Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank

Erhard Hinrichs, Sandra Kübler, Karin Naumann,
Heike Telljohann, Julia Trushkina
Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 19
72074 Tübingen, Germany
{eh, kuebler, knaumann}@sfs.uni-tuebingen.de,
{hschulz, jul}@sfs.uni-tuebingen.de

1 Introduction

The purpose of this paper is to describe recent developments in the morphological, syntactic, and semantic annotation of the TüBa-D/Z treebank of German.

The TüBa-D/Z annotation scheme is derived from the Verbmobil treebank of spoken German [4, 10], but has been extended along various dimensions to accommodate the characteristics of written texts. TüBa-D/Z uses as its data source the 'die tageszeitung' (taz) newspaper corpus.

The Verbmobil treebank annotation scheme distinguishes four levels of syntactic constituency: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields, which characterize the word order regularities among different clause types of German, and which are widely accepted among descriptive linguists of German [3, 6]. The TüBa-D/Z annotation relies on a context-free backbone (i.e. proper trees without crossing branches) of phrase structure combined with edge labels that specify the grammatical function of the phrase in question. The syntactic annotation scheme of the TüBa-D/Z is described in more detail in [12, 11].

TüBa-D/Z currently comprises approximately 15 000 sentences, with approximately 7 000 sentences being in the correction phase. The latter will be released along with an updated version of the existing treebank before the end of this year. The treebank is available in an XML format, in the NEGRA export format [1] and in the Penn treebank bracketing format. The XML format contains all types of

information as described above, the NEGRA export format contains all sentence-internal information while the Penn treebank format includes only those layers of information that can be expressed as pure tree structures.

Over the course of the last year, more fine grained linguistic annotations have been added along the following dimensions: 1. the basic Stuttgart-Tübingen tagset, STTS, [9] labels have been enriched by relevant features of inflectional morphology, 2. named entity information has been encoded as part of the syntactic annotation, and 3. a set of anaphoric and coreference relations has been added to link referentially dependent noun phrases. In the following sections, we will describe each of these innovations in turn and will demonstrate how the additional annotations can be incorporated into one comprehensive annotation scheme.

2 Morphological Annotation

The STTS [9] provides the widely accepted inventory of part of speech (POS) categories for German. Its basic tagset distinguishes 54 POS labels but does not provide information about inflectional morphology, which is a necessary prerequisite for many natural language applications, such as, for example, the recognition of grammatical functions in German [13]. In order to incorporate such morphological information, the treebank annotation scheme has been enriched by morphological features such as *case, number, person, gender, tense*, and *mood*.

For each lexical token which exhibits inflectional morphology, a relevant combination of feature-value pairs has been assigned. Thus, for example, nouns have received information on case, number, and gender, finite verbs are annotated with person, number, mood, and tense information. A complete list of POS tags which have been assigned morphological features as well as feature combinations associated with each part of speech are provided in Table 1.

Lexical tokens	**Feature Combination**
nouns, adjectives, determiners, non-personal pronouns, prepositions with incorporated articles	case, number, gender
prepositions, postpositions	case
personal pronouns	case, number, gender, person
finite verbs	person, number, mood, tense
imperative verbs	person, number
truncated words	number, gender

Table 1: Feature combinations for lexical tokens in TüBa-D/Z.

