non-finite verbs follow the finite verb, but precede their complements and the presence of a non-finite verb introduces a greater rigidity in terms of interpretation of the clausal constituents. With respect to core arguments, only subjects may intervene between a finite and non-finite verb, as in (6), and only objects may follow the non-finite verb, as in (5):

(5) Statsministern ska hålla ett tal
    primeminister-DEF shall hold a speech
    ‘The prime minister will give a speech’

(6) Ett tal ska statsministern hålla
    a speech shall prime minister hold
    ‘A speech, the prime minister will give’

Main clauses consisting of a finite, transitive verb along with its arguments are thus structurally ambiguous, see (7), whereas the placement of a non-finite verb in the same clause clearly indicates syntactic functions, cf. (8) and (9):

(7) Vem såg Ida?
    who saw Ida
    ‘Who saw Ida / Who did Ida see?’

(8) Vem har sett Ida?
    who has seen Ida
    SUBJ OBJ
    ‘Who has seen Ida?’

(9) Vem har Ida sett?
    who has Ida seen
    OBJ SUBJ
    ‘Who has Ida seen?’

3 Data and parser

Talbanken05 is a Swedish treebank in dependency format and contains both written and spoken language (Nivre et al., 2006a). The written sections of the

---

1 In this respect Scandinavian differs from German, which positions non-finite verbs in clause final position.
treebank consist of professional prose and student essays and amount to 197,123 running tokens, spread over 11,431 sentences. Figure 1 illustrates the treebank annotation for the example sentence in 10.

(10) Därefter betalar patienten avgift med 10 kronor

‘Thereafter, the patient pays a fee of 10 kronas’

For each token, Talbanken05 contains information on word form, part of speech, head and dependency relation, as well as various morphosyntactic features. For verbs, the treebank distinguishes the categories of tense and voice, illustrated by the active, present tense verb *betalar* ‘pays’ in (10).

In the parse experiments, we employ the freely available MaltParser, which is a language-independent system for data-driven dependency parsing. It is based on a deterministic parsing strategy, in combination with treebank-induced classifiers for predicting parse transitions (Nivre, 2006). The MaltParser system allows for explicit formulation of features employed during parsing by means of a feature model. As our baseline, we use the settings optimized for Swedish in the CoNLL-X shared task Nivre et al. (2006b) where the MaltParser system was the best performing parser for Swedish. The only parameter that will be varied in the later experiments is the information contained in the features used for the prediction of the next parsing action. The baseline parser employs information on part-of-speech, lexical form and previously assigned dependency relations.

4 Error analysis of baseline parser

The written part of Talbanken05 was parsed employing the baseline parser described above, using 10-fold cross validation for training and testing. The over-
all result for unlabeled and labeled dependency accuracy is 89.87 and 84.92 respectively.\(^3\)

<table>
<thead>
<tr>
<th>Gold System Count</th>
<th>ET OA</th>
<th>450</th>
</tr>
</thead>
<tbody>
<tr>
<td>ss oo</td>
<td>446</td>
<td></td>
</tr>
<tr>
<td>OA ET</td>
<td>410</td>
<td></td>
</tr>
<tr>
<td>AA RA</td>
<td>404</td>
<td></td>
</tr>
<tr>
<td>AA OA</td>
<td>398</td>
<td></td>
</tr>
<tr>
<td>TA AA</td>
<td>372</td>
<td></td>
</tr>
<tr>
<td>RA AA</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>oo ss</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>RA OA</td>
<td>308</td>
<td></td>
</tr>
<tr>
<td>AA TA</td>
<td>290</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. 10 most frequent error types in baseline experiment, where ss=subject, oo=object, AA=other adverbial, OA=object adverbial, ET=nominal post-modifier, RA=spatial adverbial, TA=time adverbial

In an error analysis of the baseline parser we try to locate consistent patterns of errors. As Table 1 shows, the overall most frequent errors in terms of dependency relations involve either various adverbial relations or the core argument relations of subject (ss) and direct object (oo). The errors in assignment of adverbial relations contain a fair number of PP-attachment errors (et, oo). Furthermore, Talbanken05 makes numerous and fine-grained distinctions in adverbial functions (spatial, temporal, modal, comparative etc.), which clearly prove difficult for the parser to replicate.

