Revisiting the impact of different annotation schemes on PCFG parsing: A grammatical dependency evaluation

Adriane Boyd
Department of Linguistics
The Ohio State University
1712 Neil Avenue
Columbus, Ohio 43210, USA
adriane@ling.osu.edu

Detmar Meurers
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstrasse 19
72074 Tübingen, Germany
dm@sfs.uni-tuebingen.de

Abstract

Recent parsing research has started addressing the questions a) how parsers trained on different syntactic resources differ in their performance and b) how to conduct a meaningful evaluation of the parsing results across such a range of syntactic representations. Two German treebanks, Negra and TüBa-D/Z, constitute an interesting testing ground for such research given that the two treebanks make very different representational choices for this language, which also is of general interest given that German is situated between the extremes of fixed and free word order. We show that previous work comparing PCFG parsing with these two treebanks employed PARSEVAL and grammatical function comparisons which were skewed by differences between the two corpus annotation schemes. Focusing on the grammatical dependency triples as an essential dimension of comparison, we show that the two very distinct corpora result in comparable parsing performance.

1 Introduction

Syntactically annotated corpora have been produced for a range of languages and they differ significantly regarding which language properties are encoded and how they are represented. Between the two extremes of constituency treebanks for English and dependency treebanks for free word order languages such as Czech lie languages such as German, for which two different treebanks have explored different options for encoding topology and dependency, Negra (Brants et al., 1999) and TüBa-D/Z (Telljohann et al., 2005).

Recent research has started addressing the question of how parsers trained on these different syntactic resources differ in their performance. Such work must also address the question of how to conduct a meaningful evaluation of the parsing results across such a range of syntactic representations. In this paper, we show that previous work comparing PCFG parsing for the two German treebanks used representations which cannot adequately be compared using the given PARSEVAL measures and that a grammatical dependency evaluation is more meaningful than the grammatical function evaluation provided.

We present the first comparison of Negra and TüBa-D/Z using a labeled dependency evaluation based on the grammatical function labels provided in the corpora. We show that, in contrast to previous literature, a labeled dependency evaluation establishes that PCFG parsers trained on the two corpora give similar parsing performance. The focus on labeled dependencies also provides a direct link to recent work on dependency-based evaluation (e.g., Clark and Curran, 2007) and dependency parsing (e.g., CoNLL shared tasks 2006, 2007).

1.1 Previous work

The question of how to evaluate parser output has naturally already arisen in earlier work on parsing English. As discussed by Lin (1995) and others, the PARSEVAL evaluation typically used to analyze the performance of statistical parsing models has many drawbacks. Bracketing evaluation may count a single error multiple times and does not differentiate between errors that significantly affect the interpretation of the sentence and those that are less crucial.

Proceedings of the ACL-08: HLT Workshop on Parsing German (PaGe-08), pages 24–32,
Columbus, Ohio, USA, June 2008.

It also does not allow for evaluation of particular syntactic structures or provide meaningful information about where the parser is failing. In addition, and most directly relevant for this paper, PARSEVAL scores are difficult to compare across syntactic annotation schemes (Carroll et al., 2003).

At the same time, previous research on PCFG parsing using treebank training data present PARSEVAL measures in comparing the parsing performance for different languages and annotation schemes, reporting a number of striking differences. For example, Levy and Manning (2003), Kübler (2005), and Kübler et al. (2006) highlight the significant effect of language properties and annotation schemes for German and Chinese treebanks. In related work, parser enhancements that provide a significant performance boost for English, such as head lexicalization, are reported not to provide the same kind of improvement, if any, for German (Dubey and Keller, 2003; Dubey, 2004; Kübler et al., 2006).