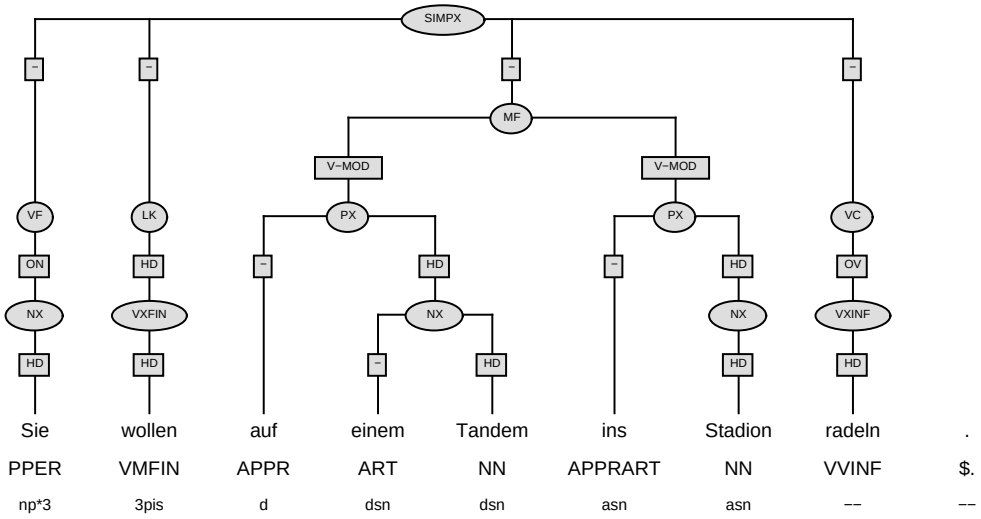

Figure 1: A morphologically annotated tree.

The tree in Figure 1 illustrates the annotation of the morphological information in the treebank for the sentence in example (1). Values of morphological features are presented in the treebank explicitly on the level below the level of lexical tokens. Features that correspond to the values can be uniquely identified by a position of a value in a cluster, given the POS tag. Thus, a cluster `3pis` assigned to the verb *wollen* in Figure 1 stands for "**3**rd person, **p**lural number, **i**ndicative mood, pre**s**ent tense". The order of features in the morphological cluster corresponds to the order in Table 1. Possible values for each feature are presented in Table 2. Apart from specific features such as *masculine* or *singular*, values for case, gender, and number features include an underspecified value. The underspecified value is used for the annotation of tokens if an appropriate concrete value cannot be recovered for a morphological feature. Typical examples of the use of an underspecified value are plural pronouns, such as *sie* (*they*) in Figure 1 or first person pronouns, such as *ich* (*I*). In both cases, gender cannot be determined.

(1) Sie wollen auf einem Tandem ins Stadion radeln.
They want to on a tandem into the stadium bike.
' They want to bike into the stadium on a tandem.'

In total, 433 distinct morphological value clusters can be generated. Combined with POS information, they result in a tagset of 1 317 tags. The number of actual tags which occur in the treebank amounts to 555 tags.

Features in TüBa-D/Z	**Values**
case	n (nominative), g (genitive), d (dative), a (accusative), * (underspecified)
gender	m (masculine), f (feminine), n (neutral), * (underspecified)
number	s (singular), p (plural), * (underspecified)
mood	i (indicative), k (subjunctive)
person	1 (first), 2 (second), 3 (third)
tense	s (present), t (past)

Table 2: Set of feature values in TüBa-D/Z.

Currently, approximately 13 000 trees have been enriched with morphological information. Annotation was performed semi-automatically by using the rule-based morphological disambiguator of Hinrichs and Trushkina [5] as a pre-filtering module that limits the number of candidate analyses for each lexical token to those that are contextually valid. This rule-based disambiguation greatly reduces the number of analyses from an overall ambiguity rate of 5.8 analyses to 1.91 analyses per token and by providing full disambiguation for 70% of all tokens. As a result, the human annotators have to consider a much smaller set of analyses, which significantly speeds up the annotation process.

The morphologically annotated treebank data have in turn been used for the training of hybrid models of morphological disambiguation that combine rule-based and statistical disambiguation [13].

3 Named Entities

For a variety of NLP applications, the robust annotation of named entities is an important prerequisite. To facilitate the use of the TüBa-D/Z data for such tasks, the level of named entity annotation has been added to the annotation scheme. This additional layer of annotation is conservative and monotonic in the following sense: It respects all syntactic boundaries that have been imposed on the elements of named entity expressions by existing layers of syntactic annotation. Named entity annotation thus amounts to mere insertion of an intermediate level of representation. At the same time, named entity annotation is fully compliant with the STTS labeling assigned to the elements of named entity expressions. These two constraints on named entity annotation ensure that it can be easily removed if such information is

irrelevant for the task to which the treebank is to be applied.

