Is it Really that Difficult to Parse German?

Sandra Kübler, Erhard W. Hinrichs, Wolfgang Maier
SfS-CL, SFB 441, University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany
{kuebler,eh,wmaier}@sfs.uni-tuebingen.de

Abstract

This paper presents a comparative study of probabilistic treebank parsing of German, using the Negra and TüBa-D/Z treebanks. Experiments with the Stanford parser, which uses a factored PCFG and dependency model, show that, contrary to previous claims for other parsers, lexicalization of PCFG models boosts parsing performance for both treebanks. The experiments also show that there is a big difference in parsing performance, when trained on the Negra and on the TüBa-D/Z treebanks. Parser performance for the models trained on TüBa-D/Z are comparable to parsing results for English with the Stanford parser, when trained on the Penn treebank. This comparison at least suggests that German is not harder to parse than its West-Germanic neighbor language English.

1 Introduction

There have been a number of recent studies on probabilistic treebank parsing of German (Dubey, 2005; Dubey and Keller, 2003; Schiehlen, 2004; Schulte im Walde, 2003), using the Negra treebank (Skut et al., 1997) as their underlying data source. A common theme that has emerged from this research is the claim that lexicalization of PCFGs, which has been proven highly beneficial for other languages[1], is detrimental for parsing accuracy of German. In fact, this assumption is by now so widely held that Schiehlen (2004) does not even consider lexicalization as a possible parameter and concentrates instead only on treebank transformations of various sorts in his experiments.

Another striking feature of all studies mentioned above are the relatively low parsing F-scores achieved for German by comparison to the scores reported for English, its West-Germanic neighbor, using similar parsers. This naturally raises the question whether German is just harder to parse or whether it is just hard to parse the Negra treebank.[2]

The purpose of this paper is to address precisely this question by training the Stanford parser (Klein and Manning, 2003b) and the LoPar parser (Schmid, 2000) on the two major treebanks available for German, Negra and TüBa-D/Z, the Tübingen treebank of written German (Telljohann et al., 2005). A series of comparative parsing experiments that utilize different parameter settings of the parsers is conducted, including lexicalization and markovization. These experiments show striking differences in performance between the two treebanks. What makes this comparison interesting is that the treebanks are of comparable size and are both based on a newspaper corpus. However, both treebanks differ significantly in their syntactic annotation scheme. Note, however, that our experiments concentrate on the original (context-free) annotations of the treebank.

The structure of this paper is as follows: section 2 discusses three characteristic grammatical features of German that need to be taken into account in syntactic annotation and in choosing an appropriate parsing model for German. Section 3 introduces the Negra and TüBa-D/Z treebanks and

[1] For English, see Collins (1999).

[2] German is not the first language for which this question has been raised. See Levy and Manning (2003) for a similar discussion of Chinese and the Penn Chinese Treebank.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 111–119,
Sydney, July 2006.

discusses the main differences between their annotation schemes. Section 4 explains the experimental setup, sections 5-7 the experiments, and section 8 discusses the results.

2 Grammatical Features of German

There are three distinctive grammatical features that make syntactic annotation and parsing of German particularly challenging: its placement of the finite verb, its flexible phrasal ordering, and the presence of discontinuous constituents. These features will be discussed in the following subsections.

2.1 Finite Verb Placement

In German, the placement of finite verbs depends on the clause type. In non-embedded assertion clauses, the finite verb occupies the second position in the clause, as in (1a). In yes/no questions, as in (1b), the finite verb appears clause-initially, whereas in embedded clauses it appears clause finally, as in (1c).

(1) a. Peter wird das Buch gelesen haben.
Peter will the book read have
'Peter will have read the book.'

b. Wird Peter das Buch gelesen haben?
Will Peter the book have read
'Will Peter have read the book?'

c. dass Peter das Buch gelesen haben wird.
that Peter the book read have will
'... that Peter will have read the book.'

Regardless of the particular clause type, any cluster of non-finite verbs, such as *gelesen haben* in (1a) and (1b) or *gelesen haben wird* in (1c), appears at the right periphery of the clause.

