The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone

Heike Telljohann, Erhard Hinrichs, Sandra Kübler

Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 19
72074 Tübingen
Germany
{hschulz,eh,kuebler}@sfs.uni-tuebingen.de

Abstract

The purpose of this paper is to describe the T¨uBa-D/Z treebank of written German and to compare it to the independently developed TIGER treebank (Brants et al., 2002). Both treebanks, TIGER and T¨uBa-D/Z, use an annotation framework that is based on phrase structure grammar and that is enhanced by a level of predicate-argument structure. The comparison between the annotation schemes of the two treebanks focuses on the different treatments of free word order and discontinuous constituents in German as well as on differences in phrase-internal annotation.

1. Introduction

The purpose of this paper is to describe the TüBa-D/Z treebank (Telljohann et al., 2003) of written German and to compare it to the independently developed TIGER treebank (Brants et al., 2002). Both treebanks use as their data source German newspaper material: the Frankfurter Rundschau newspaper corpus for TIGER and the 'die tageszeitung' (taz) newspaper corpus for TüBa-D/Z. While TIGER provides 40.000 annotated sentences, the TüBa-D/Z treebank comprises at present appr. 15.000 trees. While smaller in size, the TüBa-D/Z treebank has reached a size that has proven feasible as training material for a variety of tasks in data-driven NLP: morphological disambiguation (Hinrichs and Trushkina, 2002), partial parsing (Müller and Ule, 2002), and topological field parsing (Liepert, 2003; Ule, 2003; Veenstra et al., 2002).

Both treebanks, TIGER and TüBa-D/Z, use an annotation framework that is based on phrase structure grammar and that is enhanced by a level of predicate-argument structure. Annotation is performed semi-automatically by the graphical tool Annotate (Plaehn, 1998).

2. The TüBa-D/Z Annotation Scheme

The TüBa-D/Z treebank was made available for research and development purposes in December 2003[1] along with a detailed stylebook (Telljohann et al., 2003). It therefore seems timely to provide a detailed comparison of the TüBa-D/Z and TIGER annotation schemes so as to provide potential users of the data with an overview of the similarities and differences between these language resources.

The TüBa-D/Z annotation scheme is derived from the Verbmobil treebank of spoken German (Hinrichs et al., 2000), but has been extended along various dimensions to accommodate the characteristics of written texts.

The Verbmobil treebank annotation scheme distinguishes four levels of syntactic constituency: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields, which characterize the word order regularities among different clause types of German, and which are widely accepted among descriptive linguists of German (cf. e.g. (Drach, 1937; Höhle, 1986)). In addition to constituent structure, annotated trees contain edge labels between nodes. These edge labels encode grammatical functions (as relations between phrases) and the distinction between heads (HD) and non-heads (-) (as phrase-internal relations). The tree in Figure 1 descibes the sentence `Wir müssen uns aber davor hüten, daß sich jeder Politiker einen eigenen Tempel baut.` ("But we have to prevent that every politician builds his own temple.")[2]. The sentence (SIMPX) is grouped into the following topological fields: initial field (VF), left sentence bracket (LK), middle field (MF), verb complex (VC), and final field (NF). The finite verb constitutes the head (HD) of the clause. The grammatical relations annotated in the tree are: subject (ON), accusative object (OA), dative object (OD), verbal object (OV), prepositional object (OPP), modifier of the prepositional object (OPP-MOD) (cf. section 3.1.), and modifier (MOD). The parts of speech are given below the lexical level. For POS tagging, the STTS (Schiller et al., 1995) is used.

Syntactic and semantic ambiguity is treated in terms of underspecification, which relies on the principle of high attachment and utilizes underspecified node labels. Whenever disambiguation is possible, a non-ambiguous label is chosen, such as for the modifier of the prepositional object (OPP-MOD) in the sentence in Figure 1. If the ambiguity cannot be resolved, the ambiguous modifier receives the underspecified label MOD and is attached to the highest possible node. The sentence in Figure 1 shows an example of such an ambiguous modifier (MOD), which modifies more than one constituent.

[1] For further information, please visit the T¨uBa-D/Z website at `http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml`.

