NorGramBank: A 'Deep' Treebank for Norwegian

Helge Dyvik[1], Paul Meurer[2], Victoria Rosén[1], Koenraad De Smedt[1], Petter Haugereid[1], Gyri Smørdal Losnegaard[1], Gunn I. Lyse[1], Martha Thunes[1]

University of Bergen[1], Uni Research Computing[2]
dyvik@uib.no, paul.meurer@uni.no, desmedt@uib.no, victoria@uib.no,
petterha@gmail.com, gyri.losnegaard@uib.no, gunn.lyse@uib.no, martha.thunes@uib.no

Abstract

We present NorGramBank, a treebank for Norwegian with highly detailed LFG analyses. It is one of many treebanks made available through the INESS treebanking infrastructure. NorGramBank was constructed as a parsebank, i.e. by automatically parsing a corpus, using the wide coverage grammar NorGram. One part consisting of 350,000 words has been manually disambiguated using computer-generated discriminants. A larger part of 50 M words has been stochastically disambiguated. The treebank is dynamic: by global reparsing at certain intervals it is kept compatible with the latest versions of the grammar and the lexicon, which are continually further developed in interaction with the annotators. A powerful query language, INESS Search, has been developed for search across formalisms in the INESS treebanks, including LFG c- and f-structures. Evaluation shows that the grammar provides about 85% of randomly selected sentences with good analyses. Agreement among the annotators responsible for manual disambiguation is satisfactory, but also suggests desirable simplifications of the grammar.

Keywords: treebanks, Norwegian, Lexical Functional Grammar

1. Motivation and Design

NorGramBank is the first treebank for written Norwegian (Bokmål and Nynorsk)[1] based on 'deep' parsing. It is constructed through parsing with NorGram, a manually written grammar within the framework of Lexical Functional Grammar (LFG). NorGram contains about 380 complex syntactic rules (mostly valid for both Bokmål and Nynorsk), corresponding to a transition network with more than 160,000 states and more than 4.7 M arcs. The lexicon comprises around 180,000 lemmas for Bokmål and 110,000 lemmas for Nynorsk and has more than 200 different verb frames.

Approximately 350,000 words of parsed text have been manually disambiguated and checked using computer-generated discriminants. Through stochastic disambiguation the corpus has been extended to about 50 M word tokens.

NorGramBank and many other treebanks are made available through the INESS platform (Rosén et al., 2012), which enables advanced exploration. Here authenticated researchers have access to a fully web-based interface for treebank selection, browsing and search with visualization of syntactic structures.

Since the treebank was designed as a parsed corpus, the grammar used by the parser is decisive for the annotation. The design of NorGram is based partly on the principles developed within the Parallel Grammar Project (Butt et al., 2002), which define a cross-linguistic set of features to be used in the f-structures. The ParGram features have been established through a long-term project involving grammar development for a wide range of languages and thus ensures a degree of cross-linguistic validity for the descriptive choices. Furthermore, the design is based partly on the principles for X-bar syntax and its integration in LFG proposed by Joan Bresnan (Bresnan, 2001), which govern the design of the c-structures and constrain the possible projections of f-structures from them. While c-structure design has not been a central concern in the ParGram project, Bresnan's proposals establish a cross-linguistic set of principles on this level as well.

A different treebank for Norwegian, the Norwegian Dependency Treebank (NDT), comprising 614,000 word tokens, has been constructed through manual annotation (Solberg et al., 2014). A manually annotated treebank is not guaranteed to be fully consistent with a single linguistically principled grammar, whereas such consistency is one aim of NorGramBank. Furthermore, automatic analysis allows a higher degree of linguistic detail and more fine-grained distinctions than are usually practically possible in a manually annotated treebank. With statistical disambiguation of the parsing results (see Section 3.4.) a treebank based on automatic analysis will also be able to cover far more text than a manually annotated treebank can achieve.

To illustrate the difference in the degree of detail we may compare the analysis of a sentence from the NDT treebank with the analysis of the same sentence in NorGramBank, which also contains the NDT texts (Example 1). At the same time the analysis of this example will serve to illustrate a few aspects of NorGram.

(1) *Jeg ber bare om at anklagene ikke*
I pray only about that the accusations not
er sanne.
are true
'I only pray that the accusations aren't true.'