The discontinuous positioning of the verbal elements in verb-first and verb-second clauses is the traditional reason for structuring German clauses into so-called *topological fields* (Drach, 1937; Erdmann, 1886; Höhle, 1986). The positions of the verbal elements form the *Satzklammer* (sentence bracket) which divides the sentence into a *Vorfeld* (initial field), a *Mittelfeld* (middle field), and a *Nachfeld* (final field). The Vorfeld and the Mittelfeld are divided by the *linke Satzklammer* (left sentence bracket), which is realized by the finite verb or (in verb-final clauses) by a complementizer field. The *rechte Satzklammer* (right sentence bracket) is realized by the verb complex and consists of verbal particles or sequences of verbs. This right sentence bracket is positioned between the Mittelfeld and the Nachfeld. Thus, the theory of topological fields states the fundamental regularities of German word order.

The topological field structures in (2) for the examples in (1) illustrate the assignment of topological fields for different clause types.

(2) a. $[_{VF}$ $[_{NP}$ Peter]] $[_{LK}$ wird] $[_{MF}$ $[_{NP}$ das Buch]] $[_{RK}$ $[_{VC}$ gelesen haben.]]

b. $[_{LK}$ Wird] $[_{MF}$ $[_{NP}$ Peter] $[_{NP}$ das Buch]] $[_{RK}$ $[_{VC}$ gelesen haben?]]

c. $[_{LK}$ $[_{CF}$ dass]] $[_{MF}$ $[_{NP}$ Peter] $[_{NP}$ das Buch]] $[_{RK}$ $[_{VC}$ gelesen haben wird.]]

(2a) and (2b) are made up of the following fields: LK (for: linke Satzklammer) is occupied by the finite verb. MF (for: Mittelfeld) contains adjuncts and complements of the main verb. RK (for: rechte Satzklammer) is realized by the verbal complex (VC). Additionally, (2a) realizes the topological field VF (for: Vorfeld), which contains the sentence-initial constituent. The left sentence bracket (LK) in (2c) is realized by a complementizer field (CF) and the right sentence bracket (RK) by a verbal complex (VC) that contains the finite verb *wird*.

2.2 Flexible Phrase Ordering

The second noteworthy grammatical feature of German concerns its flexible phrase ordering. In (3), any of the three complements and adjuncts of the main verb *(ge)lesen* can appear sentence-initially.

(3) a. Der Mann hat gestern den Roman
The man has yesterday the novel
gelesen.
read
'The man read the novel yesterday.'

b. Gestern hat der Mann den Roman gelesen

c. Den Roman hat der Mann gestern gelesen

In addition, the ordering of the elements that occur in the Mittelfeld is also free so that there are two possible linearizations for each of the examples in (3a) - (3b), yielding a total of six distinct orderings for the three complements and adjuncts.

Due to this flexible phrase ordering, the grammatical functions of constituents in German, unlike for English, cannot be deduced from the constituents' location in the tree. As a consequence, parsing approaches to German need to be based on treebank data which contain a combination of constituent structure and grammatical functions – for parsing and evaluation.

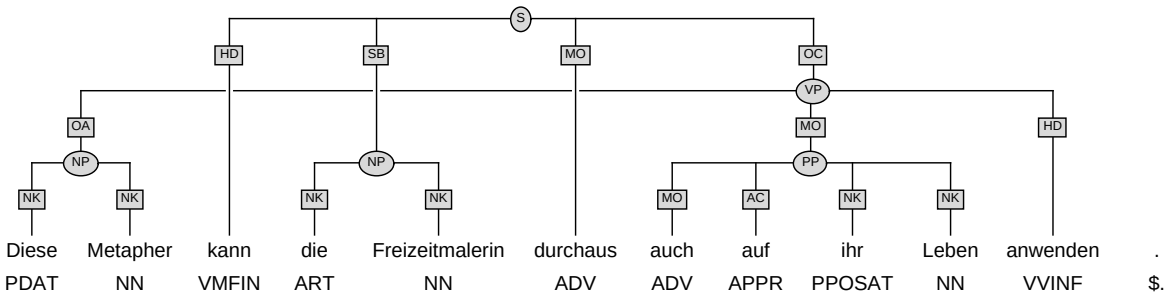

Figure 1: A sample tree from Negra.

2.3 Discontinuous Constituents

A third characteristic feature of German syntax that is a challenge for syntactic annotation and for parsing is the treatment of discontinuous constituents.

