The Norwegian Dependency Treebank

Per Erik Solberg[*], Arne Skjærholt[†], Lilja Øvrelid[†], Kristin Hagen[‡], and Janne Bondi Johannessen[‡]

[*] Språkbanken, The National Library of Norway
[†] Department of Informatics, University of Oslo
[‡] Department of Linguistics and Scandinavian Studies, University of Oslo
per.solberg@nb.no, {arnskj,liljao}@ifi.uio.no, {kristin.hagen,j.b.johannessen}@iln.uio.no

Abstract

Each article must include an abstract of 150 to 200 words in Times 9 pt with interlinear spacing of 10 pt. The heading Abstract should be centred, font Times 10 bold. This short abstract will also be used for printing a Booklet of Abstracts containing the abstracts of all papers presented at the Conference.

Keywords: keyword A, keyword B, keyword C

1. Introduction

A syntactic treebank constitutes an important language resource in establishing a set of natural language processing tools for a language and may be employed for central tasks such as part-of-speech tagging and syntactic parsing as well as for linguistic research. For the past decade, dependency analysis has become an increasingly popular form of syntactic analysis and has been claimed to strike a balance between a depth of analysis sufficient for many down-stream applications, as well as providing accuracy and efficiency in parsing with these types representations. The CoNLL shared tasks devoted to dependency parsing and joint syntactic and semantic parsing, (Nivre et al., 2007; Hajič et al., 2009), have been instrumental in establishing a common set of dependency treebanks for a range of languages such as English, Swedish, Czech and Arabic, thus enabling multilingual evaluation of different systems. The increased availability of dependency parsers has spurred down-stream use of dependency representations in diverse tasks such as Machine Translation (Ding and Palmer, 2005), Sentiment Analysis (Wilson et al., 2009) and Negation Resolution (Lapponi et al., 2012).

Until recently, no treebank has been available for Norwegian.[1] At present, however, Språkbanken, at the Norwegian National Library, is in the process of completing a two year project with the aim of producing a dependency treebank for Norwegian.

In this paper we present the result of this project, the Norwegian Dependency Treebank (NDT)[2], a syntactic treebank which encompasses treebanks for both variants of Norwegian (Bokmål and Nynorsk).[3] We describe the main annotation principles employed in the syntactic analysis of the treebank and discuss the selection of texts. We then go on to describe the annotation process in some detail. Finally, we present first results for data-driven dependency parsing of Norwegian.

2. Annotation principles

2.1. General principles

The treebank contains both morphological and syntactic annotation. The morphological annotation follows the Oslo-Bergen Tagger (Hagen et al., 2000; Solberg, 2013).

Independent syntactic annotation guidelines for the NDT have been developed in an iterative process in the beginning of the project period by the annotators working in the project (Kinn et al., 2013).The annotation guidelines are, to a large extent, based on the Norwegian Reference Grammar (Faarlund et al., 1997). The Dependency Grammar annotations are inspired by the choices made in comparable treebanks, in particular the Swedish treebank Talbanken (Nivre et al., 2006b) and the treebank of old Indo-European languages PROIEL (Haug et al., 2009).

When developing the annotation guidelines, four fundamental principles were taken into consideration:

1. **Linguistic adequacy:** The annotation should be as linguistically adequate as possible.
2. **Consistency:** It had to be possible for annotators to implement the analyses consistently.
3. **Quick annotation:** As size matters for most users, the annotators should be able to annotate quickly.
4. **Easy retrieval:** It should be easy to retrieve specific constructions.

In what remains of this section, we will show examples of how we tried to implement these principles, and compare our choices to other annotation schemes.

2.2. Adverbials

In some treebanks comparable to the NDT, e.g. Talbanken, there are separate dependency relations for different types of adverbials, such as time adverbials, manner adverbials and place adverbials. We found that it would be difficult to have such distinctions and at the same time comply with the consistency and time constraints of principle 2 and 3.

[1] A treebank of deep linguistic analysis couched in the LFG framework is however under development at the University of Bergen by the INNESS project.