The dependency graph in Figure 1 labels the arcs from head

[1] Bokmål and Nynorsk are the two written standard varieties of Norwegian

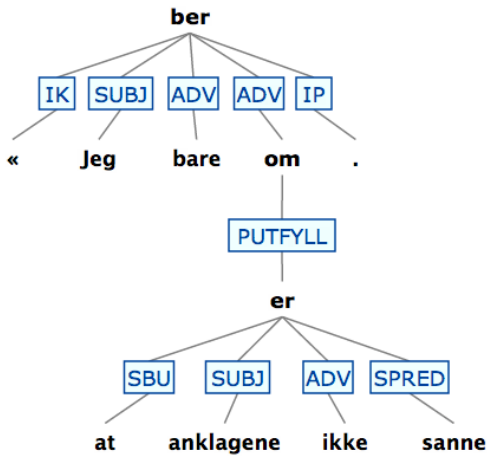

Figure 1: The NDT treebank dependency analysis of the sentence (1)

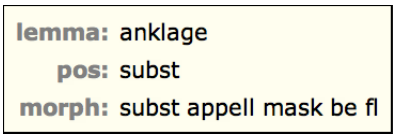

Figure 2: Morphological information associated with *anklagene* in the NDT analysis in Figure 1.

words to dependents with the functions of those dependents, with fairly coarse-grained distinctions, exemplified by 'ADV' as the label both for the sentence adverb *bare* and the selected preposition *om*. 'PUTFYLL' denotes 'prepositional complement'.

In addition to the information shown in Figure 1 morphological information is included, e.g. the information in Figure 2, associated with the word *anklagene*.

Figure 3 shows the NorGramBank LFG analysis of the same sentence. Unlike the graph in Figure 1, the analysis provides information about constituent structure with precedence and dominance relations, and makes more fine-grained distinctions among the categories, such as the fact that the preposition *om* is selected (Psel-v). Furthermore the dependency relations and syntactic functions are expressed in the f-structure, an attribute-value graph with a close affinity to dependency structures. The f-structure also contains a detailed inventory of grammatical features.

The uppercase parts of the node labels in the c-structure indicate the basic syntactic categories. The nominal subclause *at anklagene ikke er sanne* is analyzed as a CP, headed by the complementizer (C) *at*, with an S as complement. The CP projects the value of COMP in the f-structure, thus being analyzed as an argument of the main predicate 'be*om' ('pray for'). This predicate has two arguments that are mapped to the syntactic functions SUBJ and COMP, as indicated by the indices 43 and 9. Since the PP *om at anklagene ikke er sanne* is selected by the verb *be*, it is not analyzed as an adverbial adjunct, but as an argument with a preposition that does not express a predicate of its own but is rather seen as semantically incorporated in the verbal predicate. In this way the analysis captures both the syntactic independence of the preposition with respect to its governing verb (in the c-structure), and its close functional and semantic association with the same verb (in the f-structure).

The position of the finite verb is different in main clauses, since Norwegian is a verb-second language. The analysis captures this by embedding the main clause S as a complement of IP, a phrase with a finite verb as its obligatory head in second position. In the c-structure the finite verb Vfin occurs as head of IP, leaving the VPmain without a V daughter, the Vfin under I′ being its 'extended head' (Bresnan, 2001). The subject PRONP phrase occurs in the specifier position of IP where it also projects the value of TOPIC in the f-structure. In this way the syntactic differences between main and subordinate clauses are reduced to the presence vs. absence of an IP structure embedding the S.

2. Related Work

In other treebank projects 'deep' syntactic or semantically-oriented analysis is achieved either directly by parsing, as in our project, or by augmenting existing treebank analyses with such information. The LinGO Redwoods Treebank (Oepen et al., 2004), based on Head-driven Phrase Structure Grammar (HPSG), is constructed by parsing and discriminant-based disambiguation. This treebank has inspired the methodology of our project. The Alpino Dependency Treebank (van der Beek et al., 2002) is also based on parsing with an HPSG-based grammar, in this case producing dependency structures, and subsequent parse selection. The input to the parser may be partly bracketed by annotators. Augmenting existing treebanks with deeper, semantically-oriented information is the method of the Prague Dependency Treebank (Hajič, 2004), where deeper, 'tectogrammatical' information has been added to the existing more surface-oriented annotation. Similarly, (Schluter and Genabith, 2009) describes the semi-automatic addition of LFG f-structures to the French Treebank. The representations in The Sequoia French Dependency Treebank have also been supplemented with deep (or 'canonical') subcategorization frames and dependency links in addition to the existing surface frames and links (Candito et al., 2014).

3. Treebank Construction

The purpose of NorGramBank is to capture the variety of grammatical constructions that characterize the language, and their distribution in texts. Consequently, providing the best possible analyses of grammatical sentences or sentence parts takes priority over providing ad hoc analyses of ungrammatical or marginally grammatical sentences. This has consequences for the choice of texts, the efforts in preprocessing, and the degree of coverage that is possible to achieve.

3.1. Text Selection

The treebank contains both fiction and nonfiction texts, the latter including newspaper texts. In many other treebanks, including the NDT, newspaper text is the primary text type.