Previous work has compared the similar Negra and Tiger corpora of German to the very different TüBa-D/Z corpus. Kübler et al. (2006) compares the Negra and TüBa-D/Z corpora of German using a PARSEVAL evaluation and an evaluation on core grammatical function labels that is included to address concerns about the PARSEVAL measure.[1] Using the Stanford Parser (Klein and Manning, 2002), which employs a factored PCFG and dependency model, they claim that the model trained on TüBa-D/Z consistently outperforms that trained on Negra in PARSEVAL and grammatical function evaluations. Dubey (2004) also includes an evaluation on grammatical function for statistical models trained on Negra, but obtains very different results from Kübler et al. (2006).[2]

In recent related work, Rehbein and van Genabith (2007a) demonstrate using the Tiger and TüBa-D/Z corpora of German that PARSEVAL is inappropriate for comparisons of the output of PCFG parsers trained on different treebank annotation schemes because PARSEVAL scores are affected by the ratio of terminal to non-terminal nodes. A dependency-based evaluation on triples of the form *word-POS-head* shows better results for the parser trained on Tiger even though the much lower PARSEVAL scores, if meaningful, would predict that the output for Tiger is of lower quality. However, their dependency-based evaluation does not make use of the grammatical function labels, which are provided in the corpora and closely correspond to the representations used in recent work on formalism-independent evaluation of parsers (e.g., Clark and Curran, 2007).[3]

Addressing these issues, we resolve the apparent discrepancy between Kübler et al. (2006) and Dubey (2004) and establish a firm grammatical function comparison of Negra and TüBa-D/Z. We also extend the evaluation to a labeled dependency evaluation based on grammatical relations for both corpora. Such an evaluation, which abstracts away from the specifics of the annotation schemes, shows that, in contrast to the claims made in Kübler et al. (2006), the parsing results for PCFG parsers trained on these heterogeneous corpora are very similar.

2 The corpora used

As motivated in the introduction, the work discussed in this paper is based on two German corpora, Negra and TüBa-D/Z, which differ significantly in the syntactic representations used – thereby offering an interesting test bed for investigating the influence of an annotation scheme on the parsers trained.

2.1 Negra

The Negra corpus (Brants et al., 1999) consists of newspaper text from the *Frankfurter Rundschau*, a German newspaper. Version 2 of the corpus contains 20,602 sentences. It uses the STTS tag set (Schiller et al., 1995) for part-of-speech annotation. There are 25 non-terminal node labels and 46 edge labels.

The syntactic annotation of Negra combines features from phrase structure grammar and depen-

[1] The evaluation is based only on the grammatical function; it does not identify the dependency pair that it labels.

[2] While the focus of Kübler et al. (2006) is on comparing parsing results across corpora, Dubey (2004) focuses on improving parsing for Negra, including corpus-specific enhancements leading to better results. This difference in focus and additional differences in experimental setup mean that a fine-grained comparison of the results is inappropriate – the relevant point here is that the gap between the results (23% for subjects, 35% for accusative objects) warrants further attention in the context of comparing parsing results across corpora.

[3] Their evaluation also introduces an additional level of complexity by finding heads heuristically rather than relying on the head labels present on some elements in each corpus.

dency grammar using a tree-like syntactic structure with grammatical functions labeled on the edges of the tree. Flat sentence structures are used in many places to avoid attachment ambiguities and non-branching phrases are not used.

The annotation scheme emphasizes the use of the tree structure to encode grammatical dependencies, representing a head and all its dependents within a local tree regardless of whether a dependent is realized near its head or not, e.g., because it has been extraposed or fronted. Since traditional syntax trees do not permit the crossing branches needed to license discontinuous constituents, Negra uses a "syntax graph" data structure to represent the annotation. An example of a syntax graph with a discontinuous constituent (VP) due to a fronted dative object (NP) is shown in Figure 1.

