German Treebanks: TIGER and TüBa-D/Z

Stefanie Dipper and Sandra Kübler

Abstract German is a language that is closely related to English but has a richer morphology and freer word order than English. Additionally, German has four existing major treebanks, which differ considerably in their syntactic annotation schemes. All treebanks use a combination of constituent structure and grammatical functions, but the decisions with regard to other phenomena differ significantly, for example in the treatment of discontinuous structures. This makes German a good choice for a comparative analysis of treebanks. This chapter presents two major treebanks of German, TIGER and TüBa-D/Z. We describe the projects in which the two treebanks were annotated, discuss the respective annotation schemes, the processes used for annotation, and the data formats. We also discuss the usage of both treebanks, as well as other German treebanks, and we present a comparison of the two annotation schemes along with their advantages and disadvantages.

1 Introduction

German is an interesting language with regard to treebanks, for different reasons: On the one hand, it is a language that is closely related to English but has a richer morphology and freer word order than English. On the other hand, German is one of the very few languages for which more than one treebank exists, and the existing treebanks differ considerably in their syntactic annotation scheme.

Stefanie Dipper
Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, 44780 Bochum, Germany; e-mail: dipper@linguistics.rub.de

Sandra Kübler
Indiana University, Department of Linguistics, Bloomington, IN 47405, USA; e-mail: skuebler@indiana.edu

* We would like to thank Heike Zinsmeister for insightful comments and for providing us with references, and we would like to thank the two anonymous reviewers for valuable comments.

This chapter presents the two major treebanks of German, TIGER [3] and TüBa-D/Z [86].[1] Both treebanks are based on predecessors, TIGER on NEGRA [10] and TüBa-D/Z on TüBa-D/S [35], a treebank based on spontaneous dialogs (for more information see the following sections). Both TIGER and TüBa-D/Z are based on newspaper data, and both annotation schemes claim to be "theory-neutral". This means that the schemes refer to categories and structures that are widely used in syntactic theories of German. It also means that the schemes result from pragmatic mixtures of different approaches, combining their advantages. However, the resulting annotation schemes differ significantly, as shown in Sect. 2. This situation allows for a comparison of how different decisions made in treebank annotation impact later applications (see Sect. 5 for more details). For parsing, for example, first results show that there are significant differences in parsing quality between the two treebanks and that the standard evaluation metric is biased towards trees with a high number of nodes per word.

German syntax. In contrast to English, German has a case system of four cases: nominative, genitive, dative, and accusative (see Ex. (1a)). The assignment of grammatical functions is closely related to the case of a phrase: Subjects ('sbj') are in the nominative, direct objects ('dobj') in the accusative, and indirect objects ('iobj') in the dative case. Prepositions generally subcategorize for a specific case. This case system allows for a freer word order than in English. While the order inside phrases is fixed, the ordering of phrases is freer. Only the placement of verbs is fixed: In a main clause, the finite verb is in second (constituent) position, and all other verbal elements are clause-final. In a subordinate clause, all verbal elements are placed in final position. In the example in (1), all six possible orderings of the noun phrases are possible, with differences in information structure.

(1) a. [$_{\text{NP}_{sbj}}$ Der Arzt] hat [$_{\text{NP}_{iobj}}$ dem Patienten] [$_{\text{NP}_{dobj}}$ die Pille] gegeben.
The$_{nom}$ doctor has the$_{dat}$ patient the$_{acc}$ pill given.
(Eng.: The doctor gave the patient the pill.)
b. Der Arzt hat die Pille dem Patienten gegeben.
c. Die Pille hat dem Patienten der Arzt gegeben.
d. Die Pille hat der Arzt dem Patienten gegeben.
e. Dem Patienten hat der Arzt die Pille gegeben.
f. Dem Patienten hat die Pille der Arzt gegeben.

The fixed placement of the verbal elements in a clause lends itself to an analysis into *topological fields* [23, 24]. Example (2) shows a sentence with topological fields: VF is the initial field, LK the left bracket, MF the middle field, VC the final verb complex, and C the complementizer field in a subordinate clause. Topological fields are explicitly used in one of the major treebanks in German, TüBa-D/Z (cf. Sect. 2.2 for a description of the different fields).

[1] Project websites are available at http://www.ims.uni-stuttgart.de/forschung/projekte/tiger.html (TIGER) and http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html (TüBa-D/Z). All URLs provided in this paper have been accessed Dec 18, 2013.

(2) [$_{VF}$ Es] [$_{LK}$ ist] [$_{MF}$ schon kurios], [$_{C}$ was] [$_{MF}$ sich derzeit beim Fussball-Zweitligisten FC St. Pauli] [$_{VC}$ abspielt].
(Eng.: It is rather strange what is happening with the second league soccer team FC St. Pauli.)

The remainder of this chapter is structured as follows: In the following, we will provide a short description of the two projects in which TIGER and TüBa-D/Z were created. Then, in Sect. 2, we give an overview of the annotation schemes used in TIGER and TüBa-D/Z. Section 3 describes how both treebanks were annotated, and Sect. 4 details the physical representation of the two treebanks. In Sect. 5, we describe in which ways TIGER and TüBa-D/Z have been used, and Sect. 6 gives a short list of other treebanks for German.

1.1 The TIGER Project

The TIGER project, funded by the German Research Foundation (DFG), ran from 1999–2004. Its original goal was to extend the NEGRA corpus [10] both in size and detail of annotation. TIGER finally ended up as an independent corpus, sharing the basic annotation scheme with NEGRA but using different sets of texts. Due to this genesis, the description of TIGER also refers to the NEGRA project and corpus.

The NEGRA corpus was created by project C3: *NEGRA: Concurrent Grammar Processing* of the collaborative research center SFB 378, *Resource-Adaptive Cognitive Processes* at Saarland University. Project C3 ran from 1996–2001 and focused on combining constraint-based systems and robust statistical processing techniques. Among the outputs of the project was the first German treebank, the NEGRA corpus. Release 2 contains 350,000 tokens (20,000 sentences). The annotation scheme was designed as theory-neutral as possible, combining advantages of phrase-structure grammar and dependency grammar. Specific features were rather flat hierarchies and crossing branches, which encode discontinuous relationships (see Sect. 2.1 for more details).

The TIGER project was a joint initiative of the Department of Computational Linguistics and Phonetics at Saarland University, the Institute for Natural Language Processing (IMS) at the University of Stuttgart, and the Department of German Studies at the University of Potsdam. The project worked on different aspects of treebanking: It extended the NEGRA annotation scheme, experimented with alternative annotation methods, and created a search tool (*TIGERSearch* [54]) and an XML-based exchange format (*TIGER-XML* [56]). The TIGER annotation scheme adds lemma and morphological information and provides additional fine-grained distinctions at the level of grammatical functions and a new device called ‘secondary edges’, to encode shared constituents in coordinations and ellipses.[2]

[2] Secondary edges were already proposed in the context of the NEGRA project [80] but had not been used in the actual annotation of the NEGRA corpus.

The textual basis of TIGER is the newspaper 'Frankfurter Rundschau', covering two complete weeks from November 1995,[3] as well as further articles from selected days, e.g. one day from each month of 1997. Regional and sports news were excluded because they often contain tables and enumerations rather than complete sentences.

The first release of TIGER, published in July 2003, contained about 700,000 tokens (40,000 sentences). It was annotated with part of speech (POS) tags and syntactic trees with grammatical functions. It also contained corrections for misspelled words and meta-information (domain, date) about most of the articles. Release 2, published in December 2005, contained almost 900,000 tokens (50,000 sentences) and was further enriched with inflectional morphology and lemma annotation. Misspelled words were replaced by their corrected version in this release. In Release 2.1 (August 2007), morphological features were additionally split into their atomic parts (e.g. the complex value `morph="Nom.Sg.Masc"` became `case="Nom" number="Sg" gender="Masc"`). The current release, 2.2, published in July 2012, is a cleaned-up version of release 2.1. Release 2.2 is also provided with CoNLL-2009 dependency trees that have been derived automatically from the tree annotations. The TIGER treebank and the search tool TIGERSearch are hosted by the CLARIN-D center at the IMS Stuttgart.

The annotation levels are documented in different guidelines: POS and morphological annotation uses the *Stuttgart-Tübingen Tagset* (STTS) [88, 76], morphological and lemma annotations are further documented in [18]. Finally, there are extensive guidelines for syntactic annotation [1]. The presentation in this chapter focuses on the syntactic layer.

1.2 The TüBa-D/Z Project

The TüBa-D/Z[4] project is an ongoing project that started in 1999 at the Department of Linguistics at the University of Tübingen. The project started as an extension of the TüBa-D/S treebank [35, 36], which was developed in the *Verbmobil* project [94]. Verbmobil was a large-scale project on speech-to-speech machine translation for the languages German, English, and Japanese, specialized for the domain of scheduling business meetings. For all three languages, treebanks of the recorded and transcribed dialogues were created. The German Verbmobil treebank (TüBa-D/S) was based on a theory-neutral annotation scheme, with the restriction that the annota-

[3] This period was chosen because it covers a globally relevant event: the assassination of Israeli Prime Minister Yitzhak Rabin. The idea was to keep the option open of building a multilingual corpus, because it would be rather easy to find news about this event in many different languages. A drawback is that the there is some overlap in content among the articles of the two weeks.
The NEGRA corpus also consists of texts from 'Frankfurter Rundschau', from 1991 and 1992. As far as we know, there is no overlap in texts between the NEGRA and TIGER corpora.

[4] TüBa-D/Z is short for 'Tübinger Baumbank des Deutschen / Zeitungssprache' (Tübingen Treebank of German / Newspaper), i.e., the Z denotes newspaper texts while the S in TüBa-D/S denotes spontaneous speech.

tions should not contain any crossing branches, traces, or empty categories. This annotation scheme had to be adapted for the use in the TüBa-D/Z treebank since the TüBa-D/Z is based on written language, which covers complex phenomena that did not occur in TüBa-D/S (see below for details). Over the years, the TüBa-D/Z project was funded by different funding sources, including the *Competence Center for Text- and Information Technology* (KIT), the collaborative research center SFB 441, project A1: *Representation and Automatic Acquisition of Linguistic Data*, the collaborative research center SFB 833, project A3: *Disambiguating Discourse Connectives using Corpus-induced Semantic Relations*, and the ESFRI research infrastructure projects D-SPIN and CLARIN-D.[5]

TüBa-D/Z has been released incrementally; the current release is no. 9.1, and it covers 85,358 sentences (which is equivalent to 1,569,916 tokens or 3,444 newspaper articles). TüBa-D/Z has the newspaper 'die tageszeitung' (taz) as its textual basis. The first part covers complete days from July 1992, October 1995, and April and May 1999, the sentences for later parts were taken from individual articles from the years 1989 and 1997.

In the first release of TüBa-D/Z, which contained 15,000 sentences, the treebank contained annotations for the following linguistic levels: POS annotation, syntactic constituent annotation enriched by grammatical functions and head/non-head annotation, topological fields, and named entities. This release also contained corrections of misspelled words. In later releases, the following layers of annotation were added for all sentences: inflectional morphology, lemma annotation, anaphora and coreference, automatically generated dependency annotations (converted from constituents), and automatically converted chunk annotations. Additionally, there are partial annotations available for selected discourse particles, such as *nachdem* (after) or *seitdem* (since), as well as for explicit and implicit discourse relations. The latest release added word sense annotations for 30 nouns and 79 verbs, based on *GermaNet* [53] senses. The syntactic annotation is documented in an extensive stylebook, which was updated along with most releases; the latest version is from 2012 [87]. The annotation of anaphora and coreference is documented in its own set of guidelines [60]. The same holds for the discourse connectives [78]. The chunk annotation [48] and discourse connectives [29] are described in workshop proceedings. In the following sections, we will concentrate on the annotations of syntactic constituents and topological fields.

[5] For more information on these projects, see http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html.

2 Annotation Scheme

2.1 The TIGER Annotation Scheme

As mentioned above, the TIGER annotation scheme is an extension of the scheme that has been developed in the NEGRA project. The designers of the NEGRA scheme made the following assumptions [11, 10, 79]:

- The annotations should be theory-neutral, and sufficiently detailed as to permit the extraction of theory-specific representations.
- In purely constituency-based representations, non-local relationships (e.g. topicalization, extraposition) result in rather non-transparent structures. Hence, dependency-based representations seem preferable.
- In purely dependency-based representations, constructions without a clear syntactic head (e.g. ellipses, coordinations) are difficult to analyze. Hence, constituency-based representations seem preferable.
- Use of flat structures reduces the number of possible attachment sites, promoting consistent annotation.

The NEGRA scheme therefore opted for a hybrid approach, combining the advantages of constituents and dependency relations. Figure 1 shows an example sentence from the TIGER corpus. In the structure, phrasal nodes are displayed in circles, and grammatical and other functions in grey boxes, as edge labels. The terminal nodes show the surface tokens along with POS information according to the *Stuttgart-Tübingen Tagset* (STTS).

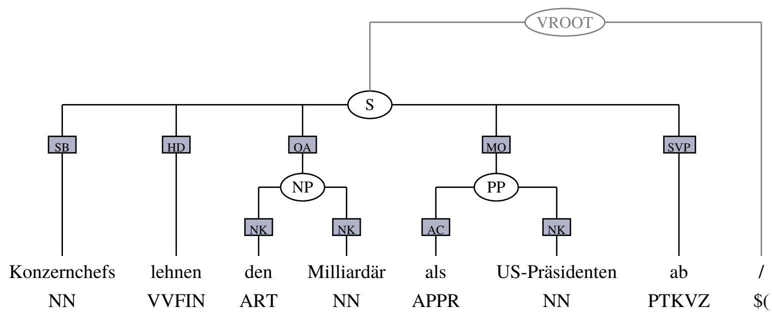


Fig. 1 The sentence *Konzernchefs lehnen den Milliardär als US-Präsidenten ab /* (Eng.: CEOs reject the billionaire as US president /) from the TIGER treebank

Flat structures. NEGRA constituents are flat, directly dominating functional and lexical heads. For instance, both the definite article and the noun of *den Milliardär* (Eng.: the billionaire) are directly dominated by an NP node. Both function as NK ('noun kernel'); the intention behind that decision is to leave open the question which one is the head. Similarly, the PP node of *als US-Präsidenten* (Eng.: as US president) has minimal internal structure. The preposition is analyzed as a kind of case marker (AC, 'adpositional case marker'), the noun again is assigned the function NK. The guiding idea is that users of the treebank can construct their preferred NP and PP analyses by combining information from the POS tags and grammatical functions.