Named entities are annotated on the morpho-syntactic level via the STTS tags and/or on the syntactic level. The STTS tagset uses the label NE for proper names and NN for common nouns. The classification of NE in the STTS guidelines comprises specific categories (e.g. first name, last name, names of companies, geographical names). By contrast, categories like names of products or compounds which consist of NE + NN are POS-tagged as NN. Moreover, complex German names have to be POS-tagged according to their distribution.

Named entities either occur as single names consisting of one lexical element or as complex names consisting of phrases or sentences. Complex names are annotated on the syntactic level by the label EN-ADD or the secondary edge EN, single elements are either marked on the morpho-syntactic level as NE or they receive the label EN-ADD.

Figure 2 gives an example of the annotation of named entities for the sentence in example (2). Here, the two person names are marked as names in the POS tags NE and as complex names by the label EN-ADD, the movie title is marked by the label EN-ADD. The geographical name within the movie title is POS-tagged as NE.

(2) Seit “Schlaflos in Seattle” gelten Tom Hanks und Meg Ryan als
Since “Sleepless in Seattle” pass Tom Hanks and Meg Ryan for
Dream-Team des Biedersinns.
dream team of petty bourgeois mentality.
’Since “Sleepless in Seattle” Tom Hanks and Meg Ryan are said to be the dream team of petty bourgeois mentality.’

In the treebank, the following classes of named entities exist:

1. Names consisting of one lexical element: They are POS-tagged as NE if they belong to one of the categories of proper names defined in the STTS guidelines. Otherwise, they are POS-tagged according to their distribution and assigned the additional node label EN-ADD. For example, nouns which are names of products (“Opel” NN) or compounds which consist of NE + NN like names of streets or places ("Sögestraße" NN), institutions ("Zeit-Stiftung" NN), or events ("Golfkrieg" NN).

2. Complex names consisting of more than one lexical element, each of them POS-tagged as NE: This class comprises complex names of persons (e.g. "Hans Taake") and foreign language material which can be recognized as a proper name (e.g. "New York", "Karel van Miert", "Tour de France"). All of them are assigned the additional node label EN-ADD.