(4) Der Mann hat gestern den Roman gelesen,
The man has yesterday the novel read
den ihm Peter empfahl.
which him Peter recommended
'Yesterday the man read the novel which Peter recommended to him.'

(5) Peter soll dem Mann empfohlen haben, den
Peter is to the man recommended have the
Roman zu lesen.
novel to read
'Peter is said to have recommended to the man to read the novel.'

(4) shows an extraposed relative clause which is separated from its head noun *den Roman* by the non-finite verb *gelesen*. (5) is an example of an extraposed non-finite VP complement that forms a discontinuous constituent with its governing verb *empfohlen* because of the intervening non-finite auxiliary *haben*. Such discontinuous structures occur frequently in both treebanks and are handled differently in the two annotation schemes, as will be discussed in more detail in the next section.

3 The Negra and the TüBa-D/Z Treebanks

Both treebanks use German newspapers as their data source: the Frankfurter Rundschau newspaper for Negra and the 'die tageszeitung' (taz) newspaper for TüBa-D/Z. Negra comprises 20 000 sentences, TüBa-D/Z 15 000 sentences. There is evidence that the complexity of sentences in both treebanks is comparable: sentence length as well as the percentage of clause nodes per sentence is comparable. In Negra, a sentence is 17.2 words long, in Tüba-D/Z, 17.5 words. Negra has an average of 1.4 clause nodes per sentence, TüBa-D/Z 1.5 clause nodes.

Both treebanks use an annotation framework that is based on phrase structure grammar and that is enhanced by a level of predicate-argument structure. Annotation for both was performed semi-automatically. Despite all these similarities, the treebank annotations differ in four important aspects: 1) Negra does not allow unary branching whereas TüBa-D/Z does; 2) in Negra, phrases receive a flat annotation whereas TüBa-D/Z uses phrase internal structure; 3) Negra uses crossing branches to represent long-distance relationships whereas TüBa-D/Z uses a pure tree structure combined with functional labels to encode this information; 4) Negra encodes grammatical functions in a combination of structural and functional labeling whereas TüBa-D/Z uses a combination of topological fields functional labels, which results in a flatter structure on the clausal level. The two treebanks also use different notions of grammatical functions: TüBa-D/Z defines 36 grammatical functions covering head and non-head information, as well as subcategorization for complements and modifiers. Negra utilizes 48 grammatical functions. Apart from commonly accepted grammatical functions, such as *SB* (subject) or *OA* (accusative object), Negra grammatical functions comprise a more extended notion, e.g. *RE* (repeated element) or *RC* (relative clause).

(6) Diese Metapher kann die Freizeitmalerin
This metaphor can the amateur painter
durchaus auch auf ihr Leben anwenden.
by all means also to her life apply.
'The amateur painter can by all means apply this metaphor also to her life.'

Figure 1 shows a typical tree from the Negra treebank for sentence (6). The syntactic categories are shown in circular nodes, the grammatical functions as edge labels in square boxes. A major

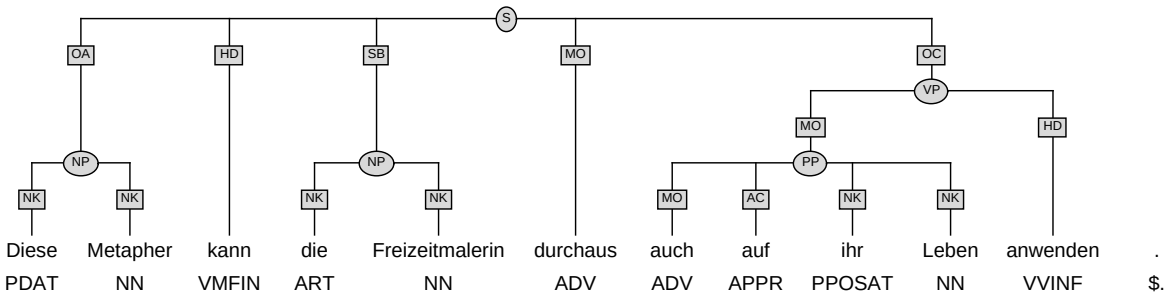

Figure 2: A Negra tree with resolved crossing branches.