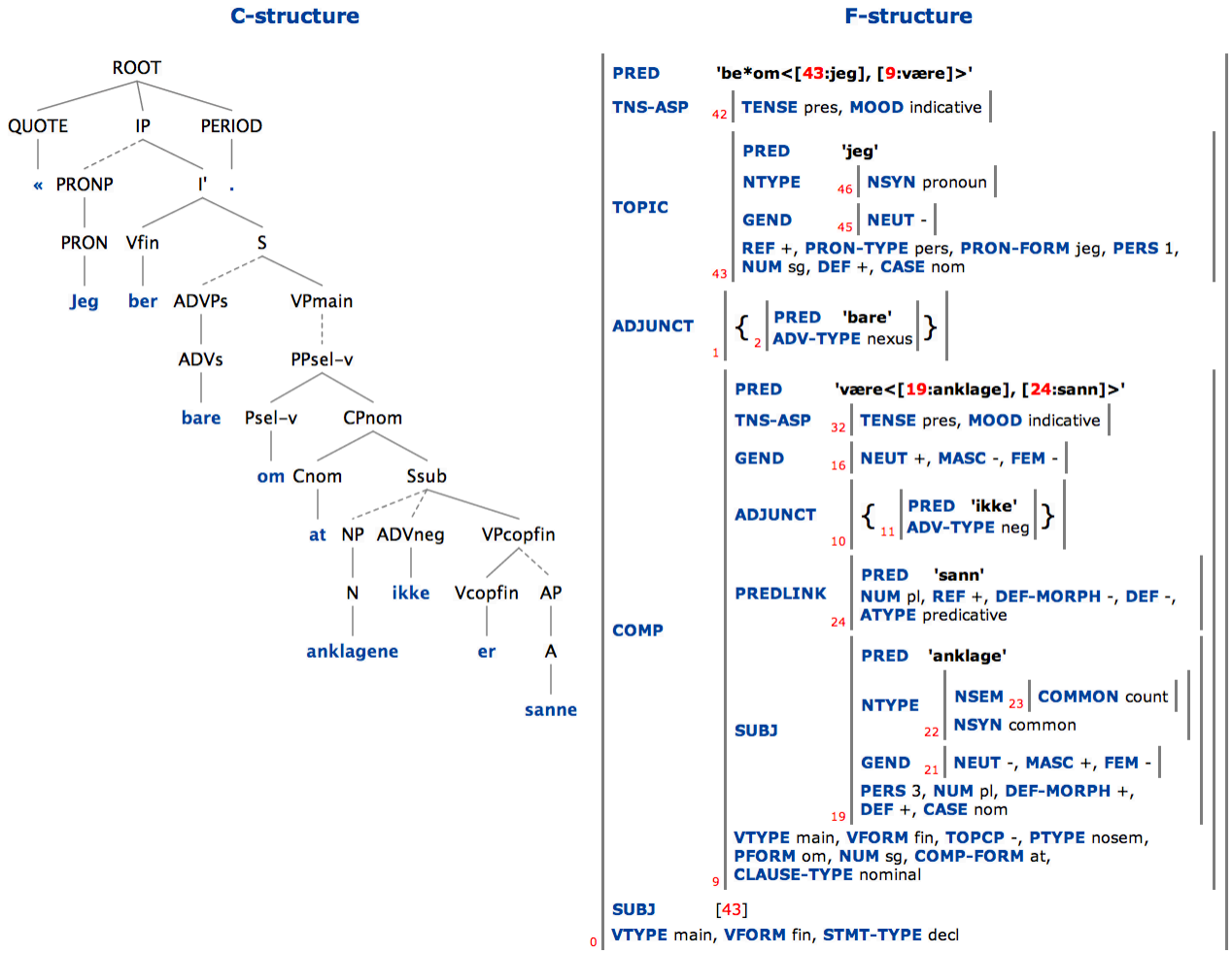

Figure 3: The NorGramBank LFG analysis of the sentence (1)

Since our approach is dependent on high-quality textual input, we also wanted to include more professionally proofread and edited text; newspaper text is usually poorly proofread and quality checked. We have therefore made fiction texts in the form of novels the most extensive text type in the treebank.

One of the main sources of fiction texts is the National Library of Norway, which has digitized large parts of its collection. Nonfiction texts in NorGramBank have been selected mainly from three large corpora of Norwegian, Leksikografisk Bokmålskorpus [Lexicographic Corpus of Bokmål], the Norwegian Newspaper Corpus, and Nynorskkorpuset ved Norsk Ordbok 2014 [the Norwegian Nynorsk Corpus].

3.2. Text Preprocessing

Since coverage in our approach is dependent on every word in a sentence being recognized, all unrecognized words must be assigned appropriate lexical and morphological information. All tokens that are not recognized by the morphological analyzer when a text is imported to the text repository are presented to annotators for preprocessing before the text is parsed (Rosén, 2014; Rosén et al., 2014). Most unrecognized words are names, productive compounds that are not captured by our compound analyzer, multiword expressions and loanwords.

