

DEPENDENCY PARSING OF SPOKEN SWEDISH

Joakim Nivre

Växjö University, School of Mathematics and Systems Engineering

Uppsala University, Department of Linguistics and Philology

E-mail: nivre@msi.vxu.se

Abstract

The tremendous improvement in robustness and accuracy of natural language parsing that we have witnessed during the last decade has almost exclusively been concerned with the analysis of written texts. The development of equally accurate syntactic parsers for spoken language is one of the greatest challenges for the parsing community. In this paper, we report the first results on parsing spoken Swedish with a data-driven dependency parser previously evaluated on written texts from a wide variety of languages. We compare two different algorithms, one restricted to projective dependency structures and one that allows non-projective structures, and compare the results to those obtained for written Swedish using the same methodology. The results show that parsing accuracy is still lower for spoken language than for written language, although part of the difference can be explained by properties of the transcribed spoken corpus used in the experiments. The results also show that the capacity to derive non-projective dependency structures is more crucial for spoken Swedish than for written Swedish.

Keywords: syntactic parsing, dependency parsing, spoken language

1. Introduction

It is something of an understatement to say that both general linguistics and computational linguistics have become more empirically minded during the last decades, a development that is reflected in particular by the central role

played by corpus data in both disciplines. However, the great majority of corpora used both in empirical linguistic investigations and in data-driven approaches to computational linguistics are still collections of written texts, which are more easily accessible in digital form. Spoken language corpora normally presuppose not only recording but also transcribing speech, which tends to be very labor-intensive. One of the pioneering projects in this area is the Gothenburg Spoken Language Corpus, which has been collected by Jens Allwood and his colleagues over a period of almost thirty years and which currently comprises about 1.5 million words of transcribed spoken Swedish from a wide range of social activities (Allwood et al. 2000). A direct consequence of the relative scarcity of spoken language corpora is that data-driven methods for natural language processing, e.g., syntactic parsing, have primarily been developed and evaluated on data from written language corpora. In addition, there are limited number of studies trying to apply probabilistic parsers to transcribed speech, in particular the English Switchboard corpus (Charniak and Johnson 2001, Hale et al. 2006).

One of the most prominent trends in current research on natural language parsing is the increasing popularity of dependency-based representations, which are sometimes claimed to be better suited for languages with free or flexible word order, since they have a natural mechanism for representing discontinuous dependencies via so-called non-projective dependencies, i.e., syntactic dependencies where a dependent d is separated from its head h by words that are not transitively dependent on h . Since spoken language is often said to be characterized by less rigid syntactic constraints than written language, it is conceivable that dependency-based representations would also be well suited for parsing spoken language. In this paper, we present the first experimental results on parsing spoken Swedish with dependency-based representations, applying the MaltParser system (Nivre et al. 2007) to data from the Swedish treebank Talbanken05 (Nivre et al. 2006). We compare two different parsing algorithms, one that is limited to projective dependency structures and one that allows non-projective dependencies, and we also report results for written Swedish for comparison.

The remainder of the paper is structured as follows. Section 2 introduces MaltParser, a language-independent system for data-driven dependency parsing, and the two parsing algorithms that are used in the experiments. Section 3 describes Talbanken05, a modern reconstruction of a treebank from the 1970s, containing data from both spoken and written Swedish. Section 4 presents the experimental results and their interpretation. Section 5 concludes and makes suggestions for future research.

2. MaltParser

MaltParser is a language-independent system for data-driven dependency parsing. In learning mode, MaltParser is applied to a labeled dependency treebank in order to induce a model for labeled dependency parsing of the language represented by the treebank. In parsing mode, MaltParser is used together with the induced model to parse new sentences in that language. MaltParser is freely available for research and educational purposes.¹

The parsing methodology implemented in MaltParser is based on three essential techniques:

1. Deterministic parsing algorithms for constructing dependency graphs (Yamada and Matsumoto 2003, Nivre 2003)
2. History-based models for predicting the next parser action (Black et al. 1992, Collins 1999)
3. Discriminative machine learning for mapping histories to parser actions (Yamada and Matsumoto 2003, Nivre et al. 2004)

The system uses no grammar but relies completely on inductive learning from treebank data for the analysis of new sentences and on deterministic parsing for disambiguation. This combination of methods guarantees that the parser is both robust, producing a well-formed analysis for every input sentence, and efficient, constructing this analysis in time that is linear or quadratic in the length of the sentence (depending on the algorithm used).