[2] All trees in this paper follow the data format for trees defi ned by the NEGRA project of the Sonderforschungsbereich 378 at the University of the Saarland, Saarbr¨ucken. They were printed by the NEGRA annotation tool (Brants and Skut, 1998).

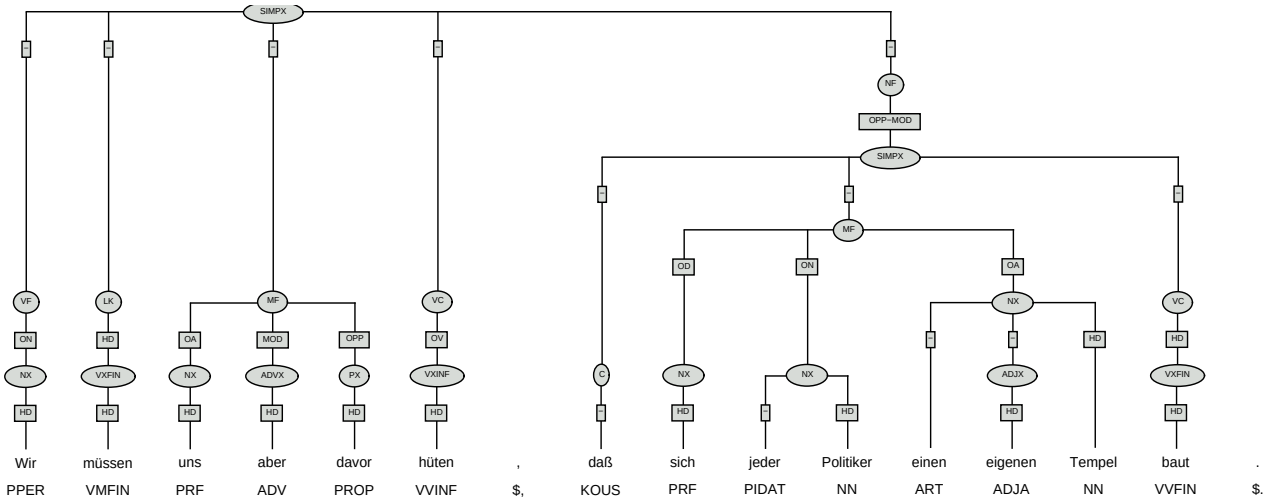

Figure 1: A TüBa-D/Z tree.

3. Differences in Annotation between TüBa-D/Z and TIGER

3.1. Context-free Backbone vs. Crossing Branches

There are two main differences in the annotation schemes of TüBa-D/Z and TIGER: the treatment of the relatively free word order in German and the treatment of discontinuous constituents.

For both phenomena, the TüBa-D/Z annotation relies on a context-free backbone (i.e. proper trees without crossing branches) of phrase structure combined with edge labels that specify the grammatical function of the phrase in question. By contrast, TIGER assumes a less constrained notion of tree, which includes the possibility of crossing branches. This allows for a purely tree structural account of word order variation without the need of dedicated function labels. In the sentence in Figure 2 taken from the TIGER treebank, `Bei den Gesprächen in London wurden zudem Hilfen für Osteuropa und die Sowjetunion behandelt.` ("During the discussions in London, support for Eastern Europe and the Sowjet Union were also addressed."), the extraposed prepositional phrase is attached directly to the VP whose head is the main verb at the end of the sentence, thus leading to a crossing branch.

At the sentence level, the differences in annotation strategy between TIGER and TüBa-D/Z manifest themselves most clearly in the treatment of subjects and of rightward extraposition. In TIGER, subjects and finite verbs are always treated as immediate daughters of the clause, regardless of their linear positions while non-finite verbs, their complements, and (prepositional) adjuncts are grouped together into single VP constituents (cf. Figure 2). This means that crossing dependencies arise in TIGER whenever the subject and/or the finite verb is surrounded by VP material. By contrast, in TüBa-D/Z, subjects and finite verbs do not occupy structurally invariant positions in the clause. Instead, they are grouped inside the relevant topological fields depending on the clause type in which they occur. Their syntactic function is indicated exclusively by grammatical function edge labels, thus obviating the need for crossing branches or additional mechanisms.