Furthermore, unary (non-branching) nodes are omitted. For instance, there are no NPs nodes that dominate one word only (e.g. the head noun or a pronoun), see the noun *Konzernchefs* (Eng.: CEOs) in Fig. 1. Again, the fact that this is an NP can be recovered by referring to the POS tag (NN, 'normal noun') and its grammatical function (SB, 'subject')—if it was part of a complex NP, it would have been assigned the function NK.

The finite verb of the sentence functions as the head (HD). Besides the subject, there is an accusative object (OA) and a modifier (MO). The final word *ab* is a separated verb particle (SVP).

Crossing branches. Figure 2 illustrates further properties of the annotation scheme. For encoding non-local dependencies, it uses crossing branches. For instance, the discontinuous sequence *so . . . wie* (Eng.: as . . . as) belongs to the same adverbial node (AVP). The first element (*so*) is the head of the phrase, the second element is the comparative complement (CC), which is headed by the comparative conjunction (CM) *wie*.

The figures also show that punctuation marks are not integrated in the actual syntactic analysis. Instead, they are all attached to a virtual root node (VROOT).

TIGER extensions of the NEGRA scheme. Figure 2 also illustrates one of the TIGER-specific extensions. The pointer from the head verb *scheint* (Eng.: seems) to the second sentential conjunct is called 'secondary edge'. It encodes the information that this verb is the head not only of the first conjunct but also of the second, elliptical conjunct.

Further TIGER-specific extensions of the original NEGRA annotation scheme concern additional labels for grammatical functions:

- TIGER distinguishes between PP arguments (prepositional objects, OP) and PP modifiers (MO), e.g. as in ***auf*** *jemanden warten* (Eng.: to wait **for** somebody; OP) vs. ***am/im/beim*** *Bahnhof warten* (Eng.: to wait **at/in/near** the station; MO). Tests for identifying PP arguments are: The preposition is morphologically simple and semantically empty. It is selected by the governing head (e.g. a verb) and cannot be replaced by another preposition without a clear change in meaning.
- Another newly introduced label is used for collocational verb constructions (CVC). In these V+PP-constructions, the verb is semantically weakened, and the main content is provided by the PP's noun. Example phrases are *zur Geltung kommen* (Eng.: be recognized; literally: to come into appreciation), or *zur Verfügung stehen* (Eng.: be available; literally: to stand at the disposal).
- TIGER provides three labels for non-referential occurrences of *es* (Eng.: it):

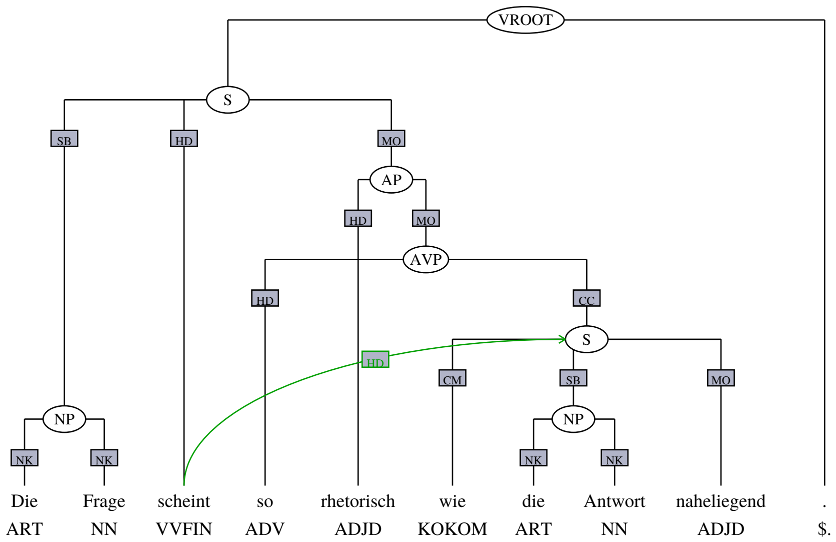


Fig. 2 The sentence *Die Frage scheint so rhetorisch wie die Antwort naheliegend.* (Eng.: The answer seems as rhetorical as the answer (seems) straightforward.) from the TIGER treebank

– *Es* which serves to fill the initial field is annotated as a placeholder (PH), as in ***Es*** *herrschte der kalte Krieg* (Eng.: The Cold War was underway).
– *Es* (PH) can also be correlated to some propositional argument, called repeated or resumptive element (RE), as in *Sie lehnen* ***es*** *ab, dass* ... (Eng.: They refuse that ...).
– Expletive *es* (EP) functions as a non-thematic argument, as in *Heute regnet* ***es*** (Eng.: Today, it is raining).

The TIGER extensions first of all aim at improving the representation of valency. Secondary edges "copy" missing constituents to elliptical constructions. Similarly, fine-grained labels for PPs and expletives support extraction of head–argument–modifier relations.

Second, these constructions (ellipses and expletives) are phenomena that are widely discussed in theoretical linguistics. Many of them would be difficult to locate in the corpus if they were not marked by specific labels and edges.

2.2 The TüBa-D/Z Annotation Scheme

The syntactic annotation scheme for the TüBa-D/Z treebank consists of a combination of surface-oriented constituent structure and topological fields, enriched by predicate-argument structure. The annotation scheme is based on the following principles:

- The *flat clustering principle* keeps the number of hierarchy levels in the constituent structure as low as possible. Thus, any degree of branching is allowed.
- The *longest match principle* requires that as many daughters as possible are grouped into a single mother node, provided that the resulting construction is syntactically and semantically well-formed.
- The *high attachment principle* is used in cases of ambiguity. It specifies that ambiguous constituents are grouped under the highest possible mother node.

The label sets are chosen so that they are based on minimal assumptions that are acceptable for most major syntactic theories. Figure 3 shows an example of a sentence with its syntactic annotation.

The figure shows a sentence with its POS tags, its constituent structure, topological fields, and its grammatical functions. Like in NEGRA and TIGER, the POS tags are based on the STTS [76, 88]. Topological fields [41] are used as the major structuring principle of clauses; they are located directly below the clause level, i.e., below any SIMPX node (or R-SIMPX in case of relative clauses). Thus the main clause in Fig. 3 is divided into an initial field (VF), the left sentence bracket (LK), containing the finite verb, the middle field (MF), and the final field (NF), which covers extraposed material.

Grammatical functions are annotated as edge labels between the maximal phrases and topological fields. Thus, the first NX in the main clause is annotated as a verb modifier (V-MOD), the finite verb in VXFIN is the head HD of the sentence, the middle field contains the subject (ON), two modifiers, and the predicate, and the final field contains a modifier of the subject (ON-MOD). Following Reis [70], the annotation scheme uses grammatical functions based on case rather than distribution. I.e., the subject is marked as nominative object (ON), the other arguments being genitive object (OG), dative object (OD), and accusative object (OA). On the phrase level, predicate-argument structure is annotated in terms of heads (HD) and non-heads (-). Thus, in the noun phrase (NX) *die 220 Albaner*, the noun (NN) constitutes the head, and the determiner (ART) and the adjectival phrase (ADJX) non-heads. The phrase labels ending in X are remnants of an original decision to annotate chunks rather than phrases, which was revised before the first release of TüBa-D/S.

Non-local phenomena. The above mentioned surface orientation of the annotation scheme resulted in a decision not to annotate crossing branches, traces, or empty categories. Thus, TüBa-D/Z trees are mostly pure tree structures; however trees do not have to be be fully connected to a spanning tree. Long-distance phenomena are handled via an extended set of grammatical functions in combination with secondary edges. The grammatical functions specify which maximal constituent is modified. For example, the sentence in Fig. 3 exhibits an extraposed noun phrase (NX), which is grouped under the final field, and the grammatical function label ON-MOD specifies that it modifies the subject. The modified phrase is always in the same clause, but can be found either in the initial field or in the middle field.

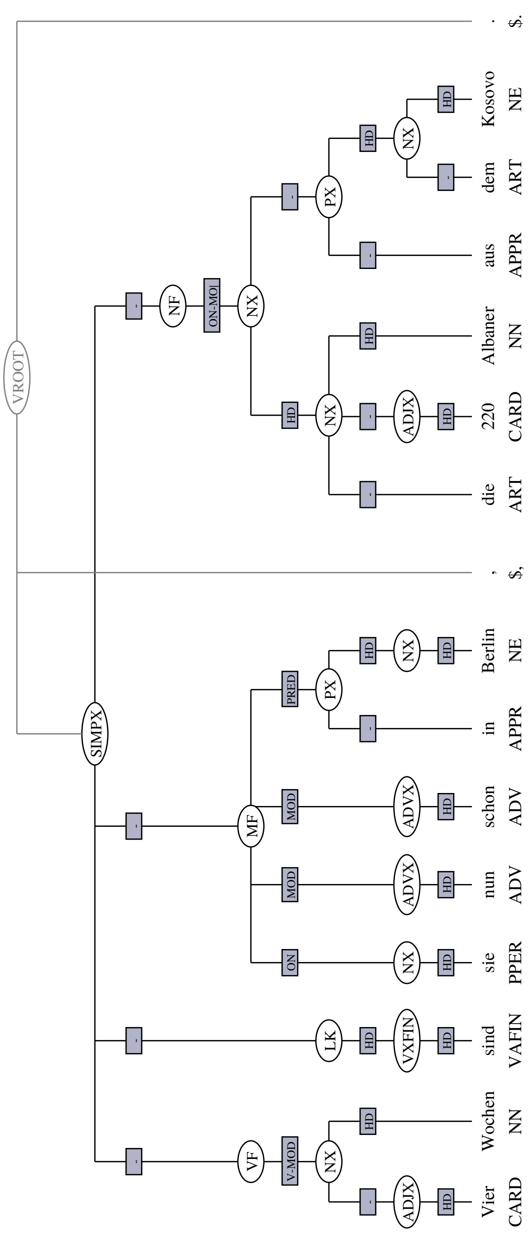


Fig. 3 The sentence *Vier Wochen sind sie nun schon in Berlin, die 220 Albaner aus dem Kosovo.* (Eng.: For four weeks, they have already been in Berlin, the 220 Albanians from the Kosovo.) from the TüBa-D/Z treebank

This definition may be underspecified, especially in cases where the extraposed constituent modifies another modifier (MOD-MOD). To handle such cases, secondary edges are used. These edges are not part of the proper tree but represent additional information; they are used for different purposes than in TIGER, to annotate headedness in verb complexes with complex internal structures, extraposition (see below), ambiguous modification, and control verb constructions. In cases where the extraposed constituent does not modify a maximal but an embedded phrase, the grammatical function refers to the maximal phrase, and the additional secondary edge connects it to the constituent that it modifies.

Figure 3 also shows that punctuation signs are not attached to any constituent, and can thus be considered to be attached to a virtual root (VROOT, as shown in the figure), parallel to the treatment of punctuation in TIGER. The reason for this is that a single punctuation sign often performs more than one function, and it is therefore often difficult to decide where to attach them. Other cases in which no single spanning tree is annotated include paratactic constructions and parentheticals. An example of the latter is shown in Fig. 4. In such cases, all sentences are projected to the SIMPX level but not grouped under a common node. In the example shown, the interjection is surrounded by direct speech[6].

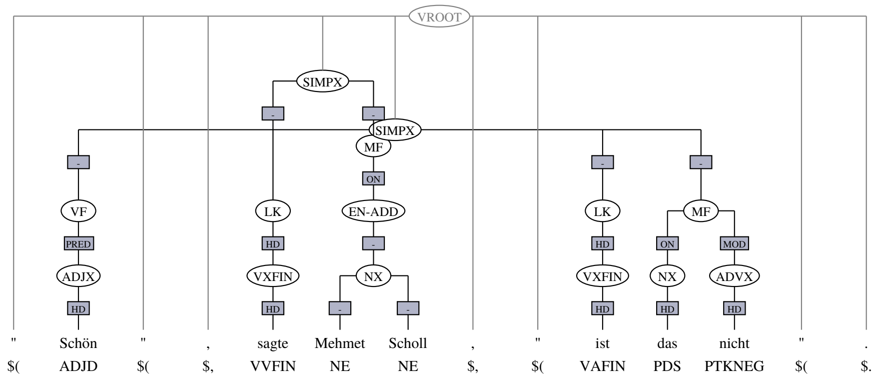


Fig. 4 The sentence *"Schön", sagte Mehmet Scholl, "ist das nicht"*. (Eng.: "Great", said Mehmet Scholl, "it is not".) from the TüBa-D/Z treebank, which has a parenthetical sentence. Note that the parenthetical directly attaches to the virtual root (VROOT), but is a separate tree, which in the graphical representation accidentally overlaps with the surrounding sentence

TüBa-D/Z extensions. The annotation scheme was originally developed for use in the TüBa-D/S treebank of spoken German. It then underwent minor adaptations to cover phenomena of written language that did not occur in the spoken data. One

[6] Note that we only have a parenthetical construction if the matrix clause is embedded into the direct speech. If the parenthetical were annotated as the head of the direct speech, this would result in a crossing branch, which is not an option in the TüBa-D/Z annotation scheme.

phenomenon in this category concerns presumptive constructions, as in (3a). These are annotated as grouped under a field LV (left dislocation), which is located to the left of the initial field. Another phenomenon concerns split coordinations, as in (3b), which necessitated the introduction of specific labels, such as OAK for an extraposed conjunct of the direct object.

(3) a. Doch wie es weitergehen soll, da herrscht kein Konsens.
(Eng.: But how it is supposed to continue, there is no consensus.)
b. 450 verschiedene Gehölze haben die Biologen registriert, 100 Vogel- und 35 Säugetierarten.
(Eng.: 450 different woods the biologists have recorded, 100 types of birds and 35 of mammals.)

However, the fact that the annotation scheme could be used with only minor modifications for spoken as well as written language can be taken as an indication of the robustness of the annotation scheme.

2.3 Comparison of the Two Schemes

TIGER and TüBa-D/Z differ in a range of decisions that were made in the annotation schemes. Here, we will discuss the major differences between the two annotation schemes, including the advantages and disadvantages of the individual decisions.