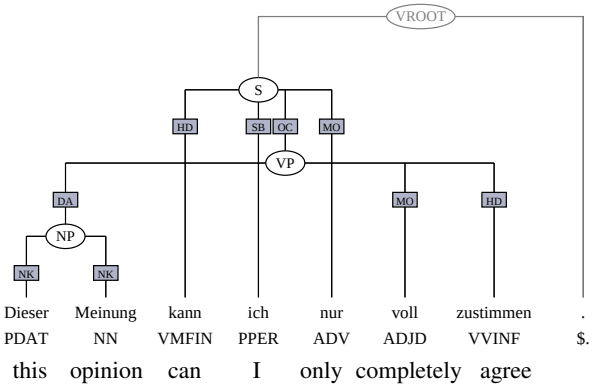

Figure 1: Negra tree for '*I can only agree with this opinion completely.*'

Negra uses flat NP and PP annotation with no marked heads. For example, both *Dieser* and *Meinung* in Figure 1 have the grammatical function label "NK". Since unary branching is not used in Negra, a bare noun or pronoun argument is not dominated by an NP node, as shown by the pronoun *ich* above.

A verbal head in Negra is always marked with the edge label "HD" and its arguments are its sisters in the local tree. The subject is always the sister of the finite verb, which is a daughter of S. If the finite verb is the main verb in the clause, the objects are also its sisters, i.e., the finite verb, subject and objects are all daughters of S. If the main verb is an auxiliary governing a non-finite main verb, the non-finite verb and its objects and modifiers form a VP where the objects are sisters of the non-finite verb as in Figure 1. The VP is then a sister of the finite verb.

The finite verb in a German declarative clause appears in the so-called verb-second position, immediately following the fronted constituent. As a result, the VP in Negra is discontinuous whenever one of its children has been fronted, as in the common word orders exemplified in (1a) and (1b).

(1) a. **Die Tür** hat Anna **wieder zugeschlagen**.
the door has Anna again slammed-shut
'Anna slammed the door shut again.'

b. **Wieder** hat Anna **die Tür zugeschlagen**.
again has Anna the door slammed-shut
'Anna slammed the door shut again.'

The sentence we saw in Figure 1 contains a discontinuous VP with a fronted dative object (*Dieser Meinung*). The dative object and a modifier (*voll*) form a VP with the non-finite verb (*zustimmen*).

2.2 TüBa-D/Z

The TüBa-D/Z corpus, version 2, (Telljohann et al., 2005) consists of 22,091 sentences of newspaper text from the German newspaper *die tageszeitung*. Like Negra, it uses the STTS tag set (Schiller et al., 1995) for part-of-speech annotation. Syntactically it uses 27 non-terminal node labels and 47 edge labels.

The syntactic annotation incorporates a topological field analysis of the German clause (Reis, 1980; Höhle, 1986), which segments a sentence into topological units depending on the position of the finite verb (verb-first, verb-second, verb-last). In a verb-first and verb-second sentence, the finite verb is the left bracket (LK), whereas in a verb-last subordinate clause, the subordinating conjunction occupies that field. In all clauses, the non-finite verb cluster forms the right bracket (VC), and arguments and modifiers can appear in the middle field (MF) between the two brackets. Extraposed material is found to the right of the right bracket, and in a verb-second sentence one constituent appears in the fronted field (VF) preceding the finite verb. By specifying constraints on the elements that can occur in the different fields, the word order in any type of German clause can be concisely characterized.

Each clause in the TüBa-D/Z corpus is divided into topological fields at the top level, and each topological field contains phrase-level annotation. An

example sentence from TüBa-D/Z is shown in Figure 2, where the topological fields VF, LK, MF, and VC are visible under the SIMPX clause node.

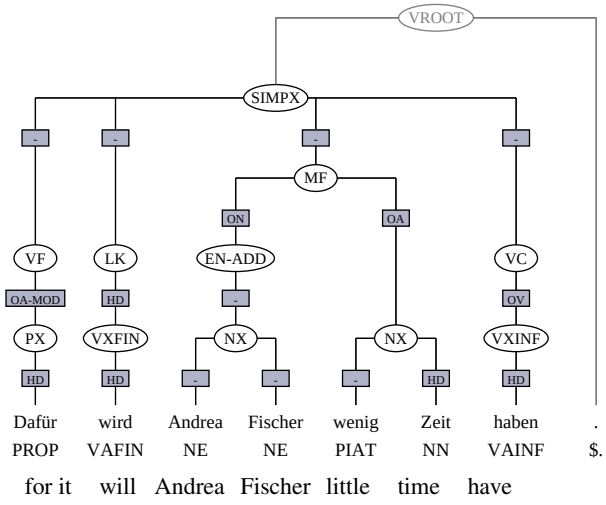

Figure 2: TüBa-D/Z tree for '*Andrea Fischer will have little time for it.*'