MaltParser has to date been evaluated on data from nineteen different languages, consistently achieving parsing accuracy at or close to the level of the best available parsers for each language (Nivre et al. 2006, Nivre et al. 2007, Hall et al. 2007), but almost exclusively on data from written language.² Most of these results have been obtained with the parsing algorithm first proposed in Nivre (2003), which is a stack-based algorithm that parses sentences in linear time, but which can only derive projective dependency structures, i.e., structures where no syntactic dependent d is

¹ URL: <http://w3.msi.vxu.se/~nivre/research/MaltParser.html>

² The notable exception is the Japanese data set in Nivre et al. (2006), which consists of transcribed spoken dialogues from the Verbmobil project (Kawata and Bartels 2000).

separated from its head h by a word that is not transitively dependent on h .³ More recently, competitive results have also been achieved with a version of the algorithm first described by Covington (2001), which is a list-based algorithm that parses sentences in quadratic time, but which also allows non-projective dependency structures to be derived (Nivre 2007). In the experiments below, we will evaluate both these algorithms, referring to them as the *projective* and the *non-projective* algorithm, respectively.

3. Talbanken05

The name Talbanken05 refers to the reconstruction and conversion into modern annotation formats of several syntactically annotated corpora developed at Lund University in the 1970s and referred to collectively here as Talbanken76 (Einarsson 1976a, Einarsson 1976b). Talbanken05 is freely available for research and educational purposes.⁴ The complete material consists of a written language part and a spoken language part of roughly equal size. The written language part in turn consists of two sections, the so-called professional prose section (P), with data from textbooks, brochures, newspapers, etc., and a collection of high school students' essays (G). The spoken language part also has two sections, interviews (IB) and conversations and debates (SD). Altogether, the corpus contains about 360,000 running tokens.

The syntactic annotation in Talbanken76 follows the MAMBA scheme, which is described by its creators as an eclectic combination of dependency grammar, topological field analysis and immediate constituent analysis (Teleman 1974). In the creation of Talbanken05, the MAMBA annotation has been converted into four different representations, two phrase structure representations, which differ with respect to the depth of the constituency analysis and are therefore referred to as *flat phrase structure* and *deep phrase structure*, respectively, and two dependency representations, which differ only with respect to the encoding format (Malt-XML vs. CoNLL). The experiments below will make use of the dependency representation in Malt-XML, which records the following information for each token: word

³ It is worth noting that the projective algorithm can be used for non-projective parsing with the technique known as *pseudo-projective* parsing, where non-projective structures are encoded as projective structures during training and then converted back to non-projective structures in a post-processing step (Nivre and Nilsson 2005). However, this technique will not be used in the experiments below.

⁴ URL: <http://w3.msi.vxu.se/~nivre/research/Talbanken05.html>

form, part of speech, syntactic head, and dependency type. An example of a dependency graph from the IB section is given in figure 1, where tokens are represented by ovals containing transcribed word forms (or the special symbol >> for junction) with part-of-speech tags below,⁵ while dependency relations are represented by arrows pointing from head to dependent, with labels indicating dependency types.⁶

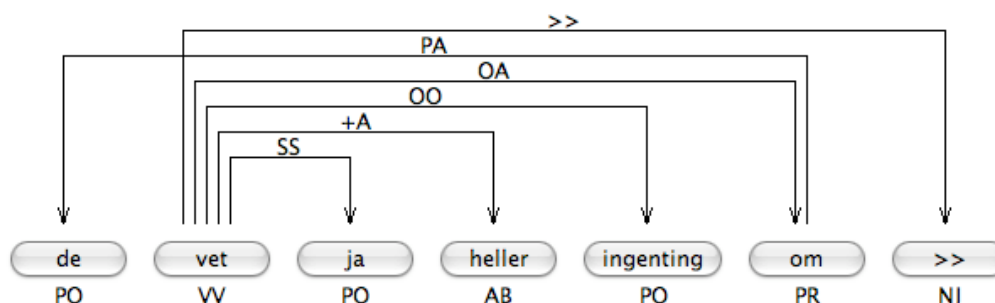


Figure 1. Dependency graph for the utterance *de vet ja heller ingenting om >>* (Talbanken05, section IB). Gloss: “that know I either nothing about [JUNCTION]”. Translation: “I don’t know anything about that either”.

Note that the dependency relation connecting the head *om* (“about”) to the dependent *de* (“that”) is non-projective, since none of the tokens occurring in between (*vet*, *ja*, *heller*) is a descendant of *om*. In phrase structure terms, *om de* (“about that”) is a discontinuous prepositional phrase. Note also that transcriptions are orthographic but often depart from standard orthography, as exemplified by the forms *de* (“that”) and *ja* (“I”), which correspond to *det* and *jag* in standard orthography. However, there is no indication as to whether a form is in standard orthography or not, which means that it is impossible to distinguish the pronoun *ja* (“I”, written *jag* in standard orthography) from the feedback word *ja* (“yes”, written *ja* in standard orthography). This is different from the Gothenburg Spoken Language Corpus, which uses an elaborate system to make sure that transcribed word forms can always be disambiguated to the level of standard orthography.

Before we move on to the parsing experiments, it should be said that the conversion from the original MAMBA annotation to dependency structures has been developed and tuned primarily for the written language sections

⁵ PO = pronoun, VV = verb, AB = adverb, PR = preposition, NJ = level juncture.

⁶ PA = preposition argument, SS = subject, +A = conjunctive adverbial, OO = direct object, OA = object adverbial, >> = juncture.

(P and G) and that it is therefore likely that the dependency representations for spoken utterances are suboptimal in certain cases. For example, the special juncture symbols (exemplified by >> in figure 1) have been treated as ordinary tokens that always attach to the nearest root of the dependency graph, which is often a source of non-projective dependency relations. This and other representational choices can probably be improved in the future, which means that the parsing results reported for spoken language below have to be taken with a pinch of salt.

4. Experiments

All four sections of Talbanken05 have been used in the experiments:

- P: Professional prose, 103,435 tokens
- G: High school essays, 105,119 tokens
- IB: Interviews, 93,261 tokens
- SD: Conversations and debates, 61,945 tokens

Each data set was divided into 80% for training, 10% for development, and 10% for testing, using a pseudo-randomized split. The results presented below are all on the development sets, leaving the test sets untouched for future studies. It is worth noting that the written data sets are larger than the spoken ones and that the SD section is by far the smallest.

For each section, MaltParser was trained on the training set and evaluated on the development set, first using the projective parsing algorithm and then using the non-projective parsing algorithm. The feature model and the learning algorithm parameters were kept constant across sections, adopting the baseline settings from Hall et al. (2007) for the projective parser and adding two additional context features for the non-projective parser. The parsing accuracy was evaluated by computing the *labeled attachment score* (LAS), i.e., the percentage of tokens that are assigned both the correct head and the correct dependency label, as well as the *unlabeled attachment score* (UAS), i.e., the percentage of tokens that are assigned the correct head (regardless of dependency label). The results for both parsers on all four sections are shown in table 1.

Table 1. Labeled attachment score (LAS) and unlabeled attachment score (UAS) for the projective and non-projective parsing algorithms on the four different sections of Talbanken05: P+G = Written, IB+SD = Spoken.

	LAS				UAS			
	Written		Spoken		Written		Spoken	
	P	G	IB	SD	P	G	IB	SD
Projective	80.1	80.2	74.7	73.4	86.3	86.5	79.7	79.6
Non-projective	79.5	80.0	78.3	76.9	85.7	86.2	83.8	83.4

The first thing to note is that the parsing accuracy is lower for spoken than for written language, with a difference of 6–7 percentage points for the projective parser and 2–3 percentage points for the non-projective parser (for both labeled and unlabeled attachment score). However, it would be premature to conclude from these results that parsing spoken language is more difficult than parsing written language. First of all, the training sets are smaller for the spoken sections, in particular for the SD section, which has the lowest accuracy scores overall. Secondly, the parser settings have not been optimized for spoken language but are based on previous research carried out almost exclusively on written language. Thirdly, as previously noted in section 3, the dependency representations for spoken language can probably be improved in several respects. When all these factors are taken into account, the results for spoken language are in fact very encouraging, indicating that it may be possible to reach the same level of accuracy for spoken language as for written language.