One characteristic of free word order is the frequent occurrence of discontinuous constituents, such as extraposed phrases or clauses. In TIGER, dependent constituents are always directly attached to their governing node. Thus, the annotation of discontinuous constituents results in crossing branches. In TüBa-D/Z, the annotation of such phenomena is restricted to pure tree structures, strictly within the bounds of the topological fields. Thus, discontinuous constituents are marked via specific edge labels, e.g. OA-MOD for a discontinuous modifier of an accusative object. In the sentence in Figure 3, `Für diese Behauptung hat Beckmeyer bisher keinen Nachweis geliefert.` ("For this claim, Beckmeyer has not provided evidence yet."), the extraposed constituent `Für diese Behauptung` modifies the accusative object in the middle field, `keinen Nachweis`. The relationship between these constituents is marked by the label OA-MOD. The sentence in Figure 1 contains a discontinuous modifier of a prepositional object, labeled OPP-MOD.

3.2. Label Sets and Phrase Internal Structure

Further differences concern the set of node and edge labels and the internal structure of phrases: TüBa-D/Z defines 25 node labels for syntactic categories and 36 edge labels for grammatical functions covering head and non-head information, as well as subcategorization for complements and modifiers. TIGER utilizes 25 node labels and 50 edge labels. Apart from commonly accepted grammatical functions, such as SB (subject) or OA (accusative object), the TIGER edge labels also comprise an extended notion of grammatical functions, e.g. RE (repeated element) or RS (reported speech). Within phrases, TIGER adopts flat structures due to direct attachment of premodifiers and postmodifiers to the phrase node. Additionally, the flat structure is achieved by the notion of noun kernel elements: pronominal, nominal, and adjectival elements. Thus, there is a tendency to describe phrase internal structures via functional labeling. In the sentence in Figure 4, `Die Kritik der`

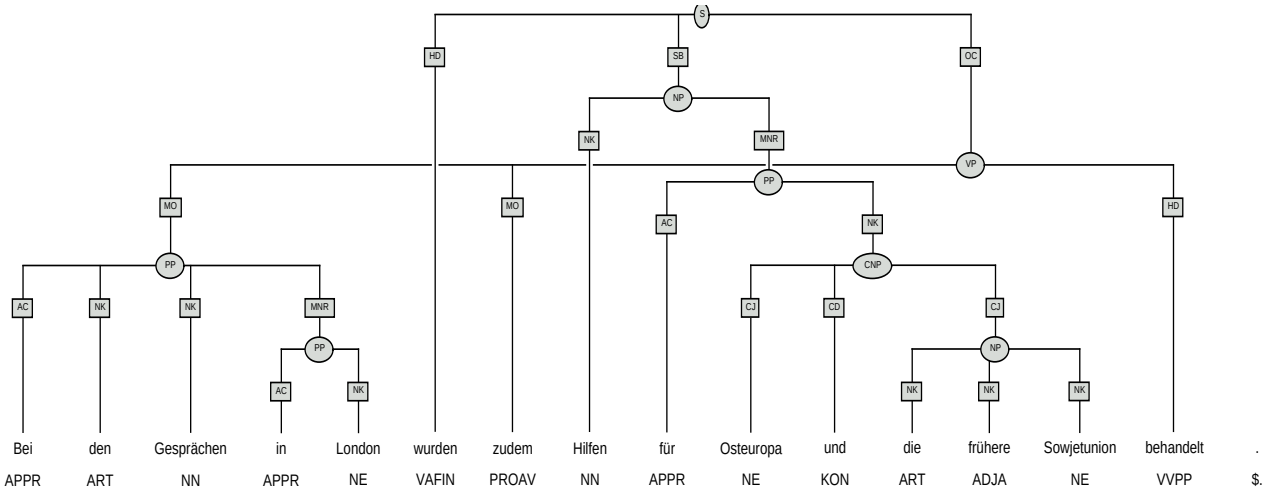

Figure 2: A TIGER tree with a long-distance relationship.

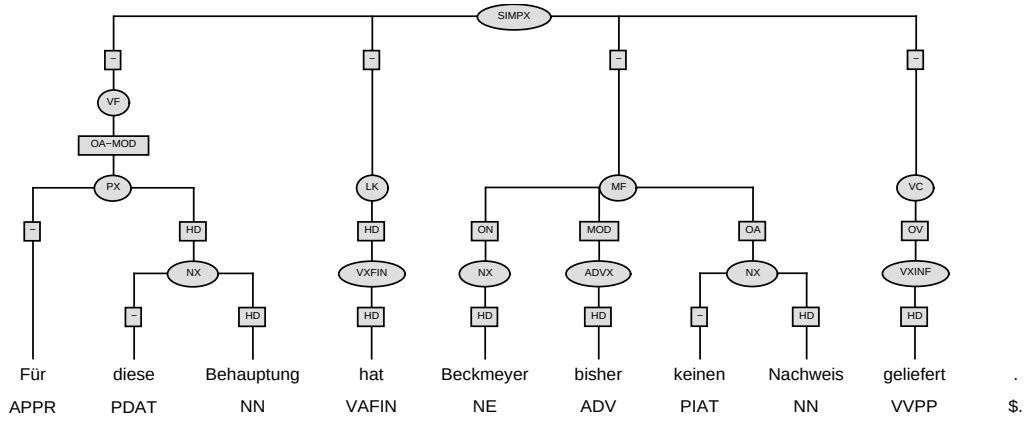

Figure 3: A TüBa-D/Z tree with a long-distance relationship.

`Ökonomenzunft an Perot entzündet sich vor allem an dessen vagen Äußerungen zur Wirtschaftspolitik.` ("The criticism of the guild of economists towards Perot is mainly kindled by his vage comments concerning economic policies."), the prepositional phrase contains the preposition as well as the premodifier (MO) of the PP and the elements of the included noun phrase on the same level. The latter elements are marked as belonging to the noun kernel (NK) or modifying it (OP), which replaces an explicitly marked noun phrase constituent.

By contrast, phrases in TüBa-D/Z are more hierarchically structured, thus representing this type of information structurally. The noun phrase in the sentence in Figure 5, `Allein die nicht mehr benötigte Anzuchtgärtnerei am Rande des Parks solle bebaut werden.` ("Only the seedling nursery, which is no longer in use, on the border of the park is supposed to be developed."), is structured so that the postmodifying prepositional phrase (PX) is attached high while the premodifiers, such as the adjectival phrase (ADJX), are attached low.

Despite the above mentioned differences, there is a multitude of similarities in the annotation schemes of the two treebanks: both use the STTS (Schiller et al., 1995) for POS-tagging, both make a distinction between pre- and postmodifiers in phrases, and both annotate named entities and elliptical structures.

In fact, it is precisely the similarities and differences between the two treebanks that make them a valuable resource for one and the same language. Depending on the intended application and depending on the available NLP tools, one treebank may be more feasible than the other. For example, due to its context-free backbone, TüBa-D/Z trees can be directly used for the training of PCFG-grammars while TIGER trees need to be converted first to pure tree structures by eliminating all crossing branches, which occur in approximately 30% of all trees.

4. Acknowledgements

This research was conducted jointly as part of the Sonderforschungsbereich 441 *Linguistic Data Structures* funded by the German Resaerch Foundation and as part of the *Kompetenzzentrum für Text- und Informationstechnologie* funded by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg, Germany.

5. References

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith, 2002. The TIGER treebank.

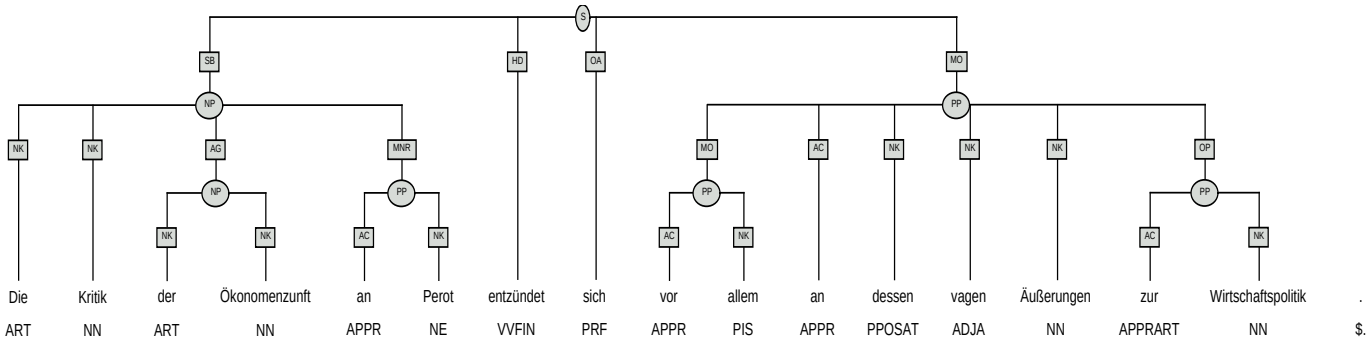

Figure 4: A TIGER tree with a complex noun phrase.

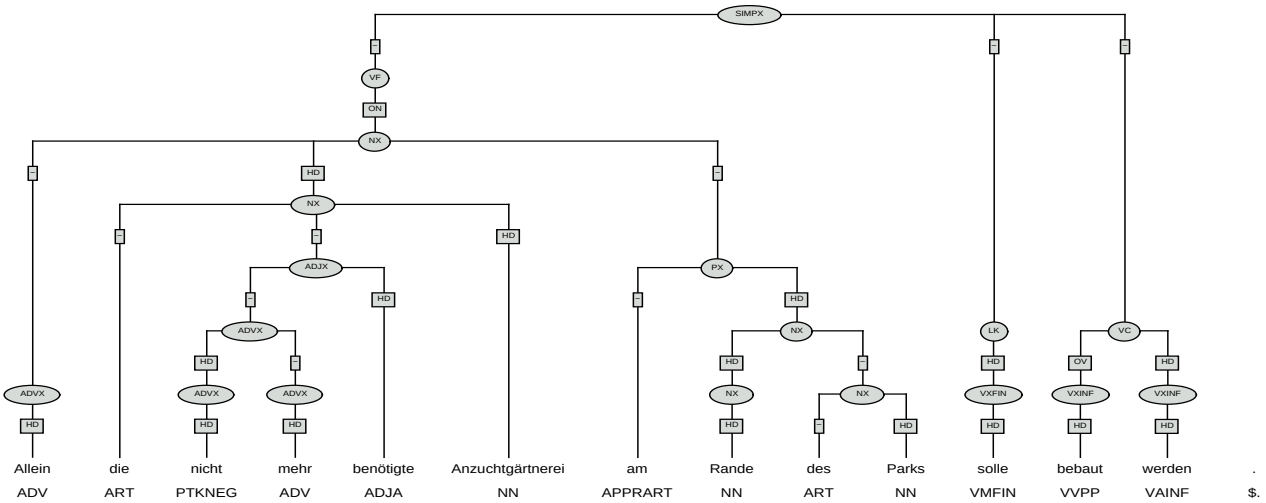

Figure 5: A TüBa-D/Z tree with a complex noun phrase.

In Erhard Hinrichs and Kiril Simov (eds.), *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*. Sozopol, Bulgaria.

Brants, Thorsten and Wojciech Skut, 1998. Automation of treebank annotation. In *Proceedings of NeMLaP-3/CoNLL98*. Sydney, Australia.

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

Hinrichs, Erhard W., Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann, 2000. The Tübingen treebanks for spoken German, English, and Japanese. In Wolfgang Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer.

Hinrichs, Erhard W. and Julia S. Trushkina, 2002. Forging agreement: Morphological disambiguation of noun phrases. In Erhard Hinrichs and Kiril Simov (eds.), *Proceedings of the First Workshop on Treebanks and Linguistic Theory (TLT 2002)*. Sozopol, Bulgaria.

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

Liepert, Martina, 2003. Topological fields chunking for German with SVMs: Optimizing SVM-parameters with GAs. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2003*. Borovets, Bulgaria.

Müller, Frank Henrik and Tylman Ule, 2002. Annotating topological fields and chunks–and revising POS tags at the same time. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan.

Plaehn, Oliver, 1998. *Annotate Bedienungsanleitung*. Universität des Saarlandes, Sonderforschungsbereich 378, Projekt C3, Saarbrücken, Germany.

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

Telljohann, Heike, Erhard W. Hinrichs, and Sandra Kübler, 2003. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

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

Veenstra, Jorn, Frank Henrik Müller, and Tylman Ule, 2002. Topological fields chunking for German. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL 2002)*. Taipei, Taiwan.