Edge labels are used to mark heads and grammatical functions, even though it can be nontrivial to figure out which grammatical function belongs to which head given that heads and their arguments often are in separate topological fields. For example, in Figure 2 the subject noun chunk (NX) has the edge label ON (object - nominative) and the object noun chunk has the edge label OA (object - accusative); both are realized within the middle field (MF), while the finite verb (VXFIN) marked as HD (head) is in the left sentence bracket (LK). This issue becomes relevant in section 3.4.2, discussing an evaluation based on labeled dependency triples.

Where Negra uses discontinuous constituents, TüBa-D/Z uses special edge labels to annotate grammatical relations which are not locally realized. For example, the fronted prepositional phrase (PX) in Figure 2 has the edge label OA-MOD which needs to be matched with the noun phrase (NX) with label OA that is found in the MF field.

2.3 Comparing Negra and TüBa-D/Z

To give an impression of how the different annotation schemes affect the appearance of a typical tree in the two corpora, Table 1 provides statistics on average sentence length and the number of non-terminals per sentence.

	Negra	TüBa-D/Z
No. of Sentences	20,602	22,091
Terminals/Sentence	17.2	17.3
Non-terminals/Sentence	7.0	20.7

Table 1: General Characteristics of the Corpora

While the sentences in Negra and TüBa-D/Z on average have the same number of words, the average TüBa-D/Z sentence has nearly three times as many non-terminal nodes as the average Negra sentence. This difference is mainly due to the extra level of topological fields annotation and the use of more contoured structures in many places where Negra uses flatter structures.

3 Experiments

The goal of the following experiments is a comparison of parsing performance across different types of evaluation metrics for parsers trained on Negra (Ver. 2) and TüBa-D/Z (Ver. 2).

3.1 Data Preparation

Following Kübler et al. (2006), only sentences with fewer than 35 words were used, which results in 20,002 sentences for Negra and 21,365 sentences for TüBa-D/Z. Because punctuation is not attached within the sentence in the corpus annotation, punctuation was removed.

To be able to train PCFG parsing models, it is necessary to convert the syntax graphs encoding trees with discontinuities in Negra into traditional syntax trees. Around 30% of sentences in Negra contain at least one discontinuity. To remove discontinuities, we used the conversion program included with the Negra corpus annotation tools (Brants and Plaehn, 2000), the same tool used in Kübler et al. (2006), which raises non-head elements to a higher tree until there are no more discontinuities. For example, for the discontinuous tree with a fronted object we saw in Figure 1, the PP containing the fronted NP *Dieser Meinung* is raised to become a daughter of the top S node.[4]

Additionally, the edge labels used in both corpora need to be folded into the node labels to become a

[4] An alternate method that avoids certain problems with this raising method is discussed in Boyd (2007).

part of context-free grammar rules used by a PCFG parser. In the Penn Treebank-style versions of the corpora appropriate for training a PCFG parser, each edge label is joined with the phrase or POS label on the phrase or word immediately below it. Both corpora include edge labels above all phrases and words. However the flatter structures in Negra result in 39 different edge labels on words while TüBa-D/Z has only 5.

Unlike Kübler et al. (2006), which ignored edge labels on words, we incorporate all edge labels present in both corpora. As a consequence of this, providing a parser with perfect lexical tags would also provide the edge label for that word. TüBa-D/Z does not annotate grammatical functions other than HD on words, but Negra includes many grammatical functions on words. Including edge labels in the perfect lexical tags would artificially boost the results of a grammatical function evaluation for Negra since it amounts to providing the correct grammatical function for the 38% of arguments in Negra that are single words.