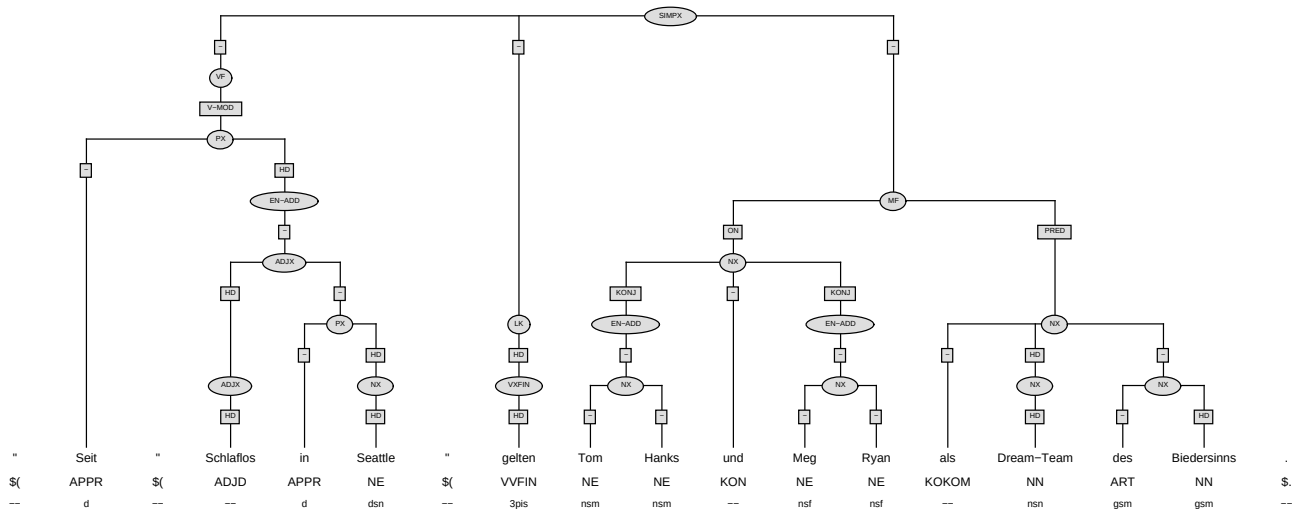

Figure 2: A tree containing named entities.

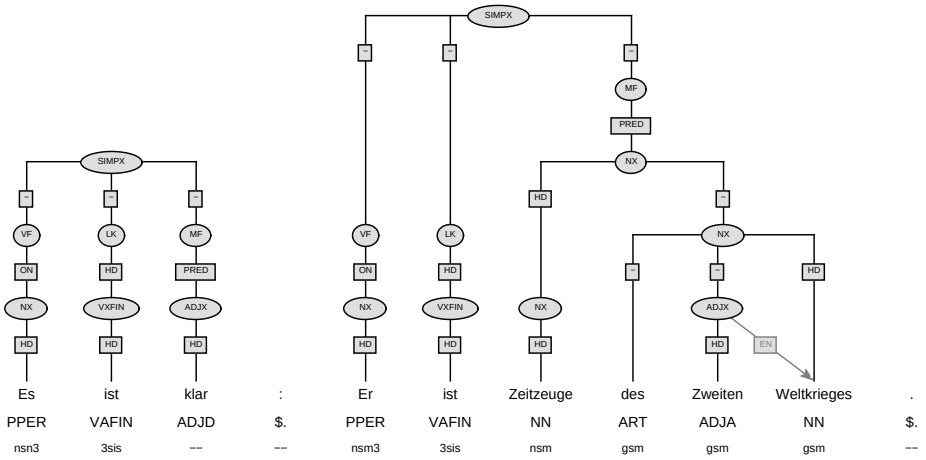

Figure 3: A tree containing a phrase internal named entity.

3. Complex names which are POS-tagged according to their distribution: titles, institutions, events, etc. (e.g. "Schlaflos in Seattle", "Zweiter Weltkrieg"). They are either assigned the additional node label EN-ADD or the secondary edge label EN.

The labels EN-ADD and EN are general markers of named entities, which have no syntactic function. Thus, they do not effect the syntactic structure if they are deleted. The internal structure of named entities is always governed by the general annotation rules, which allows recursive structure (named entities within named entities).

EN-ADD is inserted between two nodes to indicate that the node below represents a named entity. It is either directly attached to a phrase or a field. If this named entity has a pre- or postmodifier, its mother node is NX which represents the nominal status of the named entity.

The secondary edge label EN is used when the insertion of EN-ADD would cause a change of the syntactic structure. It gives information about the relation between two parts of a named entity within a complex phrase. The named entity is premodified, for instance, by an article and/or an attributive adjective which do not belong to the named entity itself (e.g. "vor den zweiten [Deutschen Existenzgründertagen]"), and may also be postmodified by an element which is part of the named entity (e.g. "das [Bundesinstitut für Arzneimittel]"). EN always points from the dependent part to the head noun of te named entity.

Figure 3 gives an example of a phrase internal named entity ("Zweiten Weltkrieges") in the sentences in example (3). The article ("des") is no part of

the named entity itself.

(3) Es ist klar: Er ist Zeitzeuge des Zweiten Weltkrieges.
It is clear: He is contemporary witness of the Second World War.
'It is clear: He is a contemporary witness of World War II.'