3.3. Parsing with XLE

The grammar is developed on the Xerox Linguistic Environment platform (XLE) (Crouch et al., 2011), and parsing is done with the XLE chart parser, which offers some options for optimization. One option utilized in parsing with NorGram is pruning (Cahill et al., 2007b; Cahill et al., 2008). This means that phrase structure rule daughters have been weighted according to their relative frequency of occurrence in a parsed and disambiguated training corpus, and then daughters below a chosen cutoff-point are disregarded during parsing. In NorGramBank pruned parsing is used in all cases where regular parsing fails to produce a result. Pruned parsing speeds up the parsing process radically and increases coverage significantly. Of the 500 test sentences discussed in Section 5.1. below, 88, i.e., 17.6%, were analyzed by pruned parsing and got a parsing result.

NorGram also includes rules for fragment parsing, which is an option included in the XLE platform. When parsing produces no result, fragment parsing is attempted, whereby analyses are assigned to maximal contiguous chunks.

3.4. Discriminant-based and Stochastic Disambiguation

The large degree of lexical and syntactic ambiguity in natural languages leads to a potentially large number of analyses. In the Bokmål part of NorGramBank, the median number of analyses per sentence is 16, while the average number of analyses is as high as 1,258. Efficient disambiguation is achieved through the use of discriminants, which are simple properties of analyses (Carter, 1997; Oepen et al., 2004). We use discriminants designed and implemented for LFG grammars (Rosén et al., 2007). Annotators use the automatically derived discriminants to choose efficiently among the set of complete alternative analyses provided by the parser; the annotators do not build or modify any analyses themselves. Since it is not feasible to disambiguate a treebank comprising 50 M words manually, the larger part of NorGramBank is being disambiguated stochastically. The stochastic parse ranking module, which is part of the XLE distribution (Riezler and Vasserman, 2004), was trained on 34,000 manually disambiguated sentences (330,000 words). In contrast to previous work on parse ranking for LFG treebanks (Cahill et al., 2007a), we are using discriminants as features in the log-linear model of the module. In our setting, discriminants are a natural choice, and they prove to perform well, as the evaluation in Section 5.2. shows.

3.5. Work Cycle and System Design

Parsebanking is used as a method for grammar development through a work cycle in which the grammar and lexicon are being continuously improved and extended in interaction with the annotators responsible for disambiguating the parsing results (Losnegaard et al., 2012). By reparsing of the entire treebank at certain intervals compatibility with the latest grammar and lexicon versions is nevertheless ensured. For sentences that have been manually disambiguated before, the discriminant choices of the annotators have been stored and are automatically reapplied to the new set of parses, which is frequently sufficient to select a unique solution again. Our methodology is inspired by the LinGO Redwoods approach to parsebanking (Oepen et al., 2004).

The INESS treebanking platform is implemented as a web-based client-server system. The treebanking software resides on an HPC cluster consisting of computing nodes that are used for parsing with XLE and server nodes that run the underlying relational database, the indexing and querying machinery, and the middleware.

The assignment of sentences to computing nodes for parsing and indexing is mediated by the database: each XLE process independently polls the database for unprocessed sentences and fetches an arbitrary one. When a sentence has been parsed, its discriminants and index data are computed and stored back to the database together with the parse result. Each time the analysis of a sentence changes (as a result of parsing or discriminant choice) the relevant part of the search index is recalculated, thereby securing that query results always reflect the current status of the treebanks.

4. Search

4.1. Query Language and Documentation

The treebanks in the INESS platform can be queried using INESS Search (Meurer, 2012). INESS Search is based on TIGERSearch, but handles a variety of treebank formats, and, in particular, supports search in LFG f-structures, which are directed graphs. Queries may also refer to c-structure configurations and to the relationship between c- and f-structures. The query language implements full first-order predicate logic, with existential and universal quantification and negation. Existentially quantified variables are prefixed with '#', universally quantified variables are prefixed with '%'. Documentation of the query language and the interface to perform search in one or more treebanks are provided online.

We discuss a few features of the query possibilities and give examples illustrating these. The direct dominance operator is '>' (general dominance, direct or indirect, is denoted by '>*'). The expression `CPnom > Ssub` will find all sentences with c-structures in which a node labelled 'CPnom' directly dominates a node labelled 'Ssub'. The dominance operator may be labelled, which is relevant in f-structure search, where the f-structure attributes can be seen as labelling the dominance relation between f-structures. Thus, the expression `#x >(SUBJ PRED) 'jeg'` will find all sentences whose f-structures at some level contain an attribute path <SUBJ PRED> whose value is 'jeg' (cp. Figure 3), i.e., all sentences containing the word *jeg* functioning as subject.

The projection relation which holds between a c-structure node and its associated f-structure is expressed by the operator '>>'. The expression `PRON* >> #f >NUM 'sg'` will find all sentences where a c-structure node whose label begins with 'PRON' projects an f-structure #f in which the attribute NUM has the value 'sg', i.e., all sentences containing a pronoun with singular number.