To avoid this problem, we introduced non-branching phrasal nodes into Negra to prevent the correct grammatical function label from being provided with the perfect lexical tag in the cases of single-word arguments, which are mostly bare nouns and pronouns. We added phrasal nodes above all single-word subject, accusative object, dative object, and genitive object[5] arguments, with the category of the inserted phrase depending on the POS tag on the word. The introduced phrasal node is given the word's original grammatical function label; the grammatical function label of the word itself becomes NK for NPs and HD for APs and VPs. In total, 14,580 nodes were inserted into Negra in this way. TüBa-D/Z has non-branching phrases above all single-word arguments, so that no such modification was needed.[6]

3.2 Experimental Setup

We trained unlexicalized PCFG parsing models using LoPar (Schmid, 2000). Unlexicalized models were used to minimize the impact of other corpus differences on parsing. A ten-fold cross validation was performed for all experiments.[7]

3.3 PARSEVAL Evaluation

As a reference point for comparison with previous work, the PARSEVAL results[8] are given in Table 2.

	Negra	TüBa-D/Z
Unlabeled Precision	78.69	89.92
Unlabeled Recall	82.29	86.48
Labeled Precision	64.08	75.36
Labeled Recall	67.01	72.47
Coverage	97.00	99.90

Table 2: PARSEVAL Evaluation

The parser trained on TüBa-D/Z performs much better than the one trained on Negra on all labeled and unlabeled bracketing scores. As we saw in section 2, Negra and TüBa-D/Z use very different syntactic annotation schemes, resulting in over 2.5 times as many non-terminals per sentence in TüBa-D/Z as in Negra with the additional unary nodes. As mentioned previously, Rehbein and van Genabith (2007a) showed that PARSEVAL is affected by the ratio of terminal to non-terminal nodes, so these results are not expected to indicate the quality of the parses. The comparison with grammatical function and dependency evaluations we turn to next showcases that PARSEVAL does not provide a meaningful evaluation metric across annotation schemes.

3.4 Dependency Evaluation

Complementing the issue of the ratio of terminals to non-terminals raised in the last section, one can question whether counting all brackets in the sentence equally, as done by the PARSEVAL metric, provides a good measure of how accurately the basic functor-argument structure of the sentence has been captured in a parse. Thus, it is useful to per-

[5]Genitive objects are modified for the sake of consistency among arguments even though there are too few genitive objects to provide reliable results in the evaluation.

[6]The addition of edge labels to terminal POS labels results in 337 lexical tags for Negra and 91 for TüBa-D/Z.

[7]Our experimental setup is designed to support a comparison between Negra and TüBa-D/Z for the three evaluation metrics and is intended to be comparable to the setup of Kübler et al. (2006). For Negra, Dubey (2004) explores a range of parsing models and the corpus preparation he uses differs from the one discussed in this paper so that a discussion of his results is beyond the scope of the corpus comparison in this paper.

[8]Scores were calculated using `evalb`.

form an evaluation based on the grammatical function labels that are important for determining the functor-argument structure of the sentence: subjects, accusative objects, and dative objects.[9] The first step in an evaluation of functor-argument structure is to identify whether an argument bears the correct grammatical function label.

3.4.1 Grammatical Function Label Evaluation

Kübler et al. (2006) present the results shown in Table 3 for the parsing performance of the unlexicalized model of the Stanford Parser (Klein and Manning, 2002). In this grammatical function label evaluation, TüBa-D/Z outperforms Negra for subjects, accusative objects, and dative objects based on an evaluation of phrasal arguments.

	Negra			TüBa-D/Z		
	Prec	Rec	F	Prec	Rec	F
Subj	52.50	58.02	55.26	66.82	75.93	72.38
Acc	35.14	36.30	35.72	43.84	47.31	45.58
Dat	8.38	3.58	5.98	24.46	9.96	17.21

Table 3: Grammatical Function Label Evaluation for Phrasal Arguments from Kübler et al. (2006)

Note that this grammatical function label evaluation is restricted to labels on phrases; grammatical function labels on words are ignored in training and testing. This results in an unbalanced comparison between Negra and TüBa-D/Z since, as discussed in section 2, TüBa-D/Z includes unary-branching phrases above all single-word arguments whereas Negra does not. In effect, single-word arguments in Negra – mainly pronouns and bare nouns – are not considered in the evaluation from Kübler et al. (2006). The result is thus a comparison of multi-word arguments in Negra to both single- and multi-word arguments in TüBa-D/Z. Recall from section 3.1 that this is not a minor difference: single-word arguments account for 38% of subjects, accusative objects, and dative objects in Negra.

As discussed in the data preparation section, Negra was modified for our experiment so as not to provide the parser with the grammatical function labels for single word phrases as part of the perfect tags provided. This evaluation handles multiple categories of arguments, not just NPs, so it focuses solely on the grammatical function labels, ignoring the phrasal categories. For example, in Negra an NP-OA in a parse is considered a correct accusative object even if the OA label in the gold standard has the category MPN. The results are shown in Table 4.

	Negra			TüBa-D/Z		
	Prec	Rec	F	Prec	Rec	F
Subj	69.69	69.12	69.42	65.74	72.24	68.99
Acc	48.17	50.97	49.57	41.37	46.81	44.09
Dat	20.93	15.22	18.08	21.40	11.51	16.46

Table 4: Grammatical Function Label Evaluation

In contrast to the results for NP grammatical functions of Kübler et al. (2006) we saw in Table 3, Negra and TüBa-D/Z perform quite similarly overall, with Negra slightly outperforming TüBa-D/Z for all types of arguments.

These results also form a clear contrast to the PARSEVAL results we saw in Table 2. Contrary to the finding in Kübler et al. (2006), the PARSEVAL evaluation does not echo the grammatical function label evaluation. In keeping with the results from Rehbein and van Genabith (2007a), we find that PARSEVAL is not an adequate predictor of performance in an evaluation targeting the functor-argument structure of the sentence for comparisons between PCFG parsers trained on corpora with different annotation schemes.

3.4.2 Labeled Dependency Triple Evaluation

While determining the grammatical function of an element is an important part of determining the functor-argument structure of a sentence, the other necessary component is determining the head of each function. To evaluate whether both the functor and the argument have been correctly found, an evaluation of labeled dependency triples is needed. As in the previous section, we focus on the grammatical function labels for arguments of verbs. To complete a labeled dependency triple for each argument, we additionally need to locate the lexical verbal head.

In Negra, the head is the sister of an argument marked with the function label "HD", however

[9] Genitive objects are also annotated in both corpora, but they are too infrequent to provide meaningful results. As discussed in Rehbein and van Genabith (2007b), labels such as subject (SB for Negra, ON for TüBa-D/Z) are not necessarily comparable in all instances, but such cases are infrequent.

heads are only marked for a subset of the phrase categories: S, VP, AP, and AVP.[10] This subset includes the phrase categories that contain verbs and their arguments, S and VP. In our experiment, the parser finds the HD grammatical function labels with a very high f-score: 99.5% precision and 96.5% recall. If the sister with the label HD is a word, then that word is the lexical head for the purposes of this dependency evaluation. If the sister with the label HD is a phrase, then a recursive search for heads within that phrase finds a lexical head. In 3.2% of cases in the gold standard, it is not possible to find a lexical head for an argument. Further methods could be applied to find the remaining heads heuristically, but we avoid the additional parameters this introduces for this evaluation by ignoring these cases.

For TüBa-D/Z, finding the head is not as simple because the verbal head and its arguments are in different topological fields. To create a parallel comparison to Negra, the finite verb from the local clause is chosen as the head for all subjects. The (finite or non-finite) main full verb is designated as the head for the accusative and dative objects. It is possible to automatically find an appropriate head verb for all but 2.7% of subjects, accusative objects, and dative objects.[11] As with Negra, only cases where a head verb can be found in the gold standard are considered in the evaluation.

As in the grammatical function evaluation in the previous section, only the grammatical function label, not the phrase category is considered in the evaluation. The results for the labeled dependency evaluation are shown in Table 5. The parser trained on Negra outperforms the one trained on TüBa-D/Z for all types of arguments.

	Negra			TüBa-D/Z		
	Prec	Rec	F	Prec	Rec	F
Subj	72.84	69.03	70.93	60.52	65.98	63.25
Acc	47.96	48.80	48.38	37.39	40.83	39.11
Dat	19.56	14.01	16.79	19.32	10.39	14.85

Table 5: Labeled Dependency Evaluation

4 Discussion of Results

Comparing PARSEVAL scores for a parser trained on the Negra and the TüBa-D/Z corpus with a grammatical function and a labeled dependency evaluation, we confirm that the PARSEVAL scores do not correlate with the scores in the other two evaluations, which given their closeness to the semantic functor argument structure make meaningful targets for evaluating parsers.

Shifting the focus to the grammatical function evaluation, we showed that a grammatical function evaluation based on phrasal arguments as provided by Kübler et al. (2006) is inadequate for comparing parsers trained on the Negra and TüBa-D/Z corpora. By introducing non-branching phrase nodes above single-word arguments in Negra, it is possible to provide a balanced comparison for the grammatical function label evaluation between Negra and TüBa-D/Z on both phrasal and single-word arguments. The models trained on both corpora perform very similarly in the grammatical function evaluation, in contrast to the claims in Kübler et al. (2006).

When the grammatical function label evaluation is extended into a labeled dependency evaluation by finding the verbal head to complete the labeled dependency triple, the parser trained on Negra outperforms that trained on TüBa-D/Z. The more significant drop in results for TüBa-D/Z compared to the grammatical function label evaluation may be due to the fact that a verbal lexical head in TüBa-D/Z is not in the same local tree as its dependents, whereas it is in Negra. The presence of intervening topological field nodes in TüBa-D/Z may make it difficult for the parser to consistently identify the elements of the dependency triple across several subtrees.

The Negra corpus annotation scheme makes it simple to identify the heads of verb arguments, but the flat NP and PP structures make it difficult to extend a labeled dependency analysis beyond verb arguments. On the other hand, TüBa-D/Z has marked heads in NPs and PPs, but it is not as easy to pair verb arguments with their heads because the verbs are in separate topological fields from their argu-

[10] However, some strings labeled as S and VP do not contain a head and thus lack a daughter with a HD function label.

[11] The relative numbers of instances where a lexical head is not found are comparable for Negra and TüBa-D/Z. Heads are not found for approximately 4% of subjects, 1% of accusative objects, and 1% of dative objects. These instances are frequently due to elision of the verb in headlines and coordinated clauses.

ments. For a constituent-based corpus annotation scheme to lend itself to a thorough labeled dependency evaluation, heads should be marked clearly for all phrase categories and all non-head elements need to have marked grammatical functions.

The presence of topological field nodes in TüBa-D/Z deserves more discussion in relation to a grammatical dependency evaluation. The corpus contains two very different types of nodes in its syntactic trees: nodes such as NP and PP that correspond to constituents and nodes such as VF (Vorfeld) and MF (Mittelfeld) that correspond to word order domains. Constituents such as NP have grammatical relations to other elements in the sentence and have identifiable heads within them, whereas nodes encoding word order domains have neither.[12] While constituents and word order domains sometimes coincide, such as the Vorfeld normally consisting of a single constituent, this is not the general case. For example, the Mittelfeld often contains multiple constituents which each stand in different grammatical relations to the verb(s) in the left and right sentence brackets (LK and VC).

Returning to the issue of finding dependencies between constituents, the intervening word order domain nodes can make it non-trivial to determine these relations in TüBa-D/Z. For example, word order domain nodes will always intervene between a verb and its arguments. In order to have all grammatical dependencies directly encoded in the treebank, it would be preferable for corpus annotation schemes to ensure that a homogeneous constituency representation can be easily obtained.

5 Future Work

An evaluation on arguments of verbs is just a first step in working towards a more complete labeled dependency evaluation. Because Negra and TüBa-D/Z do not have parallel uses of many grammatical function labels beyond arguments of verbs, a more detailed evaluation on more types of dependency relations will require a complex dependency conversion method to provide comparable results.

[12] While the focus in this work is on unlexicalized parsing, this also calls into question the effect of head lexicalization for a corpus that contains elements that by their nature are not the types of elements that have heads.

Since previous work on head-lexicalized parsing models for German has focused on PARSEVAL evaluations, it would also be useful to perform a labeled dependency evaluation to determine what effect head lexicalization has on particular constructions for the parsers. Because of the concerns discussed in the previous section and the difference in which types of clauses have marked heads in Negra and TüBa-D/Z, the effect of head lexicalization on the parsing results may differ for the two corpora.

6 Conclusion

Addressing the general question of how to compare parsing results for different annotation schemes, we revisited the comparison of PCFG parsing results for the Negra and TüBa-D/Z corpora. We show that these different annotation schemes lead to very significant differences in PARSEVAL scores for unlexicalized PCFG parsing models, but grammatical function label and labeled dependency evaluations for arguments of verbs show that this difference does not carry over to measures which are relevant to the semantic functor-argument structure. In contrast to Kübler et al. (2006) a grammatical function evaluation on subjects, accusative objects, and dative objects establishes that Negra and TüBa-D/Z perform similarly when all types of words and phrases appearing as arguments are taken into consideration. A labeled dependency evaluation based on grammatical relations, which links this work to current work on formalism-independent parser evaluation (e.g., Clark and Curran, 2007), shows that the parsing performance for Negra and TüBa-D/Z is comparable.

References

Adriane Boyd, 2007. Discontinuity Revisited: An Improved Conversion to Context-Free Representations. In *Proceedings of the Linguistic Annotation Workshop (LAW)*. Prague, Czech Republic.

Thorsten Brants and Oliver Plaehn, 2000. Interactive Corpus Annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Thorsten Brants, Wojciech Skut and Hans Uszkoreit, 1999. Syntactic Annotation of a German Newspaper Corpus. In *Proceedings of the ATALA Treebank Workshop*. Paris, France.

John Carroll, Guido Minnen and Ted Briscoe, 2003. Parser evaluation: using a grammatical relation annotation scheme. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, Kluwer, Dordrecht.

Stephen Clark and James Curran, 2007. Formalism-Independent Parser Evaluation with CCG and DepBank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*. Prague, Czech Republic.

Amit Dubey, 2004. Statistical Parsing for German: Modeling Syntactic Properties and Annotation Differences. Ph.D. thesis, Universität des Saarlandes.

Amit Dubey and Frank Keller, 2003. Probabilistic Parsing Using Sister-Head Dependencies. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sapporo, Japan.

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

Dan Klein and Christopher D. Manning, 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS)*. Vancouver, British Columbia, Canada.

Sandra Kübler, 2005. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgaria.

Sandra Kübler, Erhard W. Hinrichs and Wolfgang Maier, 2006. Is it really that difficult to parse German? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sydney, Australia.

Roger Levy and Christopher Manning, 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dekang Lin, 1995. A Dependency-based Method for Evaluating Broad-Coverage Parsers. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. Montreal, Quebec, Canada.

Ines Rehbein and Josef van Genabith, 2007a. Treebank Annotation Schemes and Parser Evaluation for German. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic.

Ines Rehbein and Josef van Genabith, 2007b. Why is it so difficult to compare treebanks? TIGER and TüBa-D/Z revisited. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT)*. Bergen, Norway.

Marga Reis, 1980. On Justifying Topological Frames: 'Positional Field' and the Order of Nonverbal Constituents in German. *Documentation et Recherche en Linguistique Allemande Contemporaine Vincennes (DRLAV)*, 22/23.

Anne Schiller, Simone Teufel and Christine Thielen, 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical report, Universität Stuttgart, Universität Tübingen, Germany.

Helmut Schmid, 2000. *LoPar: Design and Implementation*. Arbeitspapiere des Sonderforschungsbereiches 340 No. 149, Universität Stuttgart.

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