Crossing branches. Since German is a morphologically rich language with a case system, it exhibits a considerable amount of non-linear phenomena including fronting and extraposition. In TIGER, such phenomena are annotated via crossing branches while TüBa-D/Z uses a strict tree structure in combination with specific functional labels, for example OA-MOD for an extraposed modifier of the direct object (OA). The crossing branches in TIGER are easy to annotate since they group constituents that belong together. However, this makes it difficult to determine the linear order of constituents when searching. For example, in a search for an NP_1 which precedes an NP_2, linear precedence is not easily determined if NP_1 is modified by an extraposed relative clause which follows NP_2. Also, crossing constituents mean that standard parsing algorithms based on context-free grammars cannot be used directly. In order to parse such tree structures, either more powerful parsing algorithms [42, 65] have to be used, or the crossing branches must be resolved, e.g. [50], which requires a non-obvious mapping that changes the linguistic content of the tree.

The solution in TüBa-D/Z is a good fit for standard (context-free) parsing algorithms since mostly, a strict tree structure is preserved. However, since the grammatical function label only points to the maximal constituent, cases in which the extraposed material does not modify the full constituent are underspecified in the pure tree structure. An example of such a modification is shown in Fig. 5. In this sentence, the extraposed relative clause labeled R-SIMPX modifies the noun phrase *der Erben Melchiors* (Eng.: of the heirs of Melchior), not the whole direct object (OA).

This is shown by the secondary edge from the noun phrase to the relative clause. Additionally, the sets of constituent and grammatical function labels in TüBa-D/Z are considerably more extensive than in TIGER, which can present challenges to parsers.

Flat vs. hierarchical structure. TIGER uses a very flat structure inside noun phrases and does not annotate unary constituents, see Fig. 6. TüBa-D/Z, in contrast, employs a more hierarchical structure, see the direct object in Fig. 5. For TIGER, this means that the trees overall are very flat so that annotation is easier because less structure needs to be created, and more of the tree structure is visible at any given time. However, this also means that certain generalizations are left implicit or even underspecified, and need to be searched for via (heuristic) templates (see Sect. 5). For example, pronouns are not marked as noun phrases since such an NP would be unary. Here, the more explicit structure in TüBa-D/Z allows for more general queries.

Information in the trees. TIGER and TüBa-D/Z differ considerably in what types of information are integrated into the syntactic annotation. While TIGER focuses on morphological and morpho-syntactic annotations, TüBa-D/Z also integrates topological fields and named entity information in the trees. On the one hand, this allows for easier searches that combine these types of information with syntactic information. Thus, it is possible to easily search for subjects that are not in first position, i.e., not in the initial field (VF). Such a query will find sentences such as the ones shown in (4). However, this decision also means that different types of information are integrated into the tree, and it is not always obvious how to distinguish between them: Topological fields are nodes like any other syntactic constituent. Named entities were originally also annotated as individual nodes, but they were moved to syntactic nodes in release 8 and now are shown in a complex form, e.g., ADVX=ORG for an adverbial phrase, which is a named entity of the semantic class ‘organisation’.

(4) a. In einer anonymen Anzeige werden der Bremer Staatsanwaltschaft Details über dubiose finanzielle Transaktionen mitgeteilt.
(Eng.: In an anonymous note, the Bremen Public Attorney’s Office is told about shady financial transactions.)

b. Kurz und gut – irgendwann muss auch Andy Kreiter Urlaub vom Affenschinden machen und dann stehe ich mit Herzenswärme und Bananen als Urlaubsvertretung bereit.
(Eng.: Long story short – at some point, Andy Kreiter also has to take a break from monkey flaying, and then I will stand by as vacation replacement with a sympathetic heart and bananas.)

Queries relating to topological fields, such as subjects in positions other than the VF, can be approximated in the TIGER corpus using complex templates, see [22] and Sect. 5.

Table 1 provides some quantitative information of the two schemes and the treebanks. Note that not all numbers can be compared directly to each other since nodes

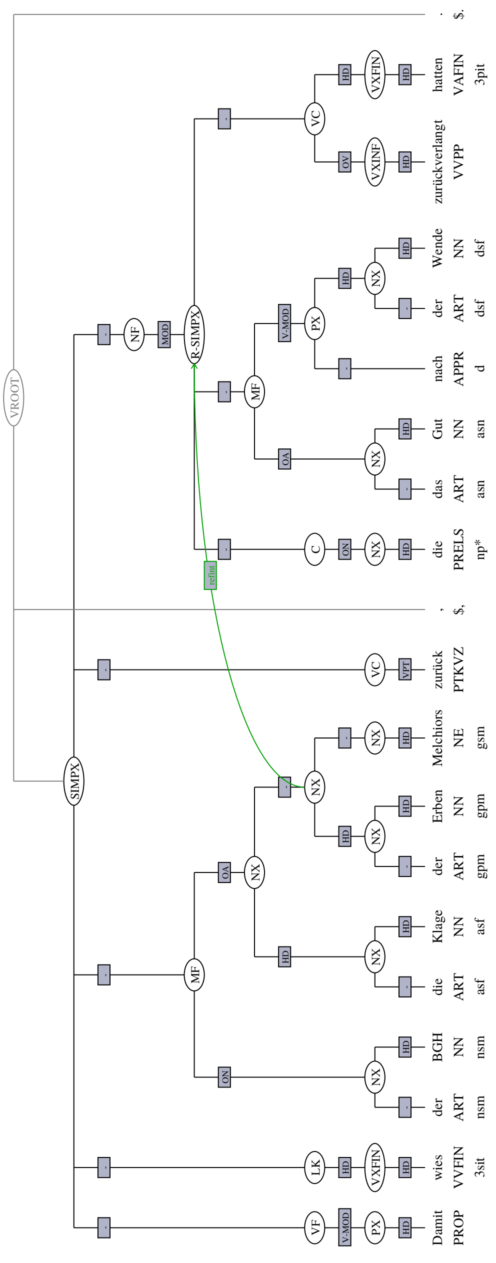


Fig. 5 The sentence *Damit wies der BGH die Klage der Erben Melchiors zurück, die das Gut nach der Wende zurückverlangen wollten.* (Eng.: Hereby the BGH turned the lawsuit of the heirs of Melchior down, who wanted to demand the property back after the reunification.) from the TüBa-D/Z treebank

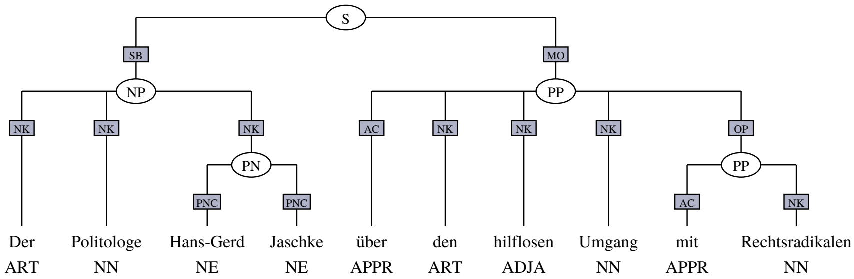


Fig. 6 The sentence *Der Politologe Hans-Gerd Jaschke über den hilflosen Umgang mit Rechtsradikalen* (Eng.: The political scientist Hans-Gerd Jaschke on the helpless handling of right-wing extremists) from the TIGER treebank

and edges encode different kinds of information in the two schemes, as described above.

Table 1 Quantitative information of the TIGER and TüBa-D/Z annotation schemes and treebanks

	TIGER scheme	TüBa-D/Z scheme
#POS tags (STTS)	53	53
#Morphology tags	584	133
#Node labels, syntactic categories	25	36
#Node labels, topological fields	–	13
#Edge labels	48	53
#Secondary-edge labels	48	4

	TIGER corpus (release 2.2)	TüBa-D/Z corpus (release 9.1)
#Words	768,677	1,340,258
#Punctuation marks	119,561	229,658
#Sentences	50,474	85,358
#Nodes, syntactic categories	373,831	1,342,924
#Nodes, topolological fields	–	506,935
#Edges (w/o punc. marks)	1,089,628	3,083,816
#Secondary edges	6,444	6,396

3 Annotation Process and Evaluation

3.1 The Annotation Process in TIGER

Large parts of the TIGER treebank were annotated by means of two semi-automatic tools, *Annotate* and *TigerMorph*. For a subset of sentences, a different path was followed: the sentences were parsed by a symbolic grammar. Both approaches are described in the following sections.

3.1.1 Annotation with *Annotate* and *TigerMorph*

As the very first step, the texts of the corpus were tokenized. The tokenized sentences were proofread once by the annotators.

For the annotation of POS tags and syntactic structures, the tool *Annotate* was used [9, 64]. This tool had been developed in the context of the NEGRA project, and since has been applied in a range of treebanking projects for German.[7]

The tool uses an SQL database to store annotations, and integrates a probabilistic POS tagger and parser. POS tagging is done by the tagger TnT [8]. The tagger marks whether the suggested tags are reliable. The parser is implemented as a cascade of Markov models [6]. Instead of generating the entire sentence structure in one step, the parser only generates one local subtree in each step, which is immediately checked by the human annotator, and modified if necessary. Based on the annotator's decision, the parser generates the next subtree, and so on. The advantage of this kind of interactive parsing is that the automatic parser can use the decisions made by the human annotator at lower levels. In this way, errors from the statistical parser do not propagate to higher levels, and can often be detected more easily since the annotator's focus is always on the node generated most recently. Another advantage of the interactive annotation process is that the annotator has to focus on sub-decisions rather than looking over a complete tree, which may disguise annotation errors. The tagger and parser are retrained at regular intervals. In an early evaluation on the NEGRA corpus, approximately 85% of the tags suggested by the TnT POS tagger were marked as reliable (and 99.2% of those were indeed correct) so that human annotators needed to proofread only 15% of the tags (which had an accuracy of 83.0%). Approximately 70% of the suggested phrases and 91% of the edge labels were correct [66].

Graphical user interface (GUI). Figure 7 shows a screenshot from the tool's GUI: Four nodes have been already annotated. Currently, the function of the highlighted node (PP) is being edited; see the field 'Edgelabel' in the bottom right corner, which is still set to 'not bound'. This means that the parser was not able to predict the PP's function. The figure also illustrates that non-local dependencies can be anno-

[7] Besides NEGRA, TIGER, TüBa-D/Z, and the Verbmobil treebanks, *Annotate* was also used for e.g. the Potsdam Commentary Corpus [84], Mercurius Treebank [19], Deutsche Diachrone Baumbank [40], and SMULTRON [93]. The tool is no longer maintained.

tated with the tool: the top AP dominates the topicalized phrase *zu abhängig* (Eng.: too dependent) and its PP argument *vom dort größten Arbeitgeber* (Eng.: from the locally largest employer).

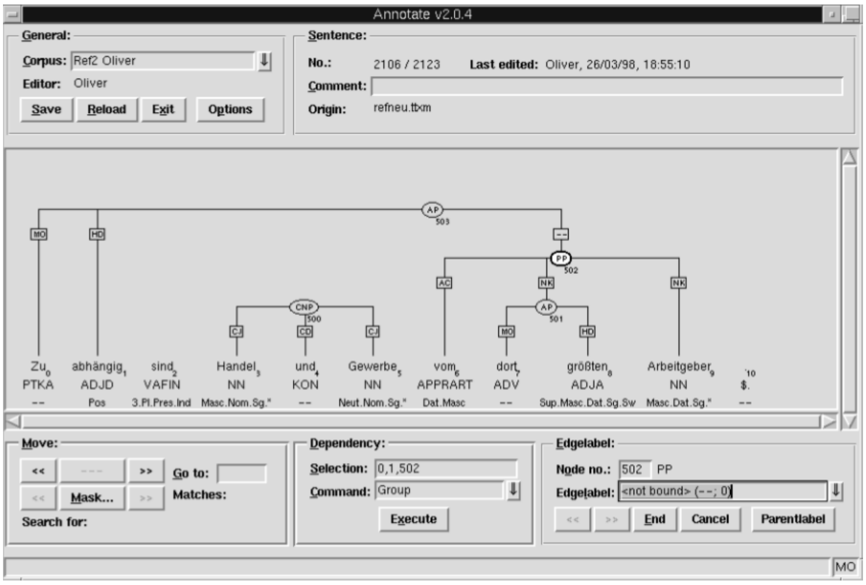


Fig. 7 The sentence *Zu abhängig sind Handel und Gewerbe vom dort größten Arbeitgeber.* (Eng.: Trade and commerce are too dependent on the locally largest employer.), in the course of being annotated by means of the tool *Annotate* (screenshot from [64])

Morphological and lemma information was added in a later stage of the project, using the tool *TigerMorph*.[8] It exploits syntactic information from the treebank (e.g. SB, OA, OD) to suggest disambiguated morphological tags (nominative, accusative, dative case).

The dependency version is created automatically via the script *Tiger2Dep* [77] from the original version of the treebank.

Annotators. The annotators were advanced undergraduate students and PhD students from German Linguistics and Computational Linguistics. Each sentence was annotated independently by two annotators, who afterwards compared their results and agreed on the final structure, using scripts that supported manual comparison and adjudication of the structures stored in the database. Difficult cases were collected and discussed in regular meetings. The TIGER treebank was annotated at three different sites: Saarbrücken, Stuttgart, and Potsdam. To ensure consistent annotation across the sites, certain parts of the treebanks were assigned to annotators

[8] *TigerMorph* was developed by Berthold Crysmann and was only used in the TIGER project. It is not available.

from different project sites, e.g. one annotator worked in Saarbrücken, the other in Stuttgart.

Twice a year, all annotators of the three sites came together for two days, and major decisions were made, such as introducing an extra label for PP arguments. At these occasions, other modifications of the annotation scheme were also decided, such as adding new tests and example sentences for difficult cases. The final version of the annotation guidelines is from 2003 and is almost 150 pages long [1]. The distinction between PP arguments and modifiers, which is often difficult to draw (and for this reason was not part of the original NEGRA scheme), is facilitated by comprehensive lists of verbs and their PP arguments or typical PP modifiers, and lists of verbs and PPs participating in collocational verb constructions.

On average, a single annotation of one sentence took about 50 sec. All steps taken together, the procedure resulted in about 10 minutes annotation time for each sentence. Inter-annotator agreement was first computed for the predecessor corpus NEGRA: Agreement for part-of-speech was 98.6%, the labeled F-score for structures was 92.4% [7]. In a following evaluation, TIGER edge labels were evaluated, resulting in an F-score of 93.89% [4].

3.1.2 Annotation with the LFG Grammar

Following a different path, parts of the corpus were parsed by a broad-coverage symbolic grammar [20], implemented in the framework of LFG (Lexical Functional Grammar [12]), using the Xerox Linguistic Environment (XLE) development platform [17]. The grammar has been developed at the University of Stuttgart, in the context of the project *Pargram* [14, 21].

An LFG grammar produces two types of output, a constituent structure and a functional structure (c- and f-structure for short). This resembles the hybrid approach taken in the TIGER annotation scheme, which mixes phrase structures with dependency structures. However, since the LFG grammar produces theory-specific structures, a range of modifications has to be applied to its output.

Figure 8 illustrates the commonalities and differences between both analyses. The LFG analysis contains more fine-grained information, such as tense and mood features (see the feature TNS-ASP in the functional structure) or information about the noun type (see the feature NSEM/COMMON, with values ‘count’ and ‘mass’). Some properties of the LFG analyses are technically motivated, as is the case for complex phrasal nodes like ‘V[v,fin]’ (which means: finite main verb) or ‘DP[std]’ (standard DP, as opposed to interrogative or relative DPs).

In general, TIGER edge labels correspond to LFG functions (displayed in the feature-value matrix on the right in Fig. 8), and TIGER nodes correspond to LFG constituents (displayed in the tree on the left). For instance, both approaches analyze the word *Angst* (Eng.: fear) as the subject (SB = SUBJ) of the sentence, and the phrase *die Szene* (Eng.: the scene) as the object (OA = OBJ). In the LFG analysis, the definite article *die* is embedded under a specifier feature, whereas in the TIGER

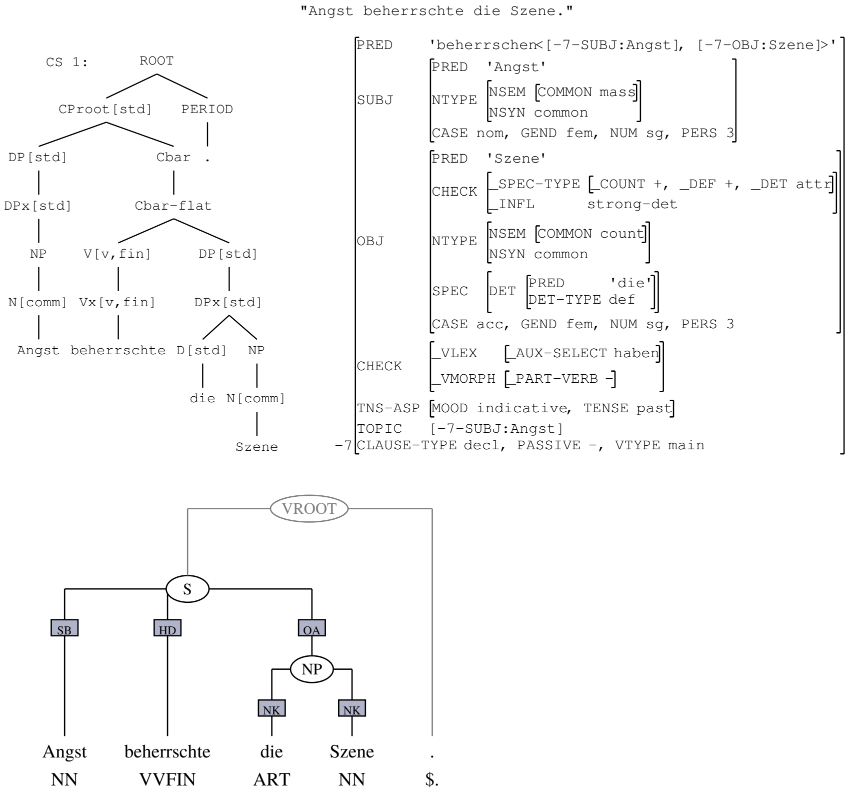


Fig. 8 LFG constituent and functional structures (top) and a TIGER analysis (bottom) of the sentence *Angst beherrschte die Szene.* (Eng.: Fear dominated the scene.)

analysis, it is a sister of the noun. The LFG node 'CProot[std]' corresponds to the 'S' node in the TIGER analysis, LFG nodes 'DP' are called 'NP' in TIGER.

Converting LFG to TIGER. To map LFG structures to the TIGER format, a transfer system was used [98].[9] The transfer system operates at the functional layer only, because this layer is assumed to be much more language-independent, as compared to the constituent layer. In a preprocessing step, constituent information had therefore to be folded into the functional layer.

Many transfer mappings concerned formal differences, such as renaming 'SUBJ' as 'SB' or 'DP[std]' as 'NP', or deleting unary nodes (e.g. NP nodes with just one daughter node). Other mappings resemble transformations known from natural language translation, such as 'head-switching'.

[9] The transfer system of the XEROX Translation Environment (XTE) by Martin Kay, which was part of the XLE development platform.

For instance, in the LFG analysis, the main verb provides the head of the clause, and auxiliaries provide aspectual and tense features. In contrast in TIGER, auxiliaries are analyzed as the head, and the main verb is embedded. Figure 9 shows an example: In the LFG analysis, the main predicate of the analysis is provided by the verb *abgelehnt* (Eng.: declined), see the feature ‘PRED’ with the value ‘ab#lehnen’. The auxiliaries *sei* and *worden* contribute features ‘MOOD subjunctive’, ‘PERF +_’ and ‘PASS-ASP dynamic_’ (embedded under the feature ‘TNS-ASP’), i.e. the sentence is in subjunctive mood, passive voice, perfect tense. In the corresponding TIGER analysis, the finite auxiliary *sei* is as the head of the sentence, embedding the second auxiliary *worden*, which, in turn, is the head of the main verb *abgelehnt*.

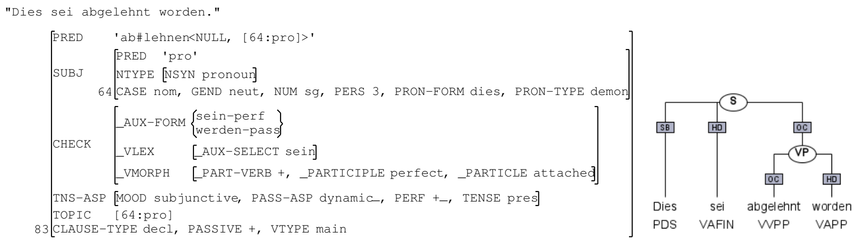


Fig. 9 Head-switching from a LFG functional structure (left) to the TIGER analysis (right) of the sentence *Dies sei abgelehnt worden.* (Eng.: This was declined.)

At the time of the TIGER project, the LFG grammar did not yet integrate a statistical disambiguation model. A symbolic ranking mechanism (similar to Optimality Theory) reduced the number of analyses to 17 on average, the median being 2 [27]. The task of the human annotators was then to disambiguate the remaining set of suggested analyses, using a range of tools provided by the XLE interface [44].

The grammar version of that time provided partial analyses for about 50% of the sentences; approximately 70% of the parsed sentences received the correct analysis (possibly among others).[10] Since producing the final output structures involved a series of successive steps, and only one third of the sentences could be analyzed this way, inter-annotator agreement was not computed.

A range of TIGER sentences were annotated this way at the University of Stuttgart. The second annotation of these sentences was done in the “traditional” way, using *Annotate*.

3.1.3 Comparison of the Approaches

Comparing both approaches is not straightforward because *Annotate* is a tool that has been developed specifically for this annotation task, and is therefore perfectly

[10] The grammar was later improved and extended, and, as of 2006, had a coverage of 86% in terms of full parses, and dependency-based F-scores of 84% [71, 25].

tailored to it. The LFG grammar has been developed independently from the TIGER project so that a considerable amount of work went into the conversion routines.

Hence, the tool-based approach using *Annotate* was clearly the easier way to go. The coverage of the LFG grammar was not broad enough so that sentences without a correct parse had to be annotated with *Annotate*. Some of the ambiguities produced by the grammar involved rather subtle differences and were difficult to spot for the annotators. Annotators not only had to know German syntax very well but also needed to know how to interpret complex LFG analyses.[11] The rather complicated mapping to the TIGER structures was another source of potential errors.

Still, it was worthwile to pursue both approaches, especially for improving the LFG grammar and creating resources for evaluating large-scale symbolic grammars. Among other things, the work initiated the creation of the TiGer Dependency Bank [26] (see Sect. 5).

3.2 The Annotation Process in TüBa-D/Z

The TüBa-D/Z treebank is annotated manually, or rather semi-automatically. In a first step, the newspaper text is segmented into sentences and tokenized. Then, the sentences are POS tagged automatically. This POS tagged version is then the basis for the syntactic annotation, which is performed in the tool *Annotate* [9, 64], as is TIGER. The interactive process of the tool suggesting individual groupings was found to provide a good balance between providing consistent annotation and forcing the annotator to look at individual annotation decisions rather than at complete trees. The morphological annotation is performed via automatic morphological analysis and disambiguation [89, 92]. These analyses are then integrated into the treebank and manually corrected. The parser within *Annotate*, which makes grouping suggestions, is regularly retrained on finished sections of the treebank.

The annotation of anaphora and coreference [60] started in 2006. To annotate these discourse phenomena, first mentions are automatically extracted from the syntax annotation: Every noun phrase (NX) generates one mention. Then, the anaphoric and coreference relations are manually annotated in *PALinkA* [62] and finally automatically integrated into the treebank (in NEGRA export format, see Sect. 4).

The dependency version [52] and the chunk version [48] are created automatically via scripts from the constituent version of the treebank.

Annotation guidelines. For the syntactic annotations, the annotation decisions are documented in an extensive stylebook, which is continuously updated. The current version, from 2012 [87]), is the fifth version and is more than 130 pages long.

[11] Flickinger et al. (chapter "Sustainable Development and Refinement of Complex Linguistic Annotations at Scale" in this volume) discuss the use of discriminants in grammar-based treebanking. Discriminants encode the features distinguishing competing analyses and can support annotators in disambiguating complex structures. Such an approach was later adapted to LFG in the INESS project, which developed the LFG Parsebanker. This tool has been applied in creating the Norwegian LFG treebank [73, 57].

The stylebook does not only cover difficult annotation decisions, but also the underlying principles of the treebank. One of the most difficult distinctions in the treebank, distinguishing between PP complements (OPP), optional complements (FOPP), and modifiers (MOD), is based on a list of verbal subcategorization frames [34]. The list is complete in the sense that it covers all verbs and all subcategorization frames that are annotated in the current release of the treebank. An example of a verb entry for *kontrollieren* (Eng.: to control) is shown in Fig. 10. This entry lists four subcategorization frames, the first having a subject (ON) and a direct object (OA), the second a subject and a clausal object (OS), the third only a subject, and the fourth a subject, a direct object, and an optional complement (FOPP). For every frame, at least one typical example from the treebank is provided along with the sentence number (e.g. R8-18: the 18th sentence in release 8). In cases where untypical examples are found, they can be added to the examples, as shown in the first frame. In the list, only complements are listed, modifiers (MOD) are not.

```
kontrollieren:
==============

  ON [kontrollieren] OA     (R8-18)
  Bsp: Ich kontrolliere solche Sachen
  Bsp: weil sich der Sport selbst kontrollieren soll (R8-42154)

  ON [kontrollieren] OS     (R8-37801)
  Bsp: InserentInnen sollten kontrollieren, "Satz"

  ON [kontrollieren]     (R8-39171)
  Bsp: Kontrollieren soll nicht ein neues Gremium

  ON [kontrollieren] OA FOPP (auf)   (R8-73574)
  Bsp: Er kontrolliert die BVG-Fahrkartenentwerter auf ihre
       Funktionstüchtigkeit
```

Fig. 10 An entry from the verb list showing all subcategorization frames for the verb *kontrollieren* (Eng.: control)

Annotators. The syntactic annotation is carried out by advanced students of Linguistics, German Linguistics, or Computational Linguistics. For (morpho-)syntax, morphology, and named entities, every sentence is annotated once by a student. During the annotation process, the annotators make notes of difficult cases or cases not covered in the stylebook. There are regular annotator meetings to discuss the difficult cases and potential additions to the stylebook. In a second round, every sentence is checked by a trained linguist (Heike Telljohann), who has accompanied the project from the very beginning. Before a new release, the whole treebank is checked for consistency via scripts and *TIGERSearch* queries. These scripts flag trees that exhibit annotations not normally found in correct annotations. Thus if an annotator accidentally had accepted a sentence with two subjects, such a sentence would be found, at the latest by the scripts. Because of the setup combining student annotators with a final check by an expert, inter-annotator agreement was not calculated.

4 Physical Representation

Both treebanks are available in a range of formats. Two of them, the *NEGRA export format* and *TIGER-XML*, are used by both treebanks. We first present the two common formats and then address others that are treebank-specific. While we are aware that the information presented here is too concise to serve as a reference,[12] the goal of this section is to familiarize the readers with the existing formats so that they can make an informed decision which formats they should use for specific applications. For instance, in the past, users often have used the Penn Treebank format (see below) of TüBa-D/Z and NEGRA without realizing that this version does not have the complete information of the original annotation format.

4.1 NEGRA Export Format

Since TIGER and TüBa-D/Z are annotated with the *Annotate* tool [9, 64] (for more details on the annotation process and the tool, see Sect. 3), the native data format for both treebanks is the *NEGRA export format*, which is the format that is automatically extracted from the database underlying *Annotate*.

The NEGRA export format is a column-based representation, which can model POS annotation, morphology, and constituent annotation, including crossing branches. A technical description of this data format can be found in [5]. In the NEGRA export format, every word and every syntactic node is represented as one row in a table. The columns of the table are predefined, covering both word nodes and syntactic nodes.