The second observation is that the non-projective parser has significantly higher accuracy for spoken language than the projective parser, while there is very little difference for written language (where the projective parser actually does slightly better). This difference can be explained by the fact that the spoken data sets contain a higher percentage of non-projective dependency graphs, 37.4% for IB and 38.3% for SD, as compared to only 6.7% for P and 14.4% for G. These figures must clearly be taken with a pinch of salt, given the observation in section 3 that some non-projective dependencies are due to the way in which junction markers are integrated into the dependency structures, but it is probably not too farfetched to believe that they also reflect the fact that syntactic ordering constraints are less rigid in spoken Swedish than in written Swedish. This interpretation is consistent with the observation that the G section contains twice as many non-projective structures as the P section, since high school essays can be expected to resemble spoken language more than professional prose. But a

deeper analysis is clearly needed before any firm conclusions can be drawn about the syntactic structure of spoken Swedish and about the adequacy of projective versus non-projective dependency parsing in this context.

5. Conclusion

In this paper, we have presented the first results on parsing spoken Swedish using a data-driven dependency parser. Although the parsing accuracy achieved is generally lower than for written Swedish texts with comparable syntactic annotation, there is reason to believe that this gap can be closed by better adaptation of the syntactic representations to spoken language, by properly optimizing parsers for the spoken language setting, and possibly also by increasing the amount of available training data. There is also some evidence that spoken Swedish contains more discontinuous constructions than written Swedish, which necessitates the use of parsing algorithms that can handle non-projective dependency structures.

A deeper analysis of the adequacy of dependency-based representations for spoken language is one of the most important areas for future research, but this should also take into account the basic representation used in spoken language transcriptions, so that spurious ambiguities that are due to the use of modified standard orthography can be eliminated. From this perspective, it would be very interesting to apply data-driven dependency parsing to the Gothenburg Spoken Language Corpus, which not only has systematic ways of encoding non-standard orthography but also provides much more data for training and tuning the parser.

References

- Allwood, Jens, Björnberg, Maria, Grönqvist, Leif, Ahlsén, Elisabeth and Ottesjö, Cajsa. 2000. The spoken language corpus at the Department of Linguistics, Göteborg University. *Forum: Qualitative Social Research* 1(3).
- Black, Ezra, Jelinek, Fred, Lafferty, John D., Magerman, David M., Mercer, Robert L. and Roukos, Salim. 1992. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the 5th DARPA Speech and Natural Language Workshop*, pages 31–37.

Charniak, Eugene and Johnson, Mark. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 118–126.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Covington, Michael A. 2001. A fundamental algorithm for dependency parsing. In *Proceedings of the 39th Annual ACM Southeast Conference*, pages 95–102.

Einarsson, Jan (1976a) Talbankens skriftspråkskonkordans. Lund University: Department of Scandinavian Languages.

Einarsson, Jan (1976b) Talbankens talspråkskonkordans. Lund University: Department of Scandinavian Languages.

Hale, John, Shafran, Izhak, Yung, Lisa, Dorr, Bonnie, Harper, Mary, Krasnyanskaya, Anna, Lease, Matthew, Liuh, Yang, Roark, Brian, Snover, Matthew and Stewart, Robin. 2006. PCFGs with syntactic and prosodic indicators of speech repairs. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 161–168.

Hall, Johan, Nilsson, Jens, Nivre, Joakim, Eryigit, Gülsen, Megyesi, Beáta, Nilsson, Mattias and Saers, Markus. 2007. Single malt or blended? A study in multilingual parser optimization. In *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*.

Kawata, Yasuhiro and Bartels, Julia. 2000. Stylebook for the Japanese treebank in VERBMOBIL. Verbmobil-Report 240, Seminar für Sprachwissenschaft, Universität Tübingen.

Nivre, Joakim. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160.

Nivre, Joakim. 2007. Incremental non-projective dependency parsing. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational*

Linguistics (NAACL HLT), pages 396–403.

Nivre, Joakim, Hall, Johan and Nilsson, Jens 2004. Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*, pages 49–56.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandra, Marinov, Svetoslav and Marsi, Erwin. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* **13**(2), 95–135.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Eryigit, Gülsen and Marinov, Svetoslav. 2006. Labeled pseudo-projective dependency parsing with support vector machines. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, pages 221–225.

Nivre, Joakim and Nilsson, Jens. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 99–106.

Nivre, Joakim, Nilsson, Jens and Hall, Johan. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

Yamada, Hiroyasu and Matsumoto, Yuji. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 195–206.