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. This is a direct consequence of the word order variation mentioned initially. As we saw in section 2, subjects and objects may both precede or follow their verbal head. These realizations are not, however, equally likely. Subjects, however, are more likely to occur preverbally (77%), whereas objects typically occupy a postverbal position (94%). Based on word order alone we would expect postverbal subjects and preverbal objects to be more dominant among the errors than in the treebank as a whole (23% and 6% respectively), since they display word order variants that depart from the canonical ordering of arguments. Table 2 shows a breakdown of the errors for confused subjects and objects and their position with respect to the

\(^3\) 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.
verbal head. We find that postverbal subjects (after) are in clear majority among

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>103 (23.1%)</td>
<td>343 (76.9%)</td>
<td>446 (100%)</td>
</tr>
<tr>
<td>OO</td>
<td>103 (33.3%)</td>
<td>206 (66.7%)</td>
<td>309 (100%)</td>
</tr>
</tbody>
</table>

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

the subjects erroneously assigned the object relation. Due to the aforementioned
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 examples (2)-(3) and the authentic error example in (10) above.

Following the error analysis, we may hypothesize that additional information
regarding properties of the verb may contribute to the resolution of these
types of ambiguities. As we saw in section 2, the V2-constraint is a categorical
constraint in Swedish. The property of being data-driven entails that there is no
grammar available for parsing where such a constraint may be stated explicitly.
Rather, analyses produced by the parser are patterned on properties found in the
treebank employed for training and these are the properties which we will be
manipulating in the following experiments.

5 Experiments

Part-of-speech tag sets commonly make reference to the feature of tense, a cat-
egory which is marked morphologically in Scandinavian, as in many other lan-
guages. In these experiments we will investigate the effect of verbal properties
on the analysis of syntactic arguments, such as subjects and objects in a purely
data-driven parser.

5.1 Experimental methodology

All parsing experiments are performed using 10-fold cross-validation for train-
ning and testing on the entire written part of Talbanken05. Overall parsing ac-
curacy will be reported using the standard metrics of labeled attachment score
(LAS) and unlabeled attachment score (UAS), i.e., the percentage of tokens
that are assigned the correct head with (labeled) or without (unlabeled) the cor-
rect dependency label, calculated using eval.pl with default settings.4 Statistical
significance is checked using Dan Bikel’s randomized parsing evaluation com-
parator.5

4 http://nextens.uvt.nl/~conll/software.html
5 http://www.cis.upenn.edu/~dbikel/software.html
We furthermore report accuracy for specific dependency relations, measured as a balanced F-score. In order to summarize improvement with respect to dependency relation assignment when comparing two parsers, we rank the relations by their frequency-weighted difference of F-scores.\(^6\)

### 5.2 Gold standard features

The Talbanken05 treebank distinguishes the morphosyntactic properties of tense (present, past, imperative, subjunctive, infinitive and supine) and voice (active or passive) for all verbs. In order to investigate the influence of these various verbal features we performed a set of experiments testing the effect of this information. Three experiments were performed with different feature sets: only voice information (Voice), only tense information (Tense) and a final experiment where the categories in the tense feature were mapped to a binary distinction between finite and non-finite verb forms (Finite). The last experiment was performed in order to test explicitly for the effect of the finiteness of the verb.

<table>
<thead>
<tr>
<th></th>
<th>Unlabeled</th>
<th>Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFeats</td>
<td>89.87</td>
<td>84.92</td>
</tr>
<tr>
<td>Voice</td>
<td>89.81</td>
<td>84.97</td>
</tr>
<tr>
<td>Tense</td>
<td>90.15</td>
<td>85.27</td>
</tr>
<tr>
<td>Finite</td>
<td>90.24</td>
<td>85.33</td>
</tr>
<tr>
<td>Tense+Voice</td>
<td>90.15</td>
<td>85.28</td>
</tr>
<tr>
<td>Finite+Voice</td>
<td>90.24</td>
<td>85.38</td>
</tr>
</tbody>
</table>

Table 3. Overall results for experiments with gold standard verbal features, expressed as average unlabeled and labeled attachment scores.

**Voice** A 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. As we see in Table 3, the addition of information on voice has little effect on the results and the overall difference from the baseline is not statistically significant. This is somewhat surprising as voice alternations have such confounding effects on the argument structure and argument realization of a verb. A closer look at the results, however, reveal that we do find an improved assignment for subjects and objects, as well as the passive agent relation following from the added information.

\(^6\) For each dependency relation, the difference in F-scores is weighted by its relative frequency, \(\sum_{\text{Deprel}}^{\text{Deprel}}\), in the treebank.
The improvement in analysis of the SS and OO relation is clearly linked to verbal argument structure; a passive transitive verb does not take an object whereas its active version does.

**Tense** An experiment (Tense) was performed where we included information on verbal tense. The results in Table 3 show a significant improvement from the baseline ($p < .0001$). The added information has a positive effect on the verbal dependency relations – ROOT, MS, VG, as well as an overall effect on the assignment of the SS and OO argument relations.

The most common error types indicate that the addition of information on tense improves on the confusion of the main argument types – SS, OO mentioned in the initial error analysis. We also find that head attachment of subjects in particular improves. The subject is always attached to the finite verb in the Talbanken05 analysis, so this is not surprising.

**Finiteness** In order to ascertain the influence of finiteness, an additional experiment was performed where the various tense features were mapped to their corresponding class of ‘finite’ or ‘non-finite’.\(^7\) We see the results in Table 3 and find a significant improvement from the baseline ($p < .0001$).

It is clear that the simple property of finiteness makes the relevant distinctions shown by the tense features. In fact, the mapping to a binary dimension of finiteness causes a further improvement ($p < .03$) compared to the use of the total set of tense features. This supports the central role of finiteness in Scandinavian syntax, and V2-languages in general. As we recall, the finite verb provides a fixed position in the positioning and ordering of clausal elements. As Table 4 shows, the addition of finiteness information causes improved analysis for verbal relations, the core argument relations (SS, OO), as well as non-argument, adverbial relations (TA, AA, NA). These are all relations whose positioning is influenced by the finiteness of the verb.

In the initial error analysis we noted that errors which confused subjects for objects and vice versa were frequent and that these were typically caused by word order variation. We find that the addition of information on finiteness results in the correct assignment of 24.4% of the subjects which were initially confused for objects by the baseline parser. These are predominantly postverbal subjects (89.9%) which directly follow a finite verb. We furthermore find that 31.4% of the objects initially confused for subjects by the baseline parser and a fair number of these (45.3%) have a non-finite head verb.

\(^7\) Note that we are not equating tense and finiteness, since there are untensed forms which are still finite, e.g. the imperative (Holmberg and Platzack, 1995). Rather we map the present and past tenses, as well as the imperative to the class ‘finite’ and the rest to the ‘non-finite’ class.
Combined features  The combination of the verbal features (Tense+Voice, Finite+Voice) causes a slight, but not significant improvement over the best of the individual features (Tense, Finite).

5.3 Automatic features

In order to assess the scalability of the results detailed above, we performed an experiment where information on voice and finiteness was assigned automatically. For part-of-speech tagging, we employ the freely available MaltTagger – a HMM part-of-speech tagger for Swedish (Hall, 2003). The pretrained model for Swedish employs the SUC tagset (Gustafson-Capková and Hartmann, 2006). The SUC part-of-speech tag set distinguishes tense and voice for verbs.

The experiments with the gold standard verbal features described above clearly showed the benefit of mapping the tense values to a binary set of finiteness-features and this mapping was performed directly for the acquired features. 8 We find that the automatically assigned verbal features of finiteness and voice are very reliable, with accuracies of 97.6 and 96.9, respectively. However, the passive feature is infrequent and shows a quite low precision (74.0) due to syncretism in the s-suffix which is employed for both passives and deponent verbs. Deponent verbs are characterized by a passive s-suffix, but have an agentive semantics. Examples include hoppas ‘hope’, trivas ‘enjoy’.

It is interesting to note that the addition of the automatically acquired information on voice actually causes a small, but significant improvement in overall results (p<.03), in contrast to the gold standard experiment. Clearly, the

---

Table 4. 10 most improved dependency relations with added information on finiteness, ranked by their weighted difference of balanced F-scores.

<table>
<thead>
<tr>
<th>Dependency relation</th>
<th>Freq</th>
<th>NoFeats</th>
<th>Finite</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT root</td>
<td>.0649</td>
<td>86.71</td>
<td>88.03</td>
</tr>
<tr>
<td>SS subject</td>
<td>.1105</td>
<td>90.25</td>
<td>90.91</td>
</tr>
<tr>
<td>VG verb group</td>
<td>.0302</td>
<td>94.65</td>
<td>96.42</td>
</tr>
<tr>
<td>OO direct object</td>
<td>.0632</td>
<td>84.53</td>
<td>85.31</td>
</tr>
<tr>
<td>+F coordinated clause</td>
<td>.0099</td>
<td>52.07</td>
<td>55.45</td>
</tr>
<tr>
<td>MS coordinated clause</td>
<td>.0096</td>
<td>63.35</td>
<td>66.63</td>
</tr>
<tr>
<td>TA time adverbial</td>
<td>.0249</td>
<td>70.29</td>
<td>71.20</td>
</tr>
<tr>
<td>AA other adverbial</td>
<td>.0537</td>
<td>68.70</td>
<td>69.04</td>
</tr>
<tr>
<td>++ conjunction</td>
<td>.0422</td>
<td>90.33</td>
<td>90.67</td>
</tr>
<tr>
<td>NA negation adverbial</td>
<td>.0422</td>
<td>92.46</td>
<td>93.56</td>
</tr>
</tbody>
</table>

---

8 Present, past, imperative and subjunctive forms are mapped to the finite feature (FV), all other forms are mapped to the non-finite feature (Ø).
overgeneration indicated by the low precision actually captures generalizations which benefit the parse results. In parallel with the gold standard results, we find that the feature of finiteness causes a significant improvement in results (p<.0001). The results are somewhat lower, as is to be expected, but we find that it influences the analysis of the argument relations, as well as the verbal relations.

<table>
<thead>
<tr>
<th></th>
<th>Gold standard</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unlabeled</td>
<td>Labeled</td>
</tr>
<tr>
<td>NoFeats</td>
<td>89.87</td>
<td>84.92</td>
</tr>
<tr>
<td>Voice</td>
<td>89.81</td>
<td>84.97</td>
</tr>
<tr>
<td>Finite</td>
<td>90.24</td>
<td>85.33</td>
</tr>
<tr>
<td>Finite+Voice</td>
<td>90.24</td>
<td>85.38</td>
</tr>
</tbody>
</table>

Table 5. Overall results for experiments with automatic features.

6 Conclusion

The above experiments have shown how properties of the verb are important in syntactic parsing of Swedish. An error analysis revealed consistent errors in the assignment of syntactic relations by the baseline parser. These errors were partly caused by structural properties of the language, and, in particular, word order variation.

The fact that the Scandinavian languages are V2-languages, which position the finite verb in second position, led us to design a set of experiments where we investigated the addition of information on the verbal properties of voice and tense. We found that the addition of tense, in particular, caused a significant improvement of overall results (p<.0001). In order to further test the extent to which tense may be reduced to finiteness, we performed an experiment where we mapped the tense features to features expressing the binary category of finiteness (finite/non-finite). We observed a further improvement of results (p<.03), supporting the central role of the property of finiteness in syntactic analysis of Scandinavian. We found an improved analysis for verbal dependency relations, as well as arguments and adverbials with verbal attachment. Corresponding experiments with automatically acquired features showed slightly lower, but similar effects, highlighting the scalability of the results.

It is clear that there are other linguistic properties which influence the assignment of syntactic relations in Swedish, such as the animacy and definiteness of arguments (Øvrelid and Nivre, 2007). The placement of adverbials are also
characterized by variation in Scandinavian and in terms of future research, we would like to examine the analysis of adverbials and their interaction with verbal features as well as different features of syntactic arguments. Scalability continues to be a main concern and additions in terms of linguistic features should be acquired automatically instead of relying on gold standard annotation.
Bibliography