Word nodes. For word nodes, the first column contains the word, the second column contains the lemma if available, the third one the POS tag, and the fourth column the morphological tag. The fifth and sixth columns are reserved for syntactic information. The fifth column contains the grammatical function of the word, and the sixth column a number that points to the word's mother node. Optionally, columns seven and eight contain the label and the pointer to a node to which the current node has a secondary edge. The last column can be followed by a comment, starting with a % sign.[13]

Syntactic nodes. For syntactic nodes, the first column contains the node's ID (e.g., #500 for the first node in the tree, i.e., the leftmost lowest node). Node IDs

[12] For discussions of these and similar formats, see also Ide et al. (chapter "Designing Annotation Schemes: From Model to Representation" in this volume).

[13] This description refers to the NEGRA export format 4. There is a previous version, export format 3, which lacks the lemma column, but is otherwise identical.

start with no. 500, and daughter nodes must have lower numbers than their parents. The second and fourth columns do not contain any information for syntactic nodes. The third column contains the label of the syntactic node, and the fifth and sixth columns contain the grammatical function and the pointer to the mother node, as for words. Thus, if the syntactic node #500 points to 510, this means that node #510 is the mother node of #500.

Figure 12 shows the NEGRA export format representation for the TüBa-D/Z tree in Fig. 11. Note that the sentence has one misspelled word, *Kandidatengetz*, which was corrected in the comment in the export format. TIGER and TüBa-D/Z both use the comment field to add information that goes beyond the NEGRA export format. In the sentence in Fig. 12, for example, the subject of the subordinate clause *es* (Eng.: it) is marked as an expletive *it*. The sentence also shows a secondary edge from the VXINF node #518 to the VXINF #517. This is marked by an arc between the two nodes in Fig. 11. In this case, the secondary edge details the head information between the participle *erwartet* (Eng.: expected) and the infinitive *werden* (Eng.: be). This is necessary because we have three verbal forms in the verb complex (VC), and only one of them carries head (HD) information.

NEGRA Header. The NEGRA export format starts with a header providing different kinds of meta information. Figure 13 shows an excerpt of the header of the TIGER treebank.

The section named `#BOT ORIGIN` provides information about the origins of the sentences, BOT is short for 'beginning of table'. In the case of the TIGER corpus, this part defines IDs of the newspaper articles that make up the corpus, along with information about the articles' domains. For instance, 'NAC' means 'Nachrichten' (Eng.: news), 'FEU' means 'Feuilleton', and 'WIR' means 'Wirtschaft' (Eng.: economy). The header also contains lists of all tags that can be used in the annotation of the corpus. Figure 13 shows selected POS tags (under the header `#BOT WORDTAG`) and morphological tags (`#BOT MORPHTAG`).

Each sentence in the corpus is preceded by a line starting with `#BOS` ('beginning of sentence'), see Fig. 14. The first number following the BOS marker (`6025`) is the sentence number, the second number the annotator's ID (`0`), the third number (`1062583297`) shows the date of the annotation, encoded in Unix format (i.e., seconds since 1/1/1970). The last number (`86`) refers to the article ID, i.e., this sentence comes from the News section (compare to the meta-information in Fig. 13). Unfortunately, for TIGER, not all article information has been preserved correctly in the NEGRA export format; some IDs were lost in the course of the annotation process.

4.2 TIGER-XML

In a collaboration between TIGER and the EU project MATE ('Multi-level Annotation Tools Engineering'), an XML-based representation format for syntactically-annotated corpora was developed: TIGER-XML [56]. Its purpose was to serve as a

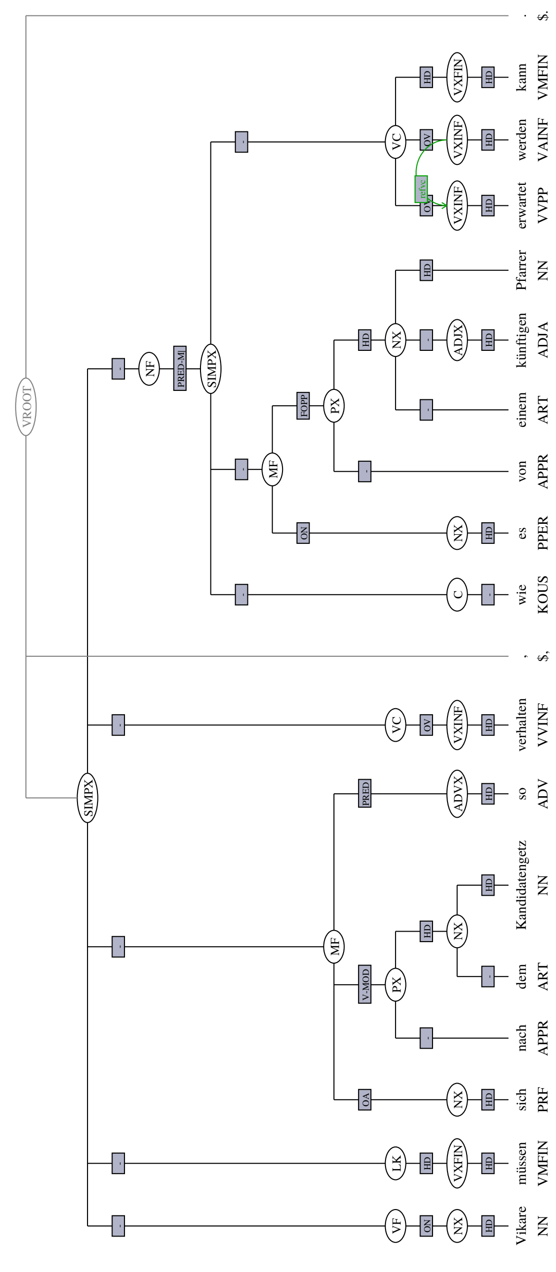


Fig. 11 The sentence *Vikare müssen sich nach dem Kandidatengetz so verhalten, wie es von einem künftigen Pfarrer erwartet werden kann.* (Eng.: According to the Candidates' Law, vicars must act as can be expected from a future priest.) from the TüBa-D/Z treebank

```
#BOS 24538 2 1134150923 1146
Vikare            Vikar              NN     npm  HD        500
müssen            müssen%aux         VMFIN  3pis HD        502
sich              #refl              PRF    ap*3 HD        504
nach              nach               APPR   d    -         506
dem               das                ART    dsn  -         505
Kandidatengetz    Kandidatengesetz   NN     dsn  HD        505 %% Kandidatengesetz
so                so                 ADV    –    HD        507
verhalten         verhalten          VVINF  –    HD        509
,                 ,                  $,     –    –         0
wie               wie                KOUS   –    -         511
es                es                 PPER   nsn3 HD        512
von               von                APPR   d    -         515
einem             ein                ART    dsm  -         514
künftigen         künftig            ADJA   dsm  HD        513
Pfarrer           Pfarrer            NN     dsm  HD        514
erwartet          erwarten           VVPP   –    HD        517
werden            werden%passiv      VAINF  –    HD        518
kann              können%aux         VMFIN  3sis HD        519
.                 .                  $.     –    –         0
#500              –                  NX     –    ON        501
#501              –                  VF     –    -         523
#502              –                  VXFIN  –    HD        503
#503              –                  LK     –    -         523
#504              –                  NX     –    OA        508
#505              –                  NX     –    HD        506
#506              –                  PX     –    V-MOD     508
#507              –                  ADVX   –    PRED      508
#508              –                  MF     –    -         523
#509              –                  VXINF  –    OV        510
#510              –                  VC     –    -         523
#511              –                  C      –    -         521
#512              –                  NX     –    ON        516 %% R=expletive
#513              –                  ADJX   –    -         514
#514              –                  NX     –    HD        515
#515              –                  PX     –    FOPP      516
#516              –                  MF     –    -         521
#517              –                  VXINF  –    OV        520
#518              –                  VXINF  –    OV        520 refvc 517
#519              –                  VXFIN  –    HD        520
#520              –                  VC     –    -         521
#521              –                  SIMPX  –    PRED-MOD  522
#522              –                  NF     –    -         523
#523              –                  SIMPX  –    –         0
#EOS 24538
```

Fig. 12 The NEGRA export representation of the tree in Fig. 11

```
%% database tiger2 (corpus tiger2)
%%
#FORMAT 4
#BOT ORIGIN
0     --          %%
86    fr951112    %% Frankfurter Rundschau 19951106 NAC D11050364
87    fr951112    %% Frankfurter Rundschau 19951106 FEU D11050368
88    fr951112    %% Frankfurter Rundschau 19951106 WIR D11050401
#EOT ORIGIN

#BOT WORDTAG
-1    UNKNOWN   N     Unbekanntes Tag aus Einlesen aus Korpusdatei
0     --        N     <Nicht zugeordnet>
1     ADJA      Y     Attributives Adjektiv
2     ADJD      Y     Adverbiales oder prdikatives Adjektiv
3     ADV       Y     Adverb
#EOT WORDTAG

#BOT MORPHTAG
-1    UNKNOWN         unknown tag
0     --              not bound
89    1.Nom.Sg.Fem    -
90    1.Nom.Sg.Masc   -
#EOT MORPHTAG
```

Fig. 13 Excerpts of the TIGER header in *NEGRA export format*

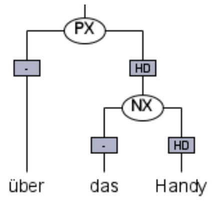
```
#BOS 6025 0 1062583297 86 %% @PO2AV@
An      an      APPR    --          AC  506
der     der     ART     Dat.Sg.Fem  NK  506
Grenze  Grenze  NN      Dat.Sg.Fem  NK  506
```

Fig. 14 A sample fragment of a TIGER sentence along with meta-information in the #BOS line

common exchange format for different treebank formats, and it serves as the native input format for the search tool *TIGERSearch*.

Straightforward use of XML for encoding tree structures would exploit embedding as the device for representing hierarchical structures, as shown in the XML code on the left in Fig. 15. Embedding cannot deal with crossing branches, though. The format TIGER-XML encodes hierarchical relations using pointers. Mother nodes point to their daughter nodes by means of `idref` attributes. The NEGRA export format uses a similar device, but pointers are reversed: in NEGRA, daughter nodes point to their mothers. TIGER-XML also provides extra elements for edges, so that they can be easily labeled with functional information, see the XML code in Fig. 15.

Figures 16 and 17 show a complete sentence from the TüBa-D/Z treebank, as a visual graph and in TIGER-XML format.




```
<PX>
 <w word="über"/>
 <NX>
  <w word="das"/>
  <w word="Handy"/>
 </NX>
</PX>
```

```
<terminals>
 <t id="s24_8" word="über" />
 <t id="s24_9" word="das" />
 <t id="s24_10" word="Handy" />
</terminals>
<nonterminals>
 <nt id="s24_507" cat="NX">
  <edge label="-" idref="s24_9" />
  <edge label="HD" idref="s24_10" />
 </nt>
 <nt id="s24_508" cat="PX">
  <edge label="-" idref="s24_8" />
  <edge label="HD" idref="s24_507" />
 </nt>
</nonterminals>
```

Fig. 15 The phrase *über das Handy* (Eng.: via the mobile phone), encoded by simple XML embedding (left) and TIGER-XML (right)

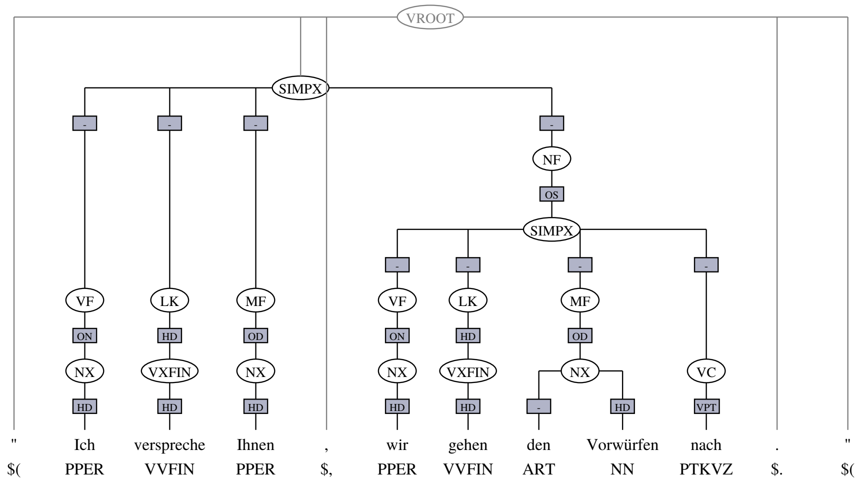


Fig. 16 The sentence *"Ich verspreche Ihnen, wir gehen den Vorwürfen nach."* (Eng.: "I promise you, we are looking into the accusations.") from the TüBa-D/Z treebank

Compared to the NEGRA export format, comments and header information (including information about article boundaries) are missing in the TIGER-XML format.

```
<s id="s5018">
<graph root="s5018_515">
<terminals>
 <t id="s5018_1" word="&quot;" lemma="&quot;" pos="$(" morph="--" />
 <t id="s5018_2" word="Ich" lemma="ich" pos="PPER" morph="ns*1" />
 <t id="s5018_3" word="verspreche" lemma="versprechen" pos="VVFIN" morph="1sis" />
 <t id="s5018_4" word="Ihnen" lemma="Sie" pos="PPER" morph="dp*3" />
 <t id="s5018_5" word="," lemma="," pos="$," morph="--" />
 <t id="s5018_6" word="wir" lemma="wir" pos="PPER" morph="np*1" />
 <t id="s5018_7" word="gehen" lemma="nach#gehen" pos="VVFIN" morph="1pis" />
 <t id="s5018_8" word="den" lemma="der" pos="ART" morph="dpm" />
 <t id="s5018_9" word="Vorw&#x00fc;rfen" lemma="Vorwurf" pos="NN" morph="dpm" />
 <t id="s5018_10" word="nach" lemma="--" pos="PTKVZ" morph="--" />
 <t id="s5018_11" word="." lemma="." pos="$." morph="--" />
 <t id="s5018_12" word="&quot;" lemma="&quot;" pos="$(" morph="--" />
</terminals>

<nonterminals>
 <nt id="s5018_500" cat="NX">
  <edge label="HD" idref="s5018_2"/>
 </nt>
 <nt id="s5018_501" cat="VF">
  <edge label="ON" idref="s5018_500"/>
 </nt>
 <nt id="s5018_502" cat="VXFIN">
  <edge label="HD" idref="s5018_3"/>
 </nt>
 <nt id="s5018_503" cat="LK">
  <edge label="HD" idref="s5018_502"/>
 </nt>
 <nt id="s5018_504" cat="NX">
  <edge label="HD" idref="s5018_4"/>
 </nt>
 <nt id="s5018_505" cat="MF">
  <edge label="OD" idref="s5018_504"/>
 </nt>
 <nt id="s5018_506" cat="NX">
  <edge label="HD" idref="s5018_6"/>
 </nt>
 <nt id="s5018_507" cat="VF">
  <edge label="ON" idref="s5018_506"/>
 </nt>
 <nt id="s5018_508" cat="VXFIN">
  <edge label="HD" idref="s5018_7"/>
 </nt>
 <nt id="s5018_509" cat="LK">
  <edge label="HD" idref="s5018_508"/>
 </nt>
 <nt id="s5018_510" cat="NX">
  <edge label="-" idref="s5018_8"/>
  <edge label="HD" idref="s5018_9"/>
 </nt>
 <nt id="s5018_511" cat="MF">
  <edge label="OD" idref="s5018_510"/>
 </nt>
 <nt id="s5018_512" cat="VC">
  <edge label="VPT" idref="s5018_10"/>
</nt>
 <nt id="s5018_513" cat="SIMPX">
  <edge label="-" idref="s5018_507"/>
  <edge label="-" idref="s5018_509"/>
  <edge label="-" idref="s5018_511"/>
  <edge label="-" idref="s5018_512"/>
 </nt>
 <nt id="s5018_514" cat="NF">
  <edge label="OS" idref="s5018_513"/>
 </nt>
 <nt id="s5018_515" cat="SIMPX">
  <edge label="-" idref="s5018_501"/>
  <edge label="-" idref="s5018_503"/>
  <edge label="-" idref="s5018_505"/>
  <edge label="-" idref="s5018_514"/>
 </nt>
</nonterminals>
</graph>
</s>
```

Fig. 17 The annotation of the sentence from Fig. 16 in TIGER-XML

Recently, the format *<tiger2/>* has been proposed, which is an extension of TIGER-XML [2, 72]. The goal of <tiger2/> is to serve as the serialization format for the ISO Syntactic Annotation Framework SynAF.[14]

4.3 The TIGER Treebank Formats

The TIGER treebank is officially available in four different formats:

1. TIGER-XML (all releases)
2. NEGRA export format (releases 1–2.1)
3. Penn Treebank format (release 1)
4. CoNLL dependency format (release 2.2)

PennTreebank format. The PennTreebank bracketing format is available officially only for TIGER release 1. The format was probably created via the script 'negra-tocfg' by Thorsten Brants, which operated on the NEGRA format.[15] The format does not contain traces. Instead, relations that give rise to crossing branches are reallocated. The standard approach for this transformation is to re-attach crossing non-head constituents as sisters of the lowest mother node that dominates all the crossing constituent and its sister nodes in the original TIGER tree [47].

CoNLL dependency format. There are several conversions of the TIGER treebank to CoNLL-style dependencies: the version used in the CoNLL 2009 Shared Task [30], the one used in the PaGe Shared Task [47], and a version that has been created recently by means of the tool *Tiger2Dep* [77].[16] In the CoNLL format [13, 61], each word is accompanied by a pointer, which indicates the word's governor, as in the NEGRA export format (see Sect. 4.1; for more details on the CoNLL format, see Sect. 4.4).

The CoNLL 2009 Shared Task data set, which includes a subset of the TIGER treebank converted to dependency relations, stays close to the original TIGER annotation scheme and uses rather flat structures. The PaGe Shared Task data set and the tool *Tiger2Dep* use heuristic rules to determine the head of each phrase (which often is not specified explicitly, see Sect. 2.1), and introduce PP-internal structures.

Figure 18 shows a (simplified) example: The first structure (left) represents the original TIGER treebank annotation. The second analysis (center) shows the CoNLL 2009 dependency version, where both the article and head noun are directly governed by the preposition. The third version (right) shows the analysis of PaGe and *Tiger2Dep*: the article *den* (Eng.: the) is governed by the head noun *USA*, which in turn is governed by the preposition *in*.

[14] SynAF is a standard developed by the International Organization for Standardisation in ISO/TC37/SC4 (Language Resources Management); http://www.tc37sc4.org/, see Ide et al. (chapter "Designing Annotation Schemes: From Model to Representation" in this volume).

[15] The script was part of the NEGRA corpus deliverable. The script could not deal correctly with some kinds of crossing branches and was not maintained after the end of NEGRA.

[16] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/Tiger2Dep.en.html

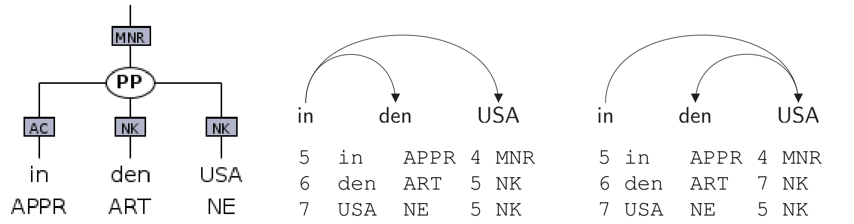


Fig. 18 The phrase *in den USA* (Eng.: in the US) in the original TIGER analysis (left), a rather flat dependency analysis (center: CoNLL 2009), and a hierarchical dependency analysis (right: PaGe and *Tiger2Dep*) (showing relevant columns only)

The TIGER treebank was also used to derive triples encoding the governor, its dependent, and the type of relation holding between them, the *TIGER Dependency Triples* [45]. For instance, the triple `mo(wäre~0, vielleicht~5)` encodes the information that the terminal node *vielleicht* (Eng.: perhaps) is a modifier ('mo') of the node *wäre* (Eng.: would be). The numbers serve as unique identifers.

Finally, there were initiatives to automatically derive "enriched" formats, i.e. formats with unary nodes (e.g. NP nodes dominating pronouns) and NP nodes within PPs (e.g., [75, 74]). Unfortunately, there is no official release in such an enriched format available.

4.4 The TüBa-D/Z Treebank Formats

The TüBa-D/Z treebank is available in five different formats:

1. NEGRA export format
2. TIGER-XML
3. Export XML
4. Penn Treebank format
5. CoNLL dependency format

Apart from the native *NEGRA Export Format*, TüBa-D/Z is also available in two XML formats: in *TIGER-XML* (see above) and in *Export XML*. In the TIGER-XML format, the focus is on the (morpho-)syntactic annotation. This means, neither the anaphora and coreference annotations nor the discourse connective and word sense annotations are available.

Export XML. The Export XML format is more closely oriented towards the NEGRA export format and the annotations in the TüBa-D/Z treebank. Thus, since TüBa-D/Z models (mostly) pure tree structures without crossing branches, the hierarchical XML structure is used to model the constituent trees. The Export XML representation of the tree in Fig. 16 is shown in Fig. 19. Note that this XML version contains all available annotations, including the ones that go beyond the syntactic annotation. The example in Fig. 19 shows, for example, that the subject of the main

```
<sentence xml:id="s5018">
 <word xml:id="s5018_1" form='"' pos="$(" lemma='"' func="--" deprel="ROOT"/>
 <node xml:id="s5018_515" cat="SIMPX" func="--">
  <node xml:id="s5018_501" cat="VF" func="-" parent="s5018_515">
   <node xml:id="s5018_500" cat="NX" func="ON" parent="s5018_501">
    <relation type="anaphoric" target="s5017_502"/>
    <word xml:id="s5018_2" form="Ich" pos="PPER" morph="ns*1"
                           lemma="ich" func="HD" parent="s5018_500"
                           dephead="s5018_3" deprel="SUBJ"/>
   </node>
  </node>
  <node xml:id="s5018_503" cat="LK" func="-" parent="s5018_515">
   <node xml:id="s5018_502" cat="VXFIN" func="HD" parent="s5018_503">
    <word xml:id="s5018_3" form="verspreche" pos="VVFIN" morph="1sis"
                           lemma="versprechen" func="HD"
                           parent="s5018_502" deprel="ROOT"/>
   </node>
  </node>
  <node xml:id="s5018_505" cat="MF" func="-" parent="s5018_515">
   <node xml:id="s5018_504" cat="NX" func="OD" parent="s5018_505">
    <word xml:id="s5018_4" form="Ihnen" pos="PPER" morph="dp*3"
                           lemma="Sie" func="HD" parent="s5018_504"
    dephead="s5018_3" deprel="OBJD"/>
   </node>
  </node>
  <word xml:id="s5018_5" form="," pos="$," lemma="," func="--" deprel="ROOT"/>
  <node xml:id="s5018_514" cat="NF" func="-" parent="s5018_515">
   <node xml:id="s5018_513" cat="SIMPX" func="OS" parent="s5018_514">
    <node xml:id="s5018_507" cat="VF" func="-" parent="s5018_513">
     <node xml:id="s5018_506" cat="NX" func="ON" parent="s5018_507">
      <word xml:id="s5018_6" form="wir" pos="PPER" morph="np*1"
                             lemma="wir" func="HD" parent="s5018_506"
                             dephead="s5018_7" deprel="SUBJ"/>
     </node>
    </node>
    <node xml:id="s5018_509" cat="LK" func="-" parent="s5018_513">
     <node xml:id="s5018_508" cat="VXFIN" func="HD" parent="s5018_509">
      <word xml:id="s5018_7" form="gehen" pos="VVFIN" morph="1pis"
                             lemma="nach#gehen" func="HD" parent="s5018_508"
      dephead="s5018_3" deprel="S"/>
     </node>
    </node>
    <node xml:id="s5018_511" cat="MF" func="-" parent="s5018_513">
     <node xml:id="s5018_510" cat="NX" func="OD" parent="s5018_511">
      <word xml:id="s5018_8" form="den" pos="ART" morph="dpm"
                             lemma="der" func="-" parent="s5018_510"
      dephead="s5018_9" deprel="DET"/>
      <word xml:id="s5018_9" form="Vorwürfen" pos="NN" morph="dpm"
                             lemma="Vorwurf" func="HD" parent="s5018_510"
                             dephead="s5018_7" deprel="OBJD"/>
     </node>
    </node>
    <node xml:id="s5018_512" cat="VC" func="-" parent="s5018_513">
     <word xml:id="s5018_10" form="nach" pos="PTKVZ" func="VPT"
                             parent="s5018_512" dephead="s5018_7"
                             deprel="AVZ"/>
    </node>
   </node>
  </node>
 </node>
 <word xml:id="s5018_11" form="." pos="$." lemma="." func="--" deprel="ROOT"/>
 <word xml:id="s5018_12" form='"' pos="$(" lemma='"' func="--" deprel="ROOT"/>
</sentence>
```

Fig. 19 The annotation from Fig. 16 in Export XML clause *Ich* (Eng.: I) has an anaphoric relation to node #502 in the previous sentence (s5017), see the XML element 'relation'.

Penn Treebank format. The fourth format in which TüBa-D/Z is available is the Penn Treebank bracketing format. The representation of the tree in Fig. 16 in this format is shown in Fig. 20. This is similar to the bracketing format of the Penn Treebank. One difference to the original format is that no indentation is provided[17], another is that all trees are grouped under a virtual root node (VROOT). Grammatical functions are separated from their syntactic node by a colon rather than the dash used in the Penn Treebank because some TüBa-D/Z node labels contain dashes. Since TüBa-D/Z does not annotate crossing branches, no traces or empty categories are necessary. In order to avoid confusion between bracketing and the word '(' or the POS tag '$(', word and POS parentheses are converted into 'LBR'. An example of the POS tag '$LBR' is shown in the example in Fig. 20.

```
(VROOT:--
  ($LBR:-- ")
  (SIMPX:--
    (VF:-
      (NX:ON
        (PPER:HD Ich)))
    (LK:-
      (VXFIN:HD
        (VVFIN:HD verspreche)))
    (MF:-
      (NX:OD
        (PPER:HD Ihnen)))
    ($,:-- ,)
    (NF:-
      (SIMPX:OS
        (VF:-
          (NX:ON
            (PPER:HD wir)))
        (LK:-
          (VXFIN:HD
            (VVFIN:HD gehen)))
        (MF:-
          (NX:OD
            (ART:- den)
            (NN:HD Vorwürfen)))
        (VC:-(PTKVZ:VPT nach)))))
  ($.:-- .)
  ($LBR:-- "))
```

Fig. 20 The tree from Fig. 16 in the Penn Treebank format

Note that this format requires true tree structures. This means that parentheticals need to be grouped under their surrounding constituents. Thus, the tree in Fig. 4 is

[17] To enhance readability, we provide indentation in the example presented in Fig. 20.

represented as shown in Fig. 21, where the parenthetical ‘SIMPX’ is grouped as a daughter under the surrounding ‘SIMPX’.

```
(VROOT:--
  ($LBR:-- ")
  (SIMPX:--
    (VF:-
      (ADJX:PRED
        (ADJD:HD Schön)))
    ($LBR:-- ")
    ($,:-- ,)
    (SIMPX:--
      (LK:-
        (VXFIN:HD
          (VVFIN:HD sagte)))
      (MF:-
        (NX=PER:ON
          (NE:- Mehmet)
          (NE:- Scholl))))
    ($,:-- ,)
    ($LBR:-- ")
    (LK:-
      (VXFIN:HD
        (VAFIN:HD ist)))
    (MF:-
      (NX:ON
        (PDS:HD das))
      (ADVX:MOD
         (PTKNEG:HD nicht))))
  ($LBR:-- ")
  ($.:-- .))
```

Fig. 21 The tree from Fig. 4 in the Penn Treebank format

CoNLL dependency format. There is also a conversion of the constituent annotation into dependencies. This conversion is based (with adaptations) on the conversion scheme suggested by Kübler and Telljohann [52]. It is carried out automatically. Since the original annotation scheme labels head/non-head relations on the phrasal level, head-finding rules are not necessary, and heuristics need to be applied only for a small number of phenomena including coordination and apposition. During the conversion, long-distance relations that are marked with special labels in the constituent version are resolved into non-projective dependencies. Like TIGER, TüBa-D/Z also uses the column-based CoNLL format. However, TüBa-D/Z uses the standard 2006/2007 CoNLL format, not the extended 2009 one. The tree in Fig. 16 is shown in its dependency representation in Fig. 22. In the CoNLL format, there are eight columns, the first one gives each word an ID, the second column represents the word, the third the lemma. The fourth and fifth columns represent coarse and fine grained POS tags, and the sixth one the morphological annotation. The seventh and eighth columns represent the dependency analysis, showing for each word its head and the label of the dependency. For example, word 2 *Ich* (Eng.: I) in Fig. 22 has word 3 *verspreche* (Eng.: promise) as its head, and it is the subject (SUBJ).