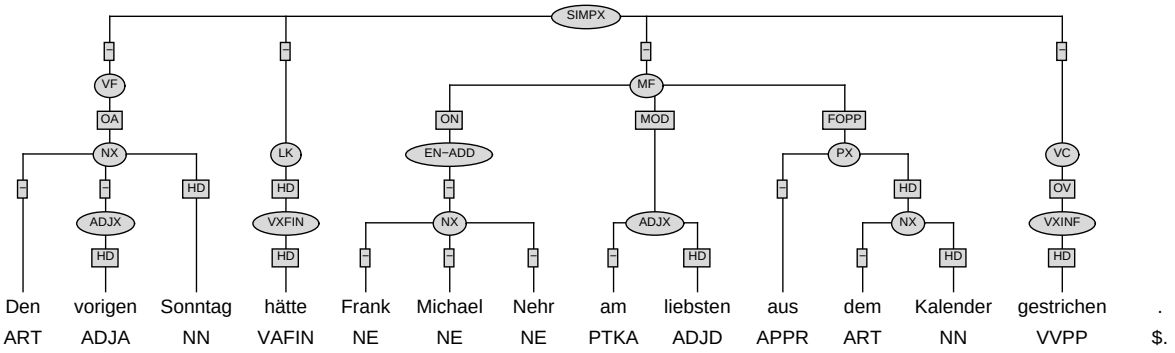

Figure 3: A sample tree from Tüba-D/Z.

phrasal category that serves to structure the sentence as a whole is the verb phrase (VP). It contains non-finite verbs (here: *anwenden*) together with their complements (here: the accusative object *Diese Metapher*) and adjuncts (here: the adverb *durchaus* and the PP modifier *auch auf ihr Leben*). The subject NP (here: *die Freizeitmalerin*) stands outside the VP and, depending on its linear position, leads to crossing branches with the VP. This happens in all cases where the subject follows the finite verb as in Figure 1. Notice also that the PP is completely flat and does not contain an internal NP.

Another phenomenon that leads to the introduction of crossing branches in the Negra treebank are discontinuous constituents of the kind illustrated in section 2.3. Extraposed relative clauses, as in (4), are analyzed in such a way that the relative clause constituent is a sister of its head noun in the Negra tree and crosses the branch that dominates the intervening non-finite verb *gelesen*.

The crossing branches in the Negra treebank cannot be processed by most probabilistic parsing models since such parsers all presuppose a strictly context-free tree structure. Therefore the Negra trees must be transformed into proper trees prior to training such parsers. The standard approach for this transformation is to re-attach crossing non-head constituents as sisters of the lowest mother node that dominates all constituents in question in the original Negra tree.

Figure 2 shows the result of this transformation of the tree in Figure 1. Here, the fronted accusative object *Diese Metapher* is reattached on the clause level. Crossing branches do not only arise with respect to the subject at the sentence level but also in cases of extraposition and fronting of partial constituents. As a result, approximately 30% of all Negra trees contain at least one crossing branch. Thus, tree transformations have a major impact on the type of constituent structures that are used for training probabilistic parsing models. Previous work, such as Dubey (2005), Dubey and Keller (2003), and Schiehlen (2004), uses the version of Negra in which the standard approach to resolving crossing branches has been applied.

(7) Den vorigen Sonntag hätte Frank Michael
The previous Sunday would have Frank Michael
Nehr am liebsten aus dem Kalender gestrichen.
Nehr preferably from the calendar deleted.
'Frank Michael Nehr would rather have deleted the previous Sunday from the calendar.'

Figure 3 shows the TüBa-D/Z annotation for sentence (7), a sentence with almost identical phrasal ordering to sentence (6). Crossing branches are avoided by the introduction of topo-