Preliminary experiments have shown that the inclusion of named entity annotation improves parsing accuracy of statistical parsers trained on the TüBa-D/Z data.

4 Anaphoric and Coreference Relations

Due to its fine grained syntactic annotation, the TüBa-D/Z data are ideally suited as a basis for the identification of markables, i.e. the set of potential anaphoric and other contextually dependent expressions referring to a nominal or pronominal antecedent. The annotation of anaphoric and coreference relations is thus a natural extension to the existing annotation scheme. In this context, the potential markables are definite NPs, personal pronouns, relative, reflexive, and reciprocal pronouns, demonstrative, indefinite and possessive pronouns as well as possessive adjectives. Compared to other annotation efforts in this area where markables have to be chosen manually, the actual manual annotation in the case of TüBa-D/Z can be restricted to the selection of the appropriate linking relations between referentially dependent expressions and their nominal antecedents. The inventory of such relations is inspired by the annotation scheme first developed in the MATE project [2] and uses the following subset of relations: *coreferential, anaphoric, cataphoric, bound, part-of, instance,* and *expletive*. Following van Deemter and Kibble [14], we define a coreference relation to hold between two NPs just in case they refer to the same extra-linguistic referent in the real world. In the following example, a *coreference relation* exists between the noun phrases [1] and [2], and an *anaphoric relation* between the noun phrase [2] and the personal pronoun [3].

(4) [1 Der neue Vorsitzende der Gewerkschaft Erziehung und Wissenschaft]
The new chairman of the union Education and Science
heißt [2 Ulli Thöne]. [3 Er] wurde gestern mit 217 von 355
is called Ulli Thöne. He was yesterday with 217 out of 355
Stimmen gewählt.
votes elected.
'The new chairman of the union of educators and scholars is called Ulli Thöne. He was elected yesterday with 217 of 355 votes.'

Cataphoric relations hold between a preceding pronoun and its antecedent within the same sentence, even if this antecedent has already been mentioned within the preceding text. An example for a cataphoric relation is shown in (5).

(5) Vier Wochen sind [sie] nun schon in Berlin, [die 220 Albaner aus
Four weeks are they now already in Berlin, the 220 Albanians from
dem Kosovo].
the Kosovo.
'They have already been in Berlin for four weeks, the 200 Albanians from Kosovo.'

The relation *bound* holds between anaphoric expressions and quantified noun phrases as their antecedents (see example (6)).

(6) [Niemandem] fällt es schwer, das Bild vor [sich] zu sehen.
To nobody is it difficult, the picture in front of himself to see.
'Nobody has trouble imagining the picture.'

The *part-of relation* holds between coordinate NPs/plural pronouns and pronouns/definite NPs referring to one member of the plural expression.

(7) [Ein paar andere Fehler] hat er aber schon vorher gemacht. [Den
A few other errors has he however already before made. The
ersten] Ende des vergangenen Jahres.
first end of the previous year.
'He had however already made a few other mistakes. The first one at the end of the previous year.'

An *instance relation* exists between a preceding/following pronoun and its NP antecedent when the pronoun refers to a particular instantiation of the class identified by the NP.

(8) Die konservativen Kräfte warten ja nur darauf, ihm [Sätze] um
The conservative powers wait just only for that, him sentences around
die Ohren zu hauen wie [jenen von den 16 Mittelstrecklern],
the ears to hit like the one about the 16 middle-distance runners,
denen er in vier Wochen die Viererkette beibringe.
to whom he in four weeks the double full-back formation teaches.
'The conservative powers are just waiting to bombard him with sentences like the one about the 16 middle-distance runners who he is teaching the double full-back formation in four weeks.'

The impersonal third person sg. pronoun ES (IT) is marked as *expletive* only if it has no proper antecedent, which is the case for presentational ES in example (9), impersonal passive as in example (10) or ES as subject for verbs without an agent as in example (11).

(9) [1 Es] zeichnet sich die konