Building gold-standard treebanks for Norwegian

Per Erik Solberg

National Library of Norway, P.O.Box 2674 Solli, NO-0203 Oslo, Norway

per.solberg@nb.no

ABSTRACT

Språkbanken at the National Library of Norway is currently building up gold-standard Dependency Grammar treebanks for Norwegian Bokmål and Nynorsk. The treebanks are manually annotated for morphological features, syntactic functions and dependency relations. This paper explains the choice of texts and format of the treebanks, some key aspects of the morphological and syntactic annotation, and it is illustrated how the treebanks can be used.

KEYWORDS: Treebanking, Dependency Grammar, Morphology, Syntax, Norwegian.

1 Introduction

Data-driven NLP tools such as part-of-speech taggers and syntactic parsers require sufficiently large manually annotated corpora for training and testing, so-called gold-standard corpora (Brants (2000), Nivre et al. (2006a)). Språkbanken has decided to provide gold-standard Dependency Grammar treebanks for the two written standards of Norwegian - Bokmål and Nynorsk. The treebanks are manually annotated for morphological features, syntactic functions and dependency relations.

The project was initiated in October 2011, and beta versions of the treebanks are released at irregular intervals at Språkbanken's web page (see `http://www.nb.no/English/Collection-and-Services/Spraakbanken/Available-resources/Text-Resources`). Stable versions will be released in October 2013.

This paper and the system demonstration at NoDaLiDa 2013 present some important features of the treebanks: Section 2 describes the format, size and choice of texts, and section 3 lays out traits of the morphological analysis. Section 4 explains some key aspects of the syntactic annotation, and section 5 exemplifies how the treebanks can be used.

2 Choice of texts and format

At the time of writing (April 2013), the Bokmål treebank consists of approximately 204 000 tokens and the Nynorsk treebank of around 156 000 tokens. These numbers will increase significantly during the the last months before the final release. The size of the two treebanks will also be leveled out.

Comparable treebanks for other languages, such as the TIGER treebank and Prague Dependency Treebank, contain, to a large extent, newspaper text (Brants et al. (2004), Böhmová et al. (2003)). The Norwegian treebanks also contain a significant amount of text of this genre, taken from Norsk aviskorpus, a corpus of Norwegian newspaper articles from the last 15 years (`http://avis.uib.no/om-aviskorpuset/english`). Other types of factual prose, such as government reports and transcripts of parliament debates, are also included. In order to include texts in a more colloquial style, we have received permission from individual bloggers to use selected posts from their blogs.

The distribution of texts in the treebanks is currently as follows:

Newspaper text	79%
Government reports	9%
Parliament debates	6%
Blogs	6%

Table 1: Distribution of texts in the treebanks

The texts are morphologically tagged, using the *Oslo-Bergen Tagger* (OBT), (Johannessen et al. (2011)) and subsequently checked and corrected. For the syntactic preprocessing, we use the MaltParser, trained on an earlier version of the treebanks ((Nivre et al. (2006a)). The output of the syntactic parsing is checked and corrected using the annotation software TrEd (`http://ufal.mff.cuni.cz/tred/`). Most texts are syntactically annotated by one annotator, but the two annotators working in the project regularly perform double annotations to check that their annotations are consistent.

The treebanks are released in the CONLL format (Nivre et al. (2007)). The CONLL format is a 10 column tab-separated table, where each new line represents a token. The columns, from right to left, are token index, word form, lemma, coarse-grained part-of-speech (POS) tag, fine-grained POS tag, morphological features, index of head, dependency relation, and two columns which are left blank in our treebanks. We do not distinguish between coarse-grained POS tags and fine-grained POS tags in our annotations, and these columns therefore always contain the same information. In the final release, the treebanks will also be available in Prague Markup Language, the XML-based format used in the Prague Dependency Treebank.

3 Morphological annotation

The lemmatization and the morphological tagset are mostly the same for OBT and Språbanken's treebanks, although we have added a few additional morphological tags (cf. Kinn et al. (2013, 7-16)), the most important being the tag *unorm*, explained below. The lemmas are taken from *Norsk ordbank*, a lexicographic database for Norwegian (`http://www.edd.uio.no/prosjekt/ordbanken`). In addition to this, OBT generates lemmas for compounds and for adjectives formed on the basis of participial forms of verbs. The morphological tagset of OBT is rather rich: In addition to features pertaining to the inflection of words, it also contains information on whether the token belongs to a particular sub-class of a POS: Pronouns are marked as demonstrative, personal, reflexive, reciprocal, question-forming etc., determiners as demonstrative, quantifying, possessive and so on. For users of the treebanks who need a distinction between coarse- and fine-grained POS tags, it should be relatively easy to make a conversion scheme based on these tags.

A phenomenon which OBT cannot handle, is spelling variants and inflectional forms which do not comply with the official norm for Bokmål and Nynorsk, and non-compounds which do not have an entry in Norsk ordbank. In such cases, the annotators manually add a lemma and the correct morphological tags, and mark the token as not complying with the norm, using the morphological tag *unorm*, (i.e. *unormert*, 'non-standard').

4 Syntactic annotation

While the morphological analyses are based on OBT, the syntactic annotation guidelines for the treebanks have been developed specifically for this project (see Kinn et al. (2013)). The formalism for the syntactic annotation is Dependency Grammar. The sentence in example (1) and its analysis in FIGURE 1 illustrates some key features of the syntactic annotation.

(1) *Den kinesiske ledelsen har tydeligvis innsett alvoret.*
the Chinese administration has evidently realized seriousness+the
'The Chinese administration has evidently realized the seriousness.'

One token serves as head for the whole sentence. In (1), the finite auxiliary verb *har* heads the sentence and carries the root-function for finite sentences FINV. Each of the other tokens are annotated for syntactic functions and unique dependency relations to another token in the sentence. The subject is dependent on the finite verb, as subjects usually only occur in finite constructions. Other arguments and modifiers, such as the object *alvoret* and the adverb *tydeligvis* in example (1), are, however, dependents on the lexical verb. While there might be cases where it would be adequate to make e.g. an adverbial dependent on the

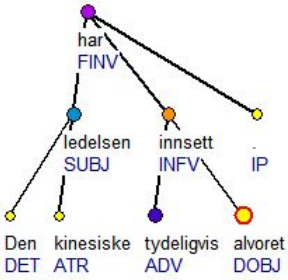

Figure 1: Analysis of (1)

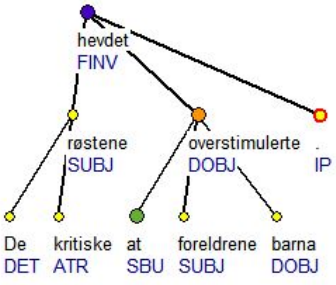

Figure 2: Analysis of (2)

auxiliary, we have chosen this analysis to ensure annotations which are as consistent and uniform as possible.

Finite subordinate clauses are always headed by the finite verb in our treebanks. Example (2) and its analysis in FIGURE 2 illustrates this.

(2) *De kritiske røstene hevdet at foreldrene overstimulerte barna.*
the critical voices claimed that parents+the overstimulated children+the
'The critical voices claimed that the parents overstimulated the children.'

The complement clause in (2) is headed by the verb *overstimulerte*, which receives the function DOBJ (*direct object*). An alternative analysis would be to make the subjunction head. In Norwegian, the subjunction can quite often be dropped in both relative and complement clauses. If all finite subordinate clauses are headed by the finite verb, it is quite simple to make queries for all subordinate clauses of a specific type, regardless of whether a subjunction is present. In the alternative analysis where the subjunction serves as head, a different analysis would have to be found for clauses lacking a subjunction, and some queries would be much more difficult to perform.

5 Use

The treebanks are mainly intended as testing and training material for data-driven NLP tools, both morphological and syntactic. Such resources have been lacking for Norwegian[1]. For example, there have been no suitable treebank for testing and training data-driven dependency parsers. To illustrate how the gold-standard treebanks can be used for this purposes, I parsed a 15 000 token text in Bokmål both with the MaltParser, and compared the results to a manually annotated version of the text, using the *CONLL-X shared task evaluation script* (`http://ilk.uvt.nl/conll/software.html`). The MaltParser was trained on the current version of the Bokmål treebank. I used the Covington's non-projective parsing algorithm and the Liblinear learning algorithm, but didn't try to optimize the parser any further (Nivre et al. (2006a), Covington (2001), Fan et al. (2008)). the MaltParser gave a

[1]But see also the INESS project at the University of Bergen: `http://iness.uib.no/`

labeled attachment score (LAS; the percentage of tokens for which both the dependency relation and the syntactic function is correct) of 85.77 %. In comparison, Nivre et al. (2006b) obtained a LAS of 84.58 % for Swedish and 84.77 % for Danish using the MaltParser, the best results for those languages in the 2006 CoNLL-X shared task on Multilingual Dependency Parsing (Buchholz and Marsi (2006)).

While the treebanks are primarily developed for NLP research and development, they can also be interesting tools for researchers within other linguistic disciplines, such as morphology, lexicography and syntax. A treebank can be used to query very specific constructions, which cannot be found in corpora with a less detailed annotation. For example, Julien (2009), who studied main clause word order in subordinate clauses in Mainland Scandinavian, wasn't able to search for fronted non-subjects in subordinate clauses in the corpora she used (Julien (2009, 5)). This could have been retrieved easily from the gold-standard treebanks, using e.g. PML Tree Query (Pajas and Štěpánek (2009)). A query in the treebanks can also immediately give all lemmas occurring e.g. as indirect object, while this kind of information is much harder to retrieve in corpora without syntactic annotation. A downside to manually annotated treebanks when used for such purposes, is that they are usually much smaller than corpora with a more shallow annotation, and it might therefore be difficult to find infrequent constructions.

6 Conclusion

This paper has presented Språkbanken's gold-standard treebanks for Norwegian Bokmål and Nynorsk, two Dependency Grammar treebanks which are manually annotated for morphological features, syntactic functions and dependency relations. The treebanks have primarily been developed with the testing and training of data-driven NLP tools in mind. Norwegian treebanks containing detailed syntactic annotation can, however, be useful and interesting resources for other purposes too, such as syntactic and lexicographic research.

Acknowledgments

Språkbanken's gold-standard treebanks are built up in close collaboration with the Text Laboratory at the University of Oslo. I would also like to thank Lilja Øverlid and Arne Skjærholt at the Department of Informatics at the University of Oslo for their valuable advice and for helping us set up a more effective syntactic preprocessing. Inquiries regarding the treebanks can be sent to *sprakbanken@nb.no*.

References

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague Dependency Treebank. In *Treebanks*, pages 103–127. Springer, Netherlands.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2(4):597–620.

Brants, T. (2000). Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231, Seattle, WA.

Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, New York, NY.

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference*, pages 95–102, Athens, GA.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Johannessen, J. B., Hagen, K., Nøklestad, A., and Lynum, A. (2011). Obt+ stat: Evaluation of a combined cg and statistical tagger. *Constraint Grammar Applications*, pages 26–34.

Julien, M. (2009). Embedded clauses with main clause word order in mainland scandinavian. *Published on LingBuzz:(http://ling. auf. net/lingBuzz/000475)*.

Kinn, K., Solberg, P. E., and Eriksen, P. K. (2013). Retningslinjer for morfologisk og syntaktisk annotasjon i Språkbankens gullkorpus. Manuscript. Språkbanken, National Library of Norway. URL: `http://www.nb.no/Tilbud/Forske/Spraakbanken/Tilgjengelege-ressursar/Tekstressursar`. [last visited on 05/04/2013].

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 915–932, Toulouse, France.

Nivre, J., Hall, J., and Nilsson, J. (2006a). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Nivre, J., Hall, J., Nilsson, J., Eryiit, G., and Marinov, S. (2006b). Labeled pseudo-projective dependency parsing with support vector machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 221–225, New York, NY.

Pajas, P. and Štěpánek, J. (2009). System for querying syntactically annotated corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, Singapore.