1	”	”	$(	$(	–	2	-PUNCT-
2	Ich	ich	PRO	PPER	ns*1	3	SUBJ
3	verspreche	versprechen	V	VVFIN	1sis	0	ROOT
4	Ihnen	Sie	PRO	PPER	dp*3	3	OBJD
5	,	,	$,	$,	–	4	-PUNCT-
6	die	die	ART	ART	apf	8	DET
7	positiven	positiv	ADJA	ADJA	apf	8	ATTR
8	Kräfte	Kraft	N	NN	apf	11	OBJA
9	der	die	ART	ART	gsf	10	DET
10	Stadt	Stadt	N	NN	gsf	8	GMOD
11	zusammenzuführen	zusammen#führen	V	VVIZU	–	3	OBJI
12	.	.	$.	$.	–	11	-PUNCT-
13	”	”	$(	$(	–	11	-PUNCT-

Fig. 22 The annotation from Fig. 16 converted to dependencies and represented in the CoNLL format

5 Usage of TIGER and TüBa-D/Z

The annotation schemes for the TIGER and TüBa-D/Z treebanks were developed to allow a wide range of applications, ranging from training a parser to serving as data sources for corpus linguistic investigations. The treebanks are available free of charge for scientific use. Licensing the treebanks is handled as follows:

- TIGER: The treebank license can be signed online, giving immediate access to the download page.[18]
- TüBa-D/Z: After signing a license agreement, the user is given access to the download web page.[19] The treebank was also integrated into Weblicht[20], an execution environment for the automatic annotation of text corpora, and can be accessed via the Tübingen aNnotated Data Retrieval Application (*TüNDRA*)[21] [55]. *TüNDRA* is a web-based syntactic query tool.

Computational applications. The treebanks have been used extensively for parsing research on German, mostly in comparison to other treebanks. There is early work on comparing parsing results for TüBa-D/Z to results for NEGRA [46, 49].

[18] The license can be signed here:
http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/license/index.html.

[19] The license is available from http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html

[20] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

[21] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra

These investigations were followed by comparisons between TüBa-D/Z and TIGER [50, 67, 68]. Both TIGER and TüBa-D/Z were used in the shared task on 'Parsing German' (PaGe), co-located with an ACL workshop with the same focus [47].

Since the annotation schemes of TüBa-D/Z and TIGER are so different, and since the investigations above showed that the standard evaluation metrics are sensitive to the average number of nodes per sentence (which is one of the major differences between TIGER and TüBa-D/Z), these investigations also resulted in investigations into better evaluation metrics for parsing [16] and in the development of a test suite for difficult phenomena in TIGER and TüBa-D/Z, TePaCoC [51].

The TIGER treebank also served in evaluating hand-crafted grammars. To this end, 2000 sentences of the corpus were used to build the *TiGer Dependency Bank* (TiGer DB) [26], which has a format similar to the *PARC 700 Dependency Bank* [43] and was designed as a dependency-based gold standard for German grammars and parsers (including the German LFG grammar, see Sect. 3.1). The TIGER *700 RMRS Bank*, which contains 700 sentences, was derived from the *TiGer Dependency Bank* [83]. It is represented in the format of *Robust Minimal Recursion Semantics* (RMRS) and thus suitable for evaluating HPSG grammars (Head-Driven Phrase Structure Grammar [15]).

TüBa-D/Z was also used in more specialized applications, such as parsing for topological fields [91, 90], anaphora resolution [37], corpus masking [69], and for word order prediction in a generation task [95]. The latter application is an example which shows the importance of the interaction between syntax and discourse phenomena.

Search tools for linguists. Both treebanks can be searched with TIGERSearch ([54], developed in the TIGER project) and ANNIS [96]. The search tools were created to facilitate use of treebanks (and other types of corpora with ANNIS) for theoretical linguists. In addition, two tutorials targeting users from linguistics were written in the TIGER project, which provide guided tours for syntacticians and lexicographers, showing how to query the treebank with TIGERSearch [81], and how to use regular expressions for searching morphological annotations [82]. TIGERSearch is very popular and frequently used by corpus linguists but less often by other linguists.

As mentioned in Sect. 2.3, searching the TIGER treebank can be tricky due to the flat structures and crossing branches. Querying is made easier by the use of *templates* and *bookmarks*, which serve to store useful queries for later reuse. Figure 23 shows a sample template 'VF', adapted and simplified from [22], that implements a query for constituents in the initial field (i.e., the VF node in the TüBa-D/Z treebank). The query expression first specifies that there is a sentence node #s which dominates some node #vf (= the target node) and a finite verb #v2. Node #vf (or its descendant) is either the leftmost daughter of the sentence or preceded by a conjunction. Moreover, #vf is either directly followed by the finite verb #v2, or a comma may intervene in the case of clausal #vf constituents.

This template covers the majority of initial constituents. It can be called in TIGERSearch, e.g., as follows: `#s:[cat="S"] & VF(#s)`. This query searches for sentential constituents in the initial field. A sample match is shown in Fig. 24.

```
// Vorfeld constituent
VF(#vf) <-
   #s:[cat="S"] &
   #s > #vf &             // #vf: Vorfeld constituent
   #v2:[pos=/V.FIN/] &    // #v2: Verb in second position
   #s >HD #v2 &

   // VF is first constituent
   ( // 1. VF is very first element in the sentence
     ( #s >@l #vf         // vf is leftmost child
     | #vf >* #childL:[T] &
       #s >@l #childL
     )
   | // 2. Or some coordinating conjunction precedes VF
     #s >@l #conj &
     [] >JU #conj &
     #conj . #vf
   )
   &
   // VF precedes VFIN
   ( // 1. VF directly precedes V2
     #vf . #v2
   | // 2. A comma may intervene after clausal VF
     #vf: [cat=("S"|"VP")] &   // either VF itself precedes comma
     #vf >* #childR:[T] &
     #childR . #comma:[word="\,"] &
     #comma . #v2              // vf followed by comma + v2
   )
;
```

Fig. 23 Template in TIGERSearch for querying VF constituents

The top S node matches the expression named #s, the embedded S node (left) the expression #vf (with the sentence-initial word *Wie* matching #childL, and the sentence-final *will* #childR), and *verrät* matches #v2.

TIGERSearch Version 2.1.1 is distributed with a set of demo corpora, including a sampler of the TIGER treebank ('TIGERSampler' in the folder DemoCorpora/German). The sampler provides a collection of predefined templates and bookmarked queries.

There is also a search tool, ICARUS, which can be used for searching the dependency versions of the treebanks [28].[22]

Linguistic Studies. Both treebanks have also been used for linguistic studies, often in combination with TIGERSearch. Meurers and Müller [58] show that by searching in the syntactic annotations of TIGER, they can find occurrences of phenomena that have been claimed to be impossible. However, they also find that "many infrequent but theoretically relevant phenomena can only be found in very large corpora", which cannot be annotated manually. Harbusch and Kempen [32] use TIGER

[22] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/icarus.html

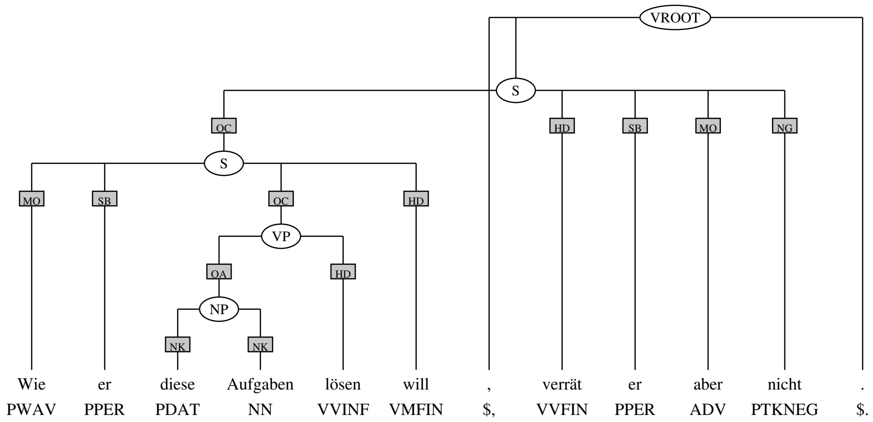


Fig. 24 The sentence *Wie er diese Aufgaben lösen will, verrät er aber nicht.* (Eng.: How he wants to solve these tasks, he does not say.) from the TIGER treebank

to investigate clausal coordinate ellipsis in German. They find examples in the corpus that are not covered in intuitionistic rules. Harbush [31] extends this investigation to a comparison of German and Dutch, this time focusing on incremental sentence production. Pappert et al. [63] investigate the validity of a set of linguistic constraints as predictors for German word order using a corpus linguistic approach in combination with psycholinguistic experiments.

TüBa-D/Z was also used for linguistic research: Hinrichs and Kübler [38, 39] investigate differences between written and spoken German, based on TüBa-D/Z and TüBa-D/S. They focus on the distribution of different types of noun phrases, direct and indirect questions, and different realizations of the Vorfeld. Zinsmeister [97] investigates coordination structures in TüBa-D/Z and presents a qualitative and quantitative survey of this phenomenon. Steiner [85] investigates partial agreement in German by comparing written data from TüBa-D/Z and spoken data from TüBa-D/S. Hinrichs and Beck [33] look at the historical development of auxiliary fronting using TüBa-D/Z in combination with the automatically annotated corpus TüPP-D/Z [59] and the German Text Archive (DTA)[23].

6 Other Treebanks for German

There are some medium- and smaller-sized treebanks for German which have been inspired by the TIGER treebank. They follow the TIGER annotation guidelines for syntactic annotations, and the STTS guidelines for POS annotations. The treebanks are:

1. The Potsdam Commentary Corpus (PCC) [84] is a corpus of German newspaper commentaries (44,000 tokens). It is annotated with various types of linguistic information: In addition to syntax, it is annotated for coreference, information structure, and discourse structure.
2. The Mercurius Treebank [19] is a treebank of a newspapers from 1597 and 1667, written in Early New High German (170,000 tokens); it is also annotated according to the TIGER guidelines.
3. Deutsche Diachrone Baumbank [40] (8,300 tokens) is a diachronic treebank with texts from Old, Middle and Early New High German. In addition to syntax, it is annotated with normalized wordforms, lemmas, and morphology.
4. SMULTRON (Stockholm MULtilingual TReebank) [93] is a parallel treebank of different languages, including German (version 3.0: 2,500 sentences). Besides syntactic annotations for both languages, the treebank contains alignments for words and phrases across the languages.

All dependency treebanks for German are the results of converting one of the treebanks NEGRA, TIGER, or TüBa-D/Z into the dependency format.[24]

A German LFG treebank has been created in the context of the Pargram project [14]. It contains automatically-created LFG analyses of almost 10,000 sentences (115,000 words) taken from the TIGER treebank and can be accessed via the INESS treebanking environment from Bergen [57].

7 Summary

In this chapter, we have presented the two major treebanks of German, TIGER and TüBa-D/Z. Even though the strategies for representing syntactic structures that the

---

[23] http://www.deutschestextarchiv.de/

[24] There is work in progress for the Copenhagen Dependency Treebank, but the annotations have not been released yet (http://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT). After the time of writing, the Hamburg Dependency Treebank was announced in 2014, which consists of approx. 2,00,000 manually annotated sentences plus 55,000 automatically parsed sentences, see https://corpora.uni-hamburg.de/drupal/de/islandora/object/treebank:hdt.

two treebanks follow are quite different, both have become quasi-standards for German treebanks. At the same time, it is obvious that neither one of the schemes satisfies the needs and requirements of all applications. This is clearly shown by the fact that both treebanks have been subjected to different conversions, targeting dependency or other formats.

A prominent difference between both treebanks is that TüBa-D/Z is still being extended, both in size and annotation layers (such as named entities, coreference, or discourse structure). Thanks to the wealth of information that is nowadays part of the TüBa-D/Z treebank, it has become a very interesting resource, simply because it is useful for a broad range of applications.

The fact that both treebanks are based on newspaper texts is certainly a major disadvantage. Extending the treebanks to include other domains and genres seems to be one of the most pressing issues.

References

1. Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pußel, Marco Rower, Bettina Schrader, Anne Schwartz, George Smith, and Hans Uszkoreit. TIGER Annotationsschema, 2003. Technical Report, Universität des Saarlandes, Universität Stuttgart, Universität Potsdam, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-syntax.pdf.
2. Sonja Bosch, Key-Sun Choi, Éric de La Clergerie, Alex Chengyu Fang, Gertrud Faass, Kiyong Lee, Antonio Pareja-Lora, Laurent Romary, Andreas Witt, Amir Zeldes, and Florian Zipser. <tiger2/> as a standardised serialisation for ISO 24615 – SynAF. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 37–60, Lisbon, Portugal, 2012.
3. Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation, Special Issue*, 2(4):597–620, 2004.
4. Sabine Brants and Silvia Hansen. Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation LREC-02*, pages 1643–1649, Las Palmas de Gran Canaria, 2002.
5. Thorsten Brants. *The NeGra Export Format for Annotated Corpora*. Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany, 1997. CLAUS Report No. 98, http://www.coli.uni-saarland.de/˜thorsten/publications/Brants-CLAUS98.pdf.
6. Thorsten Brants. Cascaded Markov models. In *Proceedings of EACL-99*, Bergen, Norway, 1999.
7. Thorsten Brants. Inter-annotator agreement for a German newspaper corpus. In *Proceedings of Second International Conference on Language Resources and Evaluation LREC-2000*, Athens, Greece, 2000.
8. Thorsten Brants. TnT — a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000*, Seattle, WA, 2000.
9. Thorsten Brants and Wojciech Skut. Automation of treebank annotation. In *Proceedings of the Joint Conference on New Methods in Natural Language Processing and Computational Language Learning, NeMLaP3/CoNLL98*, pages 49–57, Sydney, Australia, 1998.

10. Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA Treebank Workshop*, pages 69–76, Paris, France, 1999.
11. Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. Syntactic annotation of a German newspaper corpus. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, pages 73–87. Springer Netherlands, 2003.
12. Joan Bresnan. *The Mental Representation of Grammatical Relations*. MIT Press, 1982.
13. Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Language Learning (CoNLL)*, pages 149–164, New York, NY, 2006.
14. Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The Parallel Grammar Project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan, 2002.
15. Ivan A. Sag Carl Pollard. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. Chicago: University of Chicago Press, 1994.
16. Anna Corazza, Alberto Lavelli, and Giorgio Satta. An information theoretic measure to evaluate parsing difficulty across treebanks. *ACM Transactions on Speech and Language Processing*, 9(4), 2013.
17. Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman. XLE documentation. Technical Report, Palo Alto Research Center.
18. Berthold Crysmann, Silvia Hansen-Schirra, George Smith, and Dorothea Ziegler-Eisele. TIGER Morphologie-Annotationsschema. Technical report, Universität Saarbrücken, Universität Potsdam, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-morph.pdf.
19. Ulrike Demske. Das Mercurius-Projekt: eine Baumbank für das Frühneuhochdeutsche. In Gisela Zifonun and Werner Kallmeyer, editors, *Sprachkorpora — Datenmengen und Erkenntnisfortschritt*, Jahrbuch des Instituts für deutsche Sprache 2006, pages 91–104. Berlin: de Gruyter, 2007.
20. Stefanie Dipper. Grammar-based corpus annotation. In *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora (LINC-2000)*, pages 56–64, Luxembourg, 2000.
21. Stefanie Dipper. *Implementing and Documenting Large-Scale Grammars — German LFG*. PhD thesis, IMS, University of Stuttgart, 2003. Working papers of the Institut für Maschinelle Sprachverarbeitung (AIMS), vol. 9(1).
22. Stefanie Dipper. Querying topological fields in the TIGER scheme with TIGERSearch. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 37–50, Tübingen, Germany, 2014.
23. Erich Drach. *Grundgedanken der Deutschen Satzlehre*. Diesterweg, Frankfurt/M., 1937.
24. Oskar Erdmann. *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt*. Verlag der Cotta'schen Buchhandlung, Stuttgart, 1886. Erste Abteilung.
25. Martin Forst. Treebank Conversion — Establishing a testsuite for a broad-coverage LFG from the TIGER Treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC '03)*, Budapest, 2003.
26. Martin Forst, Núria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, and Valia Kordoni. Towards a dependency-based gold standard for German parsers — the TiGer dependency bank. In *Proceedings of LINC 2004*, 2004.
27. Anette Frank, Tracy Holloway King, Jonas Kuhn, and John Maxwell. Optimality Theory style constraint ranking in large-scale LFG grammars. In *Proceedings of the Third LFG Conference*, Brisbane, Australia, 1998.
28. Markus Gärtner, Gregor Thiele, Wolfgang Seeker, Anders Björkelund, and Jonas Kuhn. ICARUS – an extensible graphical search tool for dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

29. Anne Gastel, Sabrina Schulze, Yannick Versley, and Erhard Hinrichs. Annotation of explicit and implicit discourse relations in the TüBa-D/Z Treebank. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, Hamburg, Germany, 2011.
30. Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June 2009. Association for Computational Linguistics.
31. Karin Harbusch. Incremental sentence production inhibits clausal coordinate ellipsis: A treebank study into Dutch and German. *Dialogue & Discourse. Special issue on "Incremental Processing in Dialogue"*, 2(1):313–332, 2011.
32. Karin Harbusch and Gerard Kempen. Clausal coordinate ellipsis in German: The TIGER treebank as a source of evidence. In *Proceedings of NODALIDA 2007 - Sixteenth Nordic Conference of Computational Linguistics*, Tartu, Estonia, 2007.
33. Erhard Hinrichs and Kathrin Beck. Auxiliary fronting in German: A walk in the woods. In *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT)*, Sofia, Bulgaria, 2013.
34. Erhard Hinrichs and Heike Telljohann. Constructing a valence lexicon for a treebank of German. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 41–52, Groningen, The Netherlands, 2009.
35. Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. The Tübingen treebanks for spoken German, English, and Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 550–574. Berlin et al.: Springer, Berlin, 2000.
36. Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. The Verbmobil treebanks. In *Proceedings of KONVENS 2000, 5. Konferenz zur Verarbeitung natürlicher Sprache*, pages 107–112, Ilmenau, Germany, 2000.
37. Erhard W. Hinrichs, Katja Filippova, and Holger Wunsch. What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German. In Montserrat Civit, Sandra Kübler, and Maria Antònia Martí, editors, *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spain, 2005.
38. Erhard W. Hinrichs and Sandra Kübler. Treebank profiling of spoken and written German. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, pages 65–76, Barcelona, Spain, 2005.
39. Erhard W. Hinrichs and Sandra Kübler. What linguists always wanted to know about German and did not know how to ask. In Mickael Suominen, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen, and Kaius Sinnemäki, editors, *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday.*, volume 19 of *SKY Journal of Linguistics*, pages 24–33. The Linguistic Association of Finland, 2006. Special Supplement.
40. Hagen Hirschmann and Sonja Linde. Annotationsguidelines zur Deutschen Diachronen Baumbank, 2010. Technical Report, Humboldt-Universität zu Berlin, http://korpling.german.hu-berlin.de/ddb-doku.
41. Tilman Höhle. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany, 1986.
42. Laura Kallmeyer and Wolfgang Maier. Data-driven parsing using probabilistic linear context-free rewriting systems. *Computational Linguistics*, 39(1), 2013.
43. Tracy Holloway King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald M. Kaplan. The PARC700 dependency bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL-03*, 2003.
44. Tracy Holloway King, Stefanie Dipper, Anette Frank, Jonas Kuhn, and John Maxwell. Ambiguity management in grammar writing. *Research on Language and Computation*, 2004.

45. Manuel Kountz. Extraktion von Dependenztripeln aus der TIGER-Baumbank, 2006. Studienarbeit, Universität Stuttgart.
46. Sandra Kübler. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 293–300, Borovets, Bulgaria, 2005.
47. Sandra Kübler. The PaGe shared task on parsing German. In *Proceedings of the ACL Workshop on Parsing German*, pages 55–63, Columbus, OH, 2008.
48. Sandra Kübler, Kathrin Beck, Erhard Hinrichs, and Heike Telljohann. Chunking German: An unsolved problem. In *Proceedings of the Forth Linguistic Annotation Workshop (LAW)*, pages 147–151, Uppsala, Sweden, 2010.
49. Sandra Kübler, Erhard W. Hinrichs, and Wolfgang Maier. Is it really that difficult to parse German? In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 111–119, Sydney, Australia, 2006.
50. Sandra Kübler, Wolfgang Maier, Ines Rehbein, and Yannick Versley. How to compare treebanks. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2322–2329, Marrakech, Morocco, 2008.
51. Sandra Kübler, Ines Rehbein, and Josef van Genabith. TePaCoC — a corpus for testing parser performance on complex German grammatical constructions. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories*, pages 15–28, Groningen, The Netherlands, 2009.
52. Sandra Kübler and Heike Telljohann. Towards a dependency-based evaluation for partial parsing. In *Proceedings of the LREC-Workshop Beyond PARSEVAL–Towards Improved Evaluation Measures for Parsing Systems*, pages 9–16, Las Palmas, Gran Canaria, 2002.
53. Claudia Kunze and Lothar Lemnitzer. Germanet – representation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, 2002.
54. Wolfgang Lezius. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. PhD thesis, Universität Stuttgart, 2002. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4.
55. Scott Martens. TüNDRA: A web application for treebank search and visualization. In *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT)*, Sofia, Bulgaria, 2013.
56. Andreas Mengel and Wolfgang Lezius. An xml-based representation format for syntactically annotated corpora. In *In Proceedings of the International Conference on Language Resources and Evaluation*, pages 121–126, 2000.
57. Paul Meurer, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard, and Martha Thunes. The INESS treebanking infrastructure. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, number 16 in NEALT Proceedings, Olso, Norway, 2013.
58. Detmar Meurers and Stefan Müller. Corpora and syntax. In Anke Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 920–933. Mouton de Gruyter, 2009.
59. Frank H. Müller. Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, 2004.
60. Karin Naumann. Manual for the annotation of in-document referential relations. Universität Tübingen, 2007.
61. Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic, 2007.
62. Costantin Orasan. PALinkA: A highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39–43, Sapporo, Japan, 2003.

63. Sandra Pappert, Johannes Schließer, Dirk Janssen, and Thomas Pechmann. Corpus- and psycholinguistic investigations of linguistic constraints on German object order. In Andreas Späth, editor, *Interfaces and Interface Conditions*, pages 299–328. Mouton de Gruyter, 2007.
64. Oliver Plaehn. Annotate: Bedienungsanleitung. Technical Report, Universität des Saarlandes, http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate-manual.ps.gz, 1998.
65. Oliver Plaehn. Probabilistic parsing with discontinuous phrase structure grammar. Master's thesis, Department of Computational Linguistics, University of the Saarland, Saarbrücken, Germany, 1999.
66. Oliver Plaehn and Thorsten Brants. Annotate — an efficient interactive annotation tool. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-2000).*, Seattle, WA, 2000.
67. Ines Rehbein and Josef van Genabith. Treebank annotation schemes and parser evaluation for German. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 630–639, Prague, Czech Republic, 2007.
68. Ines Rehbein and Josef van Genabith. Why is it so difficult to compare treebanks? TIGER and TüBa-D/Z revisited. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT)*, Bergen, Norway, 2007.
69. Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. Masking treebanks for the free distribution of linguistic resources and other applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT)*, Bergen, Norway, 2007.
70. Marga Reis. Zum Subjektbegriff im Deutschen. In Werner Abraham, editor, *Satzglieder im Deutschen: Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung*, pages 171 – 211. Tübingen: Narr, 1982.
71. Christian Rohrer and Martin Forst. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2006)*, Genoa, Italy, 2006.
72. Laurent Romary, Amir Zeldes, and Florian Zipser. <tiger2/> — Serialising the ISO SynAF syntactic object model. *Language Resources and Evaluation*, to appear.
73. Victoria Rosén, Paul Meurer, and Koenraad De Smedt. LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories*, pages 127–133, Utrecht, 2009.
74. Adam Roussel. Documentation of the tool TIGER Tree Enricher. http://www.linguistics.ruhr-uni-bochum.de/resources /software/tte, 2014.
75. Yvonne Samuelsson and Martin Volk. Automatic node insertion for treebank deepening. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT)*, pages 127–136, Tübingen, 2004.
76. Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset), 1999. Technical report, Universität Stuttgart, Universität Tübingen, http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf.
77. Wolfgang Seeker and Jonas Kuhn. Making ellipses explicit in dependency conversion for a German treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3132—3139, Istanbul, Turkey, 2012.
78. Stefanie Simon, Erhard Hinrichs, Sabrina Schulze, and Yannick Versley. Handbuch zur Annotation expliziter und impliziter Diskursrelationen im Korpus der Tübinger Baumbank des Deutschen (TüBa-D/Z). Universität Tübingen, 2011.
79. Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, pages 705–711, 1998.
80. Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC, 1997.

81. George Smith. A brief introduction to the TIGER Treebank, version 1, 2003. Technical report, Universität Potsdam, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/tiger_introduction.pdf.
82. George Smith. Searching for morphological structure with regular expressions, 2003. Technical report, Universität Potsdam, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/tiger_regex.pdf.
83. Kathrin Spreyer and Anette Frank. The TIGER 700 RMRS Bank: RMRS construction from dependencies. In *Proceedings of LINC 2005*, pages 1–10, Jeju Island, Korea, 2005.
84. Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the ACL-04 Workshop on Discourse Annotation,*, Barcelona, 2004.
85. Ilona Steiner. Partial agreement in German: A processing issue? In *Proceedings of the International Conference on Linguistic Evidence*, Tübingen, Germany, 2009.
86. Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2229–2235, Lisbon, Portugal, 2004.
87. Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany, 2012.
88. Christine Thielen and Anne Schiller. Ein kleines und erweitertes Tagset fürs Deutsche. In Helmut Feldweg and Erhard Hinrichs, editors, *Lexikon & Text*, pages 193–203. Tübingen: Niemeyer, Tübingen, 1994.
89. Julia Trushkina. *Morpho-Syntactic Annotation and Dependency Parsing of German*. PhD thesis, Universität Tübingen, 2004.
90. Tylman Ule. *Treebank Refinement: Optimising Representations of Syntactic Analyses for Probabilistic Context-Free Parsing*. PhD thesis, Universität Tübingen, 2007.
91. Jorn Veenstra, Frank Henrik Müller, and Tylman Ule. Topological fields chunking for German. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL 2002)*, pages 56–62, Taipei, Taiwan, 2002.
92. Yannick Versley, Kathrin Beck, Erhard Hinrichs, and Heike Telljohann. A syntax-first approach to high-quality morphological analysis and lemma disambiguation for the TüBa-D/Z Treebank. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT)*, Tartu, Estonia, 2010.
93. Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank, 2010. An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments, http://www.cl.uzh.ch/research/parallelcorpora/paralleltreebanks_en.html.
94. Wolfgang Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin et al.: Springer, 2000.
95. Sina Zarrieß, Aoife Cahill, and Jonas Kuhn. To what extent does sentence-internal realisation reflect discourse context? A study on word order. In *Proceedings of the 13th Conference of the European Chapter of the ACL*, pages 767–776, Avignon, France, 2012.
96. Amir Zeldes, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*, Liverpool, UK, 2009.
97. Heike Zinsmeister. Treebank data as linguistic evidence? Coordination in TüBa–D/Z. In *Proceedings of the International Conference on Linguistic Evidence*, Tübingen, Germany, 2006.
98. Heike Zinsmeister, Jonas Kuhn, and Stefanie Dipper. Utilizing LFG parses for treebank annotation. In *Proceedings of the LFG-02 Conference*, pages 427–447, Athens, Greece, 2002. CSLI Publications.