[2] In the development phase, the treebank has also been referred to as Språkbanken's Gold Standard Corpus.

[3] These are the two written varieties of Norwegian.

When making annotation choices, we also opted for analyses which are meaningful to various end user groups. In this light, a high level of linguistic detail is not always an advantage, as it becomes more difficult to infer grammatical patterns and extract meaningful information (Marneffe and Manning, 2008). A fine-grained analysis of adverbials could in fact make such tasks more difficult, as distinctions between different types of adverbials frequently would be based on semantic and pragmatic considerations only, not on difference in syntactic structure. For example, the same preposition may express different types of adverbials in very similar contexts. We therefore opted for a more shallow analysis: All adverbials, regardless of type, receive a uniform dependency relation - ADV.

2.3. Transitive and intransitive prepositions

In other cases, the pursuit of linguistic adequacy (principle 1) has been given priority. The sentences (1) and (2) exemplify such a case:

(1) *Per setter på CD-en.*
Per puts on CD+the
'Per puts on the CD.'

(2) *Per sitter på stolen.*
Per sits on chair+the
'Per sits on the chair.'

In both (1) and (2), the preposition *på* is followed by a noun. There are, however, strong reasons for analyzing the sentences differently. In (2), the noun is clearly a complement of the preposition: The preposition and the noun are semantically connected, they behave as a single constituent, and the complement retains its position after the preposition if it is pronominalized. In (1), there is no obvious semantic connection between the preposition and the noun, the two words do not form a constituent together, and if the noun is pronominalized, it usually will precede the preposition. In the NDT, the noun in constructions like (1) would be made dependent on the verb with the dependency relation for direct objects, DOBJ, while in (2) it is made dependent on the preposition with the dependency relation of prepositional complements, PUTFYLL.
Annotators frequently meet preposition-noun sequences which are less straightforward than in these examples, and they need to deliberate whether one or the other analysis is correct. We have, however, chosen to retain this distinction, to make sure that the analyses are acceptable from a linguistic point of view and also in order to achieve a uniform analysis of sentences such as (1) and cases where the object noun or pronoun does not follow the intransitive preposition (or particle, as these are also known). To ensure consistency and a high annotation speed (principle 2 and 3), the annotation guidelines have a number of syntactic tests which the annotators use to distinguish between the constructions (Kinn et al., 2013, 54-56).

2.4. Complementizers

There is no obvious head-dependent relationship between complementizers and verbs or between lexical and function words in general, and there is therefore not a unique answer to how such a relationship should be represented in Dependency Grammar. Some annotation standards treat all relations between lexical and functional heads in the same manner. In the Standford annotation standard, the lexical word is the head whenever possible (Marneffe and Manning, 2008, 2). In NDT, we have no uniform treatment of lexical and function words, but we have made a decision for each construction, based on the four principles given above. For example, in the case of complementizers and verbs, we have chosen to let the verb be the head and the complementizer a dependent on the verb. The reason for this is that complementizers are frequently dropped in Norwegian, as the following examples show (from the NDT):

(3) *Nå tror lokale myndigheter at*
now believe local authorities that.comp
bortføringen var nøye planlagt.
abduction+the was carefully planned
'Local authorities now believe that the abduction was carefully planned.'

(4) *Jeg tror ikke det er tilfeldig.*
I believe not it is accidental
'I don't belive that it is accidental.'

Clausal complements of verbs such as *tro*, 'believe', occur both with the complementizer *at*, as in (3), and without any complementizer, as in (4). If the complementizer were the head, the complement clauses in (3) and (4) would have had different heads, despite their obvious similarities. This, in turn, would make it significantly more difficult to formulate queries using standard query tools and to deduce grammatical patterns more generally (cf. principle 4). In the NDT, sentences such as (3) and (4) are analyzed similarly: The (finite) verb of the clausal complement serves as head in both cases, and carries the dependency relation DOBJ (direct object), c.f. figure 1 and 2. Both will therefore be retrieved through a query for finite verbs with the dependency relation DOBJ.
Table 1 summarizes some choices where dependency treebanks tend to differ.