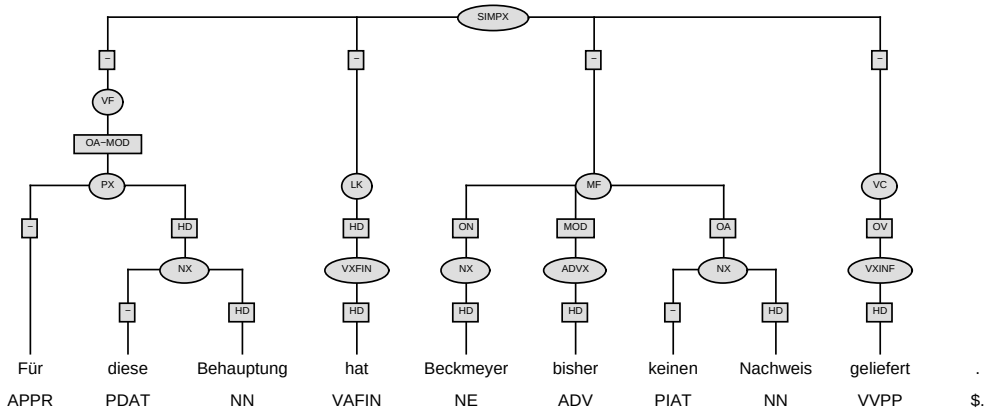

Figure 4: TüBa-D/Z annotation without crossing branches.

logical structures (here: VF, MF and VC) into the tree. Notice also that compared to the Negra annotation, TüBa-D/Z introduces more internal structure into NPs and PPs.

(8) Für diese Behauptung hat Beckmeyer bisher
For this claim has Beckmeyer yet
keinen Nachweis geliefert.
no evidence provided.
'For this claim, Beckmeyer has not provided evidence yet.'

In TüBa-D/Z, long-distance relationships are represented by a pure tree structure and specific functional labels. Figure 4 shows the TüBa-D/Z annotation for sentence (8). In this sentence, the prepositional phrase *Für diese Behauptung* is fronted. Its functional label (*OA-MOD*) provides the information that it modifies the accusative object (*OA*) *keinen Nachweis*.

4 Experimental Setup

The main goals behind our experiments were twofold: (1) to re-investigate the claim that lexicalization is detrimental for treebank parsing of German, and (2) to compare the parsing results for the two German treebanks.

To investigate the first issue, the Stanford Parser (Klein and Manning, 2003b), a state-of-the-art probabilistic parser, was trained with both lexicalized and unlexicalized versions of the two treebanks (Experiment I). For lexicalized parsing, the Stanford Parser provides a factored probabilistic model that combines a PCFG model with a dependency model.

For the comparison between the two treebanks, two types of experiments were performed: a purely constituent-based comparison using both the Stanford parser and the pure PCFG parser LoPar (Schmid, 2000) (Experiment II), and an in-depth evaluation of the three major grammatical functions *subject*, *accusative object*, and *dative object*, using the Stanford parser (Experiment III).

All three experiments use gold POS tags extracted from the treebanks as parser input. All parsing results shown below are averaged over a ten-fold cross-validation of the test data. Experiments I and II used versions of the treebanks that excluded grammatical information, thus only contained constituent labeling. For Experiment III, all syntactic labels were extended by their grammatical function (e.g NX-ON for a subject NP in TüBa-D/Z or NP-SB for a Negra subject). Experiments I and II included all sentences of a maximal length of 40 words. Due to memory limitations (7 GB), Experiment III had to be restricted to sentences of a maximal length of 35 words.

5 Experiment I: Lexicalization

Experiment I investigates the effect of lexicalization on parser performance for the Stanford Parser. The results, summarized in Table 1, show that lexicalization improves parser performance for both the Negra and the TüBa-D/Z treebank in comparison to unlexicalized counterpart models: for labeled bracketing, an F-score improvement from 86.48 to 88.88 for TüBa-D/Z and an improvement from 66.92 to 67.13 for Negra. This directly contradicts the findings reported by Dubey and Keller (2003) that lexicalization has a negative effect on probabilistic parsing models for German. We therefore conclude that these previous claims, while valid for particular configurations of

		Negra			TüBa-D/Z		
		precision	recall	F-score	precision	recall	F-score
Stanford PCFG	unlabeled	71.24	72.68	71.95	93.07	89.41	91.20
	labeled	66.26	67.59	66.92	88.25	84.78	86.48
Stanford lexicalized	unlabeled	71.31	73.12	72.20	91.60	91.21	91.36
	labeled	66.30	67.99	67.13	89.12	88.65	88.88

Table 1: The results of lexicalizing German.

		Negra			TüBa-D/Z		
		precision	recall	F-score	precision	recall	F-score
LoPar	unlabeled	70.84	72.51	71.67	92.62	88.58	90.56
	labeled	65.86	67.41	66.62	87.39	83.57	85.44
Stanford	unlabeled	71.24	72.68	71.95	93.07	89.41	91.20
	labeled	66.26	67.59	66.92	88.25	84.78	86.48
Stanford + markov	unlabeled	74.13	74.12	74.12	92.28	90.90	91.58
	labeled	69.96	69.95	69.95	89.86	88.51	89.18

Table 2: A comparison of unlexicalized parsing of Negra and TüBa-D/Z.

parsers and parameters, should not be generalized to claims about probabilistic parsing of German in general.

Experiment I also shows considerable differences in the overall scores between the two treebanks, with the F-scores for TüBa-D/Z parsing approximating scores reported for English, but with Negra scores lagging behind by an average margin of appr. 20 points. Of course, it is important to note that such direct comparisons with English are hardly possible due to different annotation schemes, different underlying text corpora, etc. Nevertheless, the striking difference in parser performance between the two German treebanks warrants further attention. Experiments II and III will investigate this matter in more depth.

6 Experiment II: Different Parsers

The purpose of Experiment II is to rule out the possibility that the differences in parser performance for the two German treebanks produced by Experiment I may just be due to using a particular parser – in this particular case the hybrid PCFG and dependency model of the Stanford parser. After all, Experiment I also yielded different results concerning the received wisdom about the utility of lexicalization from previously reported results. In order to obtain a broader experimental base, unlexicalized models of the Stanford parser and the pure PCFG parser LoPar were trained on both treebanks. In addition we experimented with two different parameter settings of the Stanford parser, one with and one without markovization. The experiment with markovization used parent information (v=1) and a second order Markov model for horizontal markovization (h=2). The results, summarized in Table 2, show that parsing results for all unlexicalized experiments show roughly the same 20 point difference in F-score that were obtained for the lexicalized models in Experiment I. We can therefore conclude that the difference in parsing performance is robust across two parsers with different parameter settings, such as lexicalization and markovization.

Experiment II also confirms the finding of Klein and Manning (2003a) and of Schiehlen (2004) that horizontal and vertical markovization has a positive effect on parser performance. Notice also that markovization with unlexicalized grammars yields almost the same improvement as lexicalization does in Experiment I.

7 Experiment III: Grammatical Functions

In Experiments I and II, only constituent structure was evaluated, which is highly annotation dependent. It could simply be the case that the TüBa-D/Z annotation scheme contains many local structures that can be easily parsed by a PCFG model or the hybrid Stanford model. Moreover, such easy to parse structures may not be of great importance when it comes to determining the correct macrostructure of a sentence. To empirically verify such a conjecture, a separate evaluation of

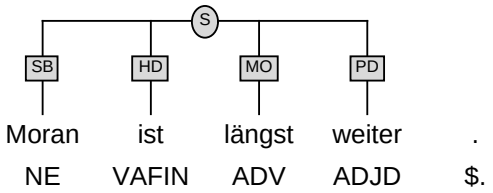

Figure 5: Negra annotation without unary nodes.

	Negra			TüBa-D/Z		
	lab. prec.	lab. rec.	lab. F-score	lab. prec.	lab. rec.	lab. F-score
without gramm. functions	69.96	69.95	69.95	89.86	88.51	89.18
all gramm. functions	47.20	56.43	51.41	75.73	74.93	75.33
subjects	52.50	58.02	55.12	66.82	75.93	71.08
accusative objects	35.14	36.30	35.71	43.84	47.31	45.50
dative objects	8.38	3.58	5.00	24.46	9.96	14.07

Table 3: A comparison of unlexicalized, markovized parsing of constituent structure and grammatical functions in Negra and TüBa-D/Z.

parser performance for different constituent types would be necessary. However, even such an evaluation would only be meaningful if the annotation schemes agree on the defining characteristics of such constituent types. Unfortunately, this is not the case for the two treebanks under consideration. Even for arguably theory-neutral constituents such as NPs, the two treebanks differ considerably. In the Negra annotation scheme, single word NPs directly project from the POS level to the clausal level, while in TüBa-D/Z, they project by a unary rule first to an NP. An extreme case of this Negra annotation is shown in Figure 5 for sentence (9). Here, all the phrases are one word phrases and are thus projected directly to the clause level.

(9) Moran ist längst weiter.
Moran is already further
'Moran is already one step ahead.'

There is an even more important motivation for not focusing on the standard constituent-based parseval measures – at least when parsing German. As discussed earlier in section 2.2, obtaining the correct constituent structure for a German sentence will often not be sufficient for determining its intended meaning. Due to the word order freeness of phrases, a given NP in any one position may in principle fulfill different grammatical functions in the sentence as a whole. Therefore grammatical functions need to be explicitly marked in the treebank and correctly assigned during parsing. Since both treebanks encode grammatical functions, this information is available for parsing and can ultimately lead to a more meaningful comparison of the two treebanks when used for parsing.

The purpose of Experiment III is to investigate parser performance on the treebanks when grammatical functions are included in the trees. For these experiments, the unlexicalized, markovized PCFG version of the Stanford parser was used, with markovization parameters v=1 and h=2, as in Experiment II. The results of this experiment are shown in Table 3. The comparison of the experiments with (line 2) and without grammatical functions (line 1) confirms the findings of Dubey and Keller (2003) that the task of assigning correct grammatical functions is harder than mere constituent-based parsing. When evaluating on all grammatical functions, the results for Negra decrease from 69.95 to 51.41, and for TüBa-D/Z from 89.18 to 75.33. Notice however, that the relative differences between Negra and TüBa-D/Z that were true for Experiments I and II remain more or less constant for this experiment as well.

In order to get a clearer picture of the quality of the parser output for each treebank, it is important to consider individual grammatical functions. As discussed in section 3, the overall inventory of grammatical functions is different for the two treebanks. We therefore evaluated those grammatical functions separately that are crucial for determining function-argument structure and

that are at the same time the most comparable for the two treebanks. These are the functions of subject (encoded as *SB* in Negra and as *ON* in TüBa-D/Z), accusative object (*OA*), and dative object (*DA* in Negra and *OD* in TüBa-D/Z). Once again, the results are consistently better for TüBa-D/Z (cf. lines 3-5 in Table 3), with subjects yielding the highest results (71.08 vs. 55.12 F-score) and dative objects the lowest results (14.07 vs. 5.00). The latter results must be attributed to data sparseness, dative object occur only appr. 1 000 times in each treebank while subjects occur more than 15 000 times.

8 Discussion

The experiments presented in sections 5-7 show that there is a difference in results of appr. 20% between Negra and TüBa-D/Z. This difference is consistent throughout, i.e. with different parsers, under lexicalization and markovization. These results lead to the conjecture that the reasons for these differences must be sought in the differences in the annotation schemes of the two treebanks.

In section 3, we showed that one of the major differences in annotation is the treatment of discontinuous constituents. In Negra, such constituents are annotated via crossing branches, which have to be resolved before parsing. In such cases, constituents are extracted from their mother constituents and reattached at higher constituents. In the case of the discontinuous VP in Figure 1, it leads to a VP rule with the following daughters: head (*HD*) and modifier (*MO*), while the accusative object is directly attached at the sentence level as a sister of the VP. This conversion leads to inconsistencies in the training data since the annotation scheme requires that object NPs are daughters of the VP rather than of S. The inconsistency introduced by tree conversion are considerable since they cover appr. 30% of all Negra trees (cf. section 3). One possible explanation for the better performance of Tüba-D/Z might be that it has more information about the correct attachment site of extraposed constituents, which is completely lacking in the context-free version of Negra. For this reason, Kübler (2005) and Maier (2006) tested a version of Negra which contained information of the original attachment site of these discontinuous constituents. In this version of Negra, the grammatical function *OA* in Figure 2 would be changed to *OA*$<$*VP* to show that it was originally attached to the VP. Experiments with this version showed a decrease in F-score from 52.30 to 49.75. Consequently, adding this information in a similar way to the encoding of discontinuous constituents in Tüba-D/Z harms performance.

By contrast, TüBa-D/Z uses topological fields as the primary structuring principle, which leads to a purely context-free annotation of discontinuous structures. There is evidence that the use of topological fields is advantageous also for other parsing approaches (Frank et al., 2003; Kübler, 2005; Maier, 2006).

Another difference in the annotation schemes concerns the treatment of phrases. Negra phrases are flat, and unary projections are not annotated. TüBa-D/Z always projects to the phrasal category and annotates more phrase-internal structure. The deeper structures in TüBa-D/Z lead to fewer rules for phrasal categories, which allows the parser a more consistent treatment of such phrases. For example, the direct attachment of one word subjects on the clausal level in Negra leads to a high number of different S rules with different POS tags for the subject phrase. An empirical proof for the assumption that flat phrase structures and the omission of unary nodes decrease parsing results is presented by Kübler (2005) and Maier (2006).

We want to emphasize that our experiments concentrate on the original context-free annotations of the treebanks. We did not investigate the influence of treebank refinement in this study. However, we would like to note that by a combination of suffix analysis and smoothing, Dubey (2005) was able to obtain an F-score of 85.2 for Negra. For other work in the area of treebank refinement using the German treebanks see Kübler (2005), Maier (2006), and Ule (2003).

9 Conclusion and Future Work

We have presented a comparative study of probabilistic treebank parsing of German, using the Negra and TüBa-D/Z treebanks. Experiments with the Stanford parser, which uses a factored PCFG and dependency model, show that, contrary to previous claims for other parsers, lexicalization of PCFG models boosts parsing performance for both treebanks. The experiments also show that there is a big difference in parsing performance, when trained on the Negra and on the TüBa-D/Z treebanks. This difference remains constant across

lexicalized, unlexicalized (also using the LoPar parser), and markovized models and also extends to parsing of major grammatical functions. Parser performance for the models trained on TüBa-D/Z are comparable to parsing results for English with the Stanford parser, when trained on the Penn treebank. This comparison at least suggests that German is not harder to parse than its West-Germanic neighbor language English.

Additional experiments with the TüBa-D/Z treebank are planned in future work. A new release of the TüBa-D/Z treebank has become available that includes appr. 22 000 trees, instead of the release with 15 000 sentences used for the experiments reported in this paper. This new release also contains morphological information at the POS level, including case and number. With this additional information, we expect considerable improvement in grammatical function assignment for the functions *subject*, *accusative object*, and *dative object*, which are marked by nominative, accusative, and dative case, respectively.

Acknowledgments

We are grateful to Helmut Schmid and to Chris Manning and his group for making their parsers publicly available as well as to Tylman Ule for providing the evaluation scripts. We are also grateful to the anonymous reviewers for many helpful comments. And we are especially grateful to Roger Levy for all the help he gave us in creating the language pack for TüBa-D/Z in the Stanford parser.

References

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

Erich Drach. 1937. *Grundgedanken der Deutschen Satzlehre*. Diesterweg, Frankfurt/M.

Amit Dubey and Frank Keller. 2003. Probabilistic parsing for German using sister-head dependencies. In *Proceedings of ACL 2003*, pages 96–103, Sapporo, Japan.

Amit Dubey. 2005. What to do when lexicalization fails: Parsing German with suffix analysis and smoothing. In *Proceedings of ACL 2005*, Ann Arbor, MI.

Oskar Erdmann. 1886. *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt*. Verlag der Cotta'schen Buchhandlung, Stuttgart, Germany.

Anette Frank, Markus Becker, Berthold Crysmann, Bernd Kiefer, and Ulrich Schäfer. 2003. Integrated shallow and deep parsing: TopP meets HPSG. In *Proceedings of ACL 2003*, Sapporo, Japan.

Tilman Höhle. 1986. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Dan Klein and Christopher Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of ACL 2003*, pages 423–430, Sapporo, Japan.

Dan Klein and Christopher Manning. 2003b. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pages 3–10, Vancouver, Canada.

Sandra Kübler. 2005. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *Proceedings of RANLP 2005*, Borovets, Bulgaria.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of ACL 2003*, pages 439–446, Sapporo, Japan.

Wolfgang Maier. 2006. Annotation schemes and their influence on parsing results. In *Proceedings of the ACL-2006 Student Research Workshop*, Sydney, Australia.

Michael Schiehlen. 2004. Annotation strategies for probabilistic parsing in German. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Helmut Schmid. 2000. LoPar: Design and implementation. Technical report, Universität Stuttgart, Germany.

Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of ANLP 1997*, Washington, D.C.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister, 2005. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

Tylman Ule. 2003. Directed treebank refinement for PCFG parsing. In *Proceedings of TLT 2003*, Växjö, Sweden.