4.2. Example-based Search Documentation

Even though INESS Search has a significantly simplified syntax as compared to TIGERSearch, the detailed nature of the information in the treebanks necessitates special measures to aid prospective users. We therefore provide a detailed example-based introduction to INESS Search which takes the comprehensive Norwegian reference grammar (*Norsk referansegrammatikk*, NRG (Faarlund et al., 1997)) as a starting point. This is the same idea as the one pursued by (Van der Wouden et al., 2015), who describe a project adding intelligent links to a grammatical database in the form of annotated queries to various Dutch language resources, thereby both making the query facilities more accessible and adding to the value of the grammatical database. However, *Norsk referansegrammatikk* is only available on paper, so in our case it is a question of structuring the documentation according to the chapter structure of the published grammar. Following the chapters and the phenomena discussed in NRG the documentation takes a selected example of each phenomenon as a starting point. The documentation then provides an analysis of the example, presents a search expression to find similar examples, gives a paraphrase and an explanation of the search expression, and presents a few search results. An example may

serve to illustrate this mode of documentation as well as providing further illustration of the grammatical analyses in NorGramBank and the query language.

Wh-questions with Clefting
The following example is taken from the chapter on constituent questions (wh-questions) (Faarlund et al., 1997, p. 942) .

(2) *Kva var det du såg?*
what was it you saw
'What was it you saw?'

Comment on the Analysis
The analysis of the example is given in Figure 4. Cleft sentences (focusing constructions) have the attributes GVN-TOP ('given-topic') and FOCUS in the f-structure. GVN-TOP is the subordinate clause (here: *du såg*), which expresses given (known) content in cleft sentences. FOCUS has the same value as PREDLINK, which is the predicative complement. In wh-questions with clefting it is the question-word which is focused. Hence these sentences are characterized by having the same element as FOCUS and FOCUS-INT (interrogative focus, the basic function of the question word), i.e., the two attributes have the same value in the f-structure. In the analysis we see that both FOCUS-INT, FOCUS and PREDLINK have the same value, with the index *17*. This can be exploited in searching for such sentences.

Query Expression
Find wh-questions with clefting (focusing)
`#x >(FOCUS & FOCUS-INT) #y & #x >STMT-TYPE 'int'`

Paraphrase
A node #x dominates, along both the attribute FOCUS and the attribute FOCUS-INT, a node #y, and the same node #x dominates, along the attribute STMT-TYPE, a node 'int'.

Explanation
The fact that two attributes in an f-structure have the same value can be expressed in a compact way by combining the two attributes with '&' as the label of the dominance symbol '>'. An alternative way of expressing the same is the more complex `#x >FOCUS #y & #x >FOCUS-INT #y`. Thus, this query expression finds sentences where the question word (FOCUS-INT) has been focused (FOCUS) by means of clefting. In addition, the expression `#x >STMT-TYPE 'int'` (i.e., statement-type = interrogative) restricts the search to direct (not embedded) questions.

Some Query Results
Men kva slags Gud er det, som krev eit offer for våre synder?
[But what kind of God is it that demands a sacrifice for our sins?]
Kva er det elles vi drassar på den jentungen for, sa han.
[What else is it we are dragging that girl-child along for, he said.]
Men kven er det som eig døden?
[But who is it that owns death?]

5. Results and Evaluation

5.1. Grammar Coverage

To test grammar coverage, 500 sentences were randomly selected from 37 M words in the Bokmål part of the corpus. Table 1 shows the number of sentences receiving full analyses, fragmented analyses, null analyses, correct analyses, 'minimally incorrect' analyses and wrong or null analyses, respectively, distributed over different sentence lengths. 'Correct analysis' means that the correct analysis is included among the full or fragment analyses of the sentence. 'Minimally incorrect' sentences are sentences whose best analysis by the parser departs minimally from the correct (desired) analysis, such as one wrong adverbial attachment or fragmenting due to one missing verb frame. The rest of the sentences either have wrong analyses or null analyses.
Of the 28 cases of null analyses, 10 are caused by shortcomings of the grammar, while 8 exceeded our preset limits on time and space resources, 5 have a higher number of analyses than our limit of 20,000, and 5 contain names or symbols which we had failed to register in the lexicon.
There are 11 correctly fragmented analyses in the fully correct category. These are cases where the sentence is deemed actually to have a fragmented structure, and where the analysis identifies the appropriate fragments. An example of such a sentence is: *Jeg vil også gjøre det helt klart for deg at jeg ikke, jeg gjentar, ikke myrdet Celia*. 'I also want to make it quite clear to you that I didn't, I repeat, didn't murder Celia.'
With 78.4% gold and 6.8% minimally incorrect analyses we may conclude that useful analyses are included among the analyses of 85.2% of the 500 test sentences.

5.2. Annotator and Disambiguator Precision

A study of the agreement of annotators and stochastic disambiguation with a gold standard produced by the grammar developer was carried out based on 500 sentences with complete analyses randomly selected from 37 M words in the Bokmål part of the corpus.[2] This set of sentences is distinct from the set of 500 sentences discussed in the previous section (5.1.). There were four annotators a1, a2, a3 and a4, while A refers to the stochastic (automatic) disambiguation. The agreement between an annotator and the gold standard was measured based on the set of discriminants (see Section 3.4.) derived from the set of alternative analyses, using the Kappa measure of agreement (Cohen, 1960). This is inspired by a similar approach in connection with an HPSG-based treebank (de Castro, 2011). In our case the base for the calculation comprised lexical, c-structure and f-structure discriminants. For each alternative analysis a certain subset of discriminants will be valid and the rest invalid. The measure finds the proportion of discriminants which the annotator and the gold standard agree on classifying as either valid or invalid, and quantifies the extent to which this proportion exceeds the result of random choice. The output is a number between 0 and 1, where values above 0.8 have top rank and are variously called 'perfect', 'good' or 'high' in different proposed scales for the interpretation of Kappa results. The

[2] We have previously carried out a pilot study of interannotator agreement (Dyvik et al., 2013).

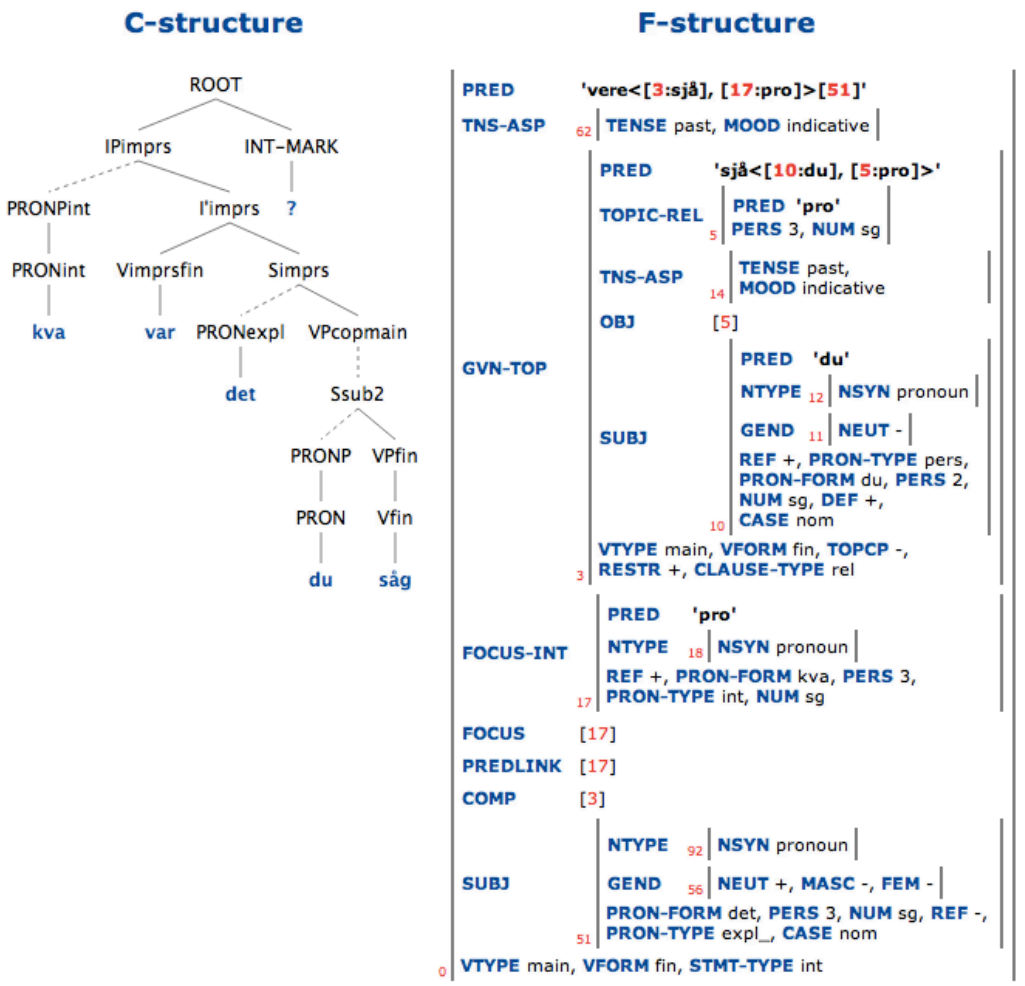

Figure 4: Documentation example: wh-question with clefting

Sentence lengths		1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	All
Sentences	#	142	145	113	61	24	9	5	1	500
Solutions, average	#	7	49	270	390	1030	745	1384	324	185
Solutions, median	#	4	14	80	96	512	592	1080	324	16
Full analyses	#	127	134	94	47	15	6	1	1	425
	%	89.4	92.4	83.2	77.0	62.5	66.7	20.0	100.0	85.0
Fragment analyses	#	10	7	16	8	4	0	2	0	47
	%	7.0	4.8	14.2	13.1	16.7	0.0	40.0	0.0	9.4
Null analyses	#	5	4	3	6	5	3	2	0	28
	%	3.5	2.8	2.7	9.8	20.8	33.3	40.0	0	5.6
Fully correct	#	125	127	87	41	7	4	0	1	392
	%	88.0	87.6	77.0	67.2	29.2	44.4	0.0	100.0	78.4
Minimally incorrect	#	7	0	12	9	5	0	1	0	34
	%	4.9	0.0	10.6	14.8	20.8	0.0	20.0	0.0	6.8
Wrong or null	#	10	18	14	11	12	5	4	0	74
	%	7.0	12.4	12.4	18.0	50.0	55.6	80.0	0.0	14.8

Table 1: Grammar coverage of 500 random sentences.

total Kappa result for a given annotator is the average of the results for each sentence, where the sentences taken into account are those for which both annotator and gold standard have disambiguated down to one single solution (between 438 and 453 sentences for the annotators and 466 for A). Table 2 shows the Kappa values for the four annotators a1–a4 and the stochastic disambiguation A.

Annotator	*Kappa agreement*
a1	0.77
a2	0.87
a3	0.86
a4	0.83
A	0.65

Table 2: Kappa agreement between annotators a1–a4 + automatic disambiguation A and gold standard

5.3. Types of Annotator Disagreement

Our previous study of interannotator agreement (Dyvik et al., 2013) contained an overview of types of discrepancies among the annotators' choices. The ranking of the types according to frequency agree fairly well with the present study, which covers more text and more types. In both studies the two most frequent types of discrepancies concern adjunct attachment and the ambiguity arising from the forms *det*, *den* and *de*, which can all be referring pronouns or determiners; in addition, *det* can be an expletive pronoun. (A fourth possibility, article as a distinct kind of determiner, was eliminated after the previous study.) In the present study adjunct attachment covers about 23% of the cases, while the *det/den/de* type covers about 13%. Other frequent types are adverb type (about 12%), selected vs. semantic prepositions (about 6%), and the choice between root and epistemic readings of modal verbs (about 5%), which sometimes has syntactic consequences.

Clearly, the more fine-grained the distinctions expressed in the annotation, the higher the risk of discrepancies among the annotators, and NorGram makes rather fine-grained distinctions. Thus, the grammar distinguishes more that 20 adverb types based on syntactic distribution, which often gives rise to alternative analyses, as in Example 3.

(3) *Sorry så bare på skjermen.*
Sorry looked only on the screen
'Sorry only looked/looked only at the screen.'

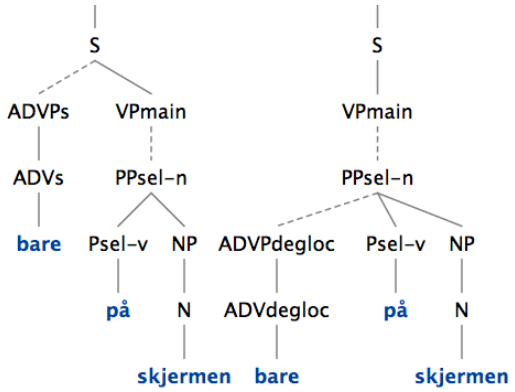

Figure 5: Ambiguity between two adverb types (see Example (3))

This gave the annotators the choice shown in Figure 5, where they chose differently. The sentence adverb (ADVs) choice means that the verb is modified (Sorry just looked), while the degree adverb (ADVdegloc, modifying locatives) means that the PP is modified (only at the screen). Only careful inspection of the textual context could resolve the ambiguity.

Studies of interannotator disagreement can help grammar developers determine the borderline where further distinctions cease to be useful. For treebanks like ours, where sentence-bounded stochastic disambiguation is responsible for about 99% of the sentences, distinctions which rely completely on an understanding of the preceding textual context will hardly be useful even if human annotators can handle them.

It is typical of many grammatical distinctions that while they are clear and easy to make on a structural basis in some cases, they may appear vague and indeterminate in others, not even implying clearly distinct readings. A relevant consideration in connection with the maintenance of a given grammatical distinction, therefore, is how frequent the vague cases are in comparison with the clear ones.

The present study points out some distinctions as candidates for reconsideration. One is the distinction between *det* as a pronoun (= 'it') and and as a demonstrative (= 'that'), as in Example 4.

(4) *Nei, det er lenge siden, svarte Maia.*
no that/it is long since answered Maia
'No, that/it is long ago, Maia answered.'

Even in context it may be unclear whether such examples should be translated with *it* or *that*.

6. Conclusion and Outlook

Our aim has been to create a treebank for Norwegian which captures those syntactic properties of texts that can be described by general grammatical rules in conjunction with a comprehensive and fine-grained lexicon, within a grammatical framework which will allow reference to 'deep' properties such as, e.g., the predicate-argument relations underlying varying syntactic expressions. We have presented the first substantial treebank for Norwegian with detailed syntactic annotation which is fully compatible with a linguistically principled, large coverage grammar. The treebank is being made available to researchers on a highly functional, fully web-based platform and will also be downloadable subject to possible text-specific licence restrictions. Enhancement and scaling up remain possible through reparsing with an improved grammar and the addition of more automatically parsed and disambiguated text. The treebank will continue to grow through the addition and analysis of new texts until the end of the INESS project in 2017.

7. References

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell, Malden, MA.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The Parallel Grammar project. In Carroll, J., Oostdijk, N., and Sutcliffe, R., editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation*

at the 19th International Conference on Computational Linguistics (COLING), pages 1–7, Taipei, Taiwan. Association for Computational Linguistics.

Cahill, A., Forst, M., and Rohrer, C. (2007a). Designing features for parse disambiguation and realization ranking. In King, T. H. and Butt, M., editors, *The Proceedings of the LFG '07 Conference*, pages 128–147. CSLI Publications, Stanford.

Cahill, A., King, T. H., and Maxwell, J. T. (2007b). Pruning the search space of a hand-crafted parsing system with a probabilistic parser. In *ACL2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic.

Cahill, A., Maxwell, J. T., Meurer, P., Rohrer, C., and Rosén, V. (2008). Speeding up LFG parsing using C-structure pruning. In *Coling 2008: Proceedings of the Workshop on Grammar Engineering Across Frameworks*, Manchester UK.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., and Villemonte de La Clergerie, E. (2014). Deep syntax annotation of the Sequoia French Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavík, Iceland.

Carter, D. (1997). The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, Rhode Island.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE Documentation. http://ling.uni-konstanz.de/pages/xle/doc/xle_toc.html/. [Online; accessed 23-October-2015].

de Castro, S. R. (2011). Developing reliability metrics and validation tools for datasets with deep linguistic information. http://hdl.handle.net/10451/8688. [Online; accessed 24-October-2015].

Dyvik, H., Thunes, M., Haugereid, P., Rosén, V., Meurer, P., De Smedt, K., and Losnegaard, G. S. (2013). Studying interannotator agreement in discriminant-based parsebanking. In Kübler, S., Osenova, P., and Volk, M., editors, *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 37–48. Bulgarian Academy of Sciences.

Faarlund, J. T., Lie, S., and Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Universitetsforlaget, Oslo, Norway.

Hajič, J. (2004). Complex corpus annotation: The Prague Dependency Treebank. Technical report, Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, Slovakia.

Losnegaard, G. S., Lyse, G. I., Thunes, M., Rosén, V., De Smedt, K., Dyvik, H., and Meurer, P. (2012). What we have learned from Sofie: Extending lexical and grammatical coverage in an LFG parsebank. In Hajič, J., De Smedt, K., Tadić, M., and Branco, A., editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 69–76, Istanbul, Turkey.

Meurer, P. (2012). INESS-Search: A search system for LFG (and other) treebanks. In Butt, M. and King, T. H., editors, *Proceedings of the LFG '12 Conference*, LFG Online Proceedings, pages 404–421, Stanford, CA. CSLI Publications.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). LinGO Redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 2(4):575–596, December.

Riezler, S. and Vasserman, A. (2004). Gradient feature testing and l_1 regularization for maximum entropy parsing. In *Proceedings of EMNLP'04*, Barcelona, Spain.

Rosén, V., Meurer, P., and De Smedt, K. (2007). Designing and implementing discriminants for LFG grammars. In King, T. H. and Butt, M., editors, *The Proceedings of the LFG '07 Conference*, pages 397–417. CSLI Publications, Stanford.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012). An open infrastructure for advanced treebanking. In Hajič, J., De Smedt, K., Tadić, M., and Branco, A., editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Rosén, V., Haugereid, P., Thunes, M., Losnegaard, G. S., Dyvik, H., and Meurer, P. (2014). The interplay between lexical and syntactic resources in incremental parsebanking. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1617–1624, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Rosén, V. (2014). Språkteknologiens behov for leksikalsk informasjon. In Fjeld, R. V. and Hovdenak, M., editors, *Rapport fra Konferanse om leksikografi i Norden, Oslo, 13.-16. august 2013*, volume 12 of *Nordiske studier i leksikografi*. Novus forlag.

Schluter, N. and Genabith, J. v. (2009). Dependency parsing resources for French: converting acquired Lexical Functional Grammar f-structure annotations and parsing f-structures directly. In *Proceedings of NODALIDA 2009*, Odense, Denmark.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. B. (2014). The Norwegian Dependency Treebank. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

van der Beek, L., Bouma, G., Malouf, R., and van Noord, G. (2002). The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

Van der Wouden, T., Bouma, G., Van de Kamp, M., Van Koppen, M., Landsbergen, F., and Odijk, J. (2015). Enriching a grammatical database with intelligent links to linguistic resources. In *CLARIN Annual Conference 2015 (Book of Abstracts)*, pages 89–92.