Head	Dependent
Preposition	Prepositional complement
Finite verb	Complementizer
First conjunct	Subsequent conjuncts
Finite auxiliary	lexical/main verb
Noun	Determiner

Table 1: Annotation choices in the NDT

3. Texts

The NDT consist of 311 000 tokens of Norwegian Bokmål and 303 000 tokens of Norwegian Nynorsk. The texts for Bokmål and Nynorsk were collected from independent sources. Since the differences between these two written

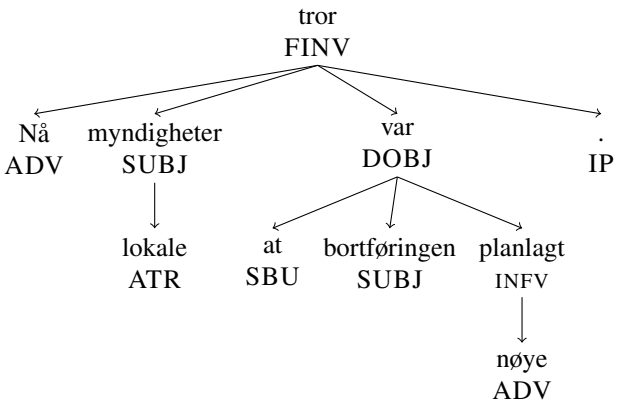

Figure 1: Analysis of (3)

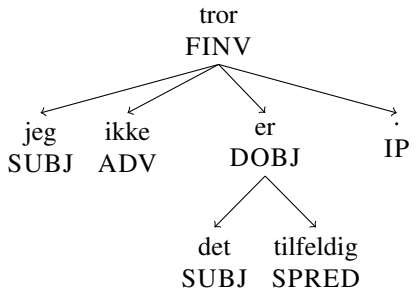

Figure 2: Analysis of (4)

standards of Norwegian are mostly lexical and morphological, the syntactic annotation is practically identical. Comparable treebanks such as the Prague Dependency Treebank and the TIGER treebank, contain mainly newspaper text (Böhmová et al., 2003; Brants et al., 2004). Other treebanks, e.g. Penn Treebank and Talbanken (Marcus et al., 1993; Nivre et al., 2006b), however, also contain texts from other sources, such as factual prose, fiction and text in a more colloquial style.

Newspaper text is frequently used for various NLP tasks and also has the advantage of being fairly standardized, unlike fiction and e.g. texts from social media. We have therefore chosen to use mostly newspaper text in the NDT, but we added small amounts of text from government reports, parliament transcripts and more colloquial texts from blogs, cf. table 2.

Source	Fraction
Newspaper text	82%
Government reports	7%
Parliament transcripts	6%
Blogs	5%

Table 2: Distribution of texts in the treebanks

Parser		UAS	LAS	Labels
CG	(BM)	79.39%	72.45%	82.10%
CG	(NN)	80.16%	74.76%	84.84%
S & Ø (2012)	(BM)	87.54%	84.63%	89.63%
Final	(BM)	89.89%	87.57%	91.70%
Final	(NN)	89.66%	87.50%	91.76%

Table 3: Preprocessor accuracies. Unlabeled (UAS) and Labeled (LAS) attachment scores, and label accuracies (Labels).

4. Annotation process

4.1. Annotators

All texts in the treebank have been manually annotated by trained linguists. A few of the texts have been syntactically annotated by two annotators, to avoid inconsistencies (cf. 4.3.). In order to speed up the annotation process, we chose to preprocess the texts using tools already available at the University of Oslo.

4.2. Preprocessing and work flow

As is standard practice when annotating syntactic corpora, the texts to be annotated are automatically PoS tagged and syntactically parsed before being annotated, an approach which has been shown to be both fast and yielding high quality annotation (Marcus et al., 1993; Fort and Sagot, 2010; Skjærholt, 2013). After tokenization, the texts are first tagged using OBT+stat, a rule-based Constraint Grammar tagger with a HMM-based overlay (Johannessen et al., 2012). The morphological annotation is then checked and corrected by an annotator using a web interface made for this particular task (Lynum, 2013). The corrected morphological annotations are then preprocessed by a dependency parser and imported into TrEd, the annotation tool developed for the Prague Dependency Treebank, which is used to correct the output of the syntactic preprocessing and create the final treebank.

The initial syntactic preprocessor was created using the syntactic module of OBT, which, while it does not create a hierarchical structure, does provide some information about heads. On top of this we built a small set of rules in the CG-3 framework (Didriksen, 2013) to build proper syntactic structures. This preprocessor was evaluated to get about 80% of heads correct (unlabeled attachment) and both head and label (labeled attachment) correct in 72–74% of cases, as shown in Table 3.

The first statistical parser trained on the corpus is that of Skjærholt and Øvrelid (2012), which was later used in inter-annotator agreement experiments by Skjærholt (2013), reported to reach a labeled accuracy of 84% and an unlabeled accuracy of 87%. Based on this, an improved parser was trained which obtains an unlabeled accuracy of nearly 90% and labeled accuracy of 87%.

	Bokmål (BM)		Nynorsk (NN)	
	UAS	LAS	UAS	LAS
Malt default	88.27	85.03	87.54	83.82
Malt optimized	92.19	89.82	91.57	88.98
MST	91.69	88.26	91.54	87.80
Bohnet&Nivre	92.94	90.56	92.78	90.18

Table 4: Dependency parsing results for the NDT

4.3. Inter-annotator agreement

To validate the consistency of the annotations produced by the different annotators, a set of experiments quantifying inter-annotator agreement were performed (Skjærholt, 2013). As is common practice in the field of syntactic annotation (Civit et al., 2003; Brants, 2000; Brants and Hansen, 2002; Hajič, 2004), the simple agreement measures labeled and unlabeled attachment accuracy were used. The reason for using an uncorrected measure rather than a chance-corrected measure such as κ or π is that these measures are not directly applicable to the task of syntactic annotation.
Skjærholt (2013) measured inter-annotator agreement as measured by labeled and unlabeled attachment, using a number of different preprocessors from the cross lingual parsers of Skjærholt and Øvrelid (2012). Here, we will concentrate on the agreement using the best parser, whose performance is shown in Table 3. Using this parser, agreement was measured to be 96.8% unlabeled and 95.3% labeled accuracy. These results are comparable to those reported for the German NEGRA (92.4% labeled F_1 (Brants, 2000)) and TIGER (93.9% labeled F_1 (Brants and Hansen, 2002)) treebanks and the Spanish Cat3LB treebank (86.9% labeled bracket precision (Civit et al., 2003)).

5. Dependency parsing

An important aspect of treebank annotation relates to its *parsability*, i.e. the quality of syntactic parsers that can be acquired based on the treebank data. In order to investigate the parser quality we can expect from the NDT, we have evaluated three state-of-the-art dependency parsers on the material: Maltparser (Nivre et al., 2006a), MSTparser (McDonald et al., 2005) and the parser of Bohnet and Nivre (2012). These implement different parsing strategies: Maltparser is a transition-based parser with local learning and greedy search, MST is a graph-based dependency parser implementing global, near-exhaustive search and the Bohnet and Nivre (2012) parser is a transition-based dependency parser with joint tagger that implements global learning and a beam search for non-projective labeled dependency parsing. This latter parser has recently outperformed pipeline systems (such as the Malt and MST parsers). For Maltparser, we trained two versions of the parser: one version with default settings and one optimized version, where the parser settings was optimized using the MaltOptimizer software (Ballesteros and Nivre, 2012). Both the MST and Bohnet and Nivre (2012) parsers were trained using default settings.
For these experiments, both portions of the treebank (Bokmål and Nynorsk) were split into 80-10-10 train, development and test sets.
Table 4 presents the dependency parsing results obtained for the NDT. We find that the Bohnet and Nivre (2012) parser outperforms the other parsers and obtains labeled accuracy scores of 90.6 and 90.2 for the BM and NN treebanks, respectively. The optimized Maltparser model performs only slightly lower, at 89.8 and 89.0. These are encouraging results which indicate that the treebank provides a good basis for parser development.

6. Conclusion

We have presented the first treebank for Norwegian, a treebank containing dependency representations for a large sample of Norwegian texts We have described the annotation principles that motivate the analyses, the collections of texts, as well as the annotation process and presented results for inter-annotator agreement, showing that the syntactic annotation is of a consistency comparable to other large treebank initiatives. Finally, we have presented the first results for Norwegian dependency parsing, contrasting three state-of-the-art data-driven dependency parsers.

7. Acknowledgements

Place all acknowledgements (including those concerning research grants and funding) in a separate section at the end of the article.

8. References

Miguel Ballesteros and Joakim Nivre. 2012. Maltoptimizer: A system for maltparser optimization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank. In *Treebanks*, pages 103–127. Springer.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1455–1465. Association for Computational Linguistics.

Sabine Brants and Silvia Hansen. 2002. Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1643–1649.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Thorsten Brants. 2000. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Montserrat Civit, Alicia Ageno, Borja Navarro, Núria Bufí, and M. Antònia Martí. 2003. Qualitative and Quantitative Analysis of Annotators' Agreement in the Development of. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 21–32, Växjö. Växjö University Press.

Tino Didriksen. 2013. Constraint Grammar manual. 3rd version of the CG formalism variant. Technical report, GrammarSoft.

Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics*, pages 541 – 548, Ann Arbor, MI, USA. Association for Computational Linguistics.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget.

Karën Fort and Benoît Sagot. 2010. Influence of Preannotation on POS-tagged Corpus Development. In Nianwen Xue and Massimo Poesio, editors, *Proceedings of the fourth linguistic annotation workshop*, pages 56–63, Stroudsburg. Association for Computational Linguistics.

Kristin Hagen, Janne Bondi Johannessen, and Anders Nøklestad. 2000. A constraint-based tagger for norwegian. In *17th Scandinavian Conference in Linguistics*, pages 31–48, Odense, Denmark.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, aluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*.

Jan Hajič. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. Jazykovedný ústav Ľ. Štúra, SAV.

Dag T. T. Haug, Marius Jøhndal, Hanne Martine Eckhoff, Eirik Welo, Mari J. B. Hertzenberg, and Angelika Muth. 2009. Computational and linguistic issues in designing a syntactically annotated parallel corpus of indo-european languages. In *Traitement Automatique des Langues*.

Janne Bondi Johannessen, Kristin Hagen, André Lynum, and Anders Nøklestad, 2012. *OBT+stat: A combined rule-based and statistical tagger*, volume 49, page 51. John Benjamins.

Kari Kinn, Pål Kristian Eriksen, and Per Erik Solberg. 2013. Retningslinjer for morfologisk og syntaktisk annotasjon i språkbankens gullkorpus. Technical report, National Library of Norway.

Emanuele Lapponi, Erik Velldal, Lilja Øvrelid, and Jonathon Read. 2012. Sequence-labeling negation using dependency features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, Montreal, Canada.

André Lynum. 2013. Tag-annotator. Software.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Catherine de Marneffe and Christopher de Manning. 2008. The Stanford typed dependencies representation. In *Proceedings of the Coling 2008 Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, UK:Manchester.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 525–530.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Joakim Nivre, 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.

Arne Skjærholt and Lilja Øvrelid. 2012. Impact of treebank characteristics on cross-lingual parser adaptation. In Iris Hendrickx, Sandra Kübler, and Kiril Simov, editors, *Proceedings of the 11th international workshop on treebanks and linguistic theories*, pages 187–198, Lisbon. Edições Colibri.

Arne Skjærholt. 2013. Influence of preprocessing on dependency syntax annotation: speed and agreement. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 28–32, Sofia. Association for Computational Linguistics.

Per Erik Solberg. 2013. Building gold-standard treebanks for norwegian. In *Proceedings of NODALIDA 2013*, Oslo, Norway.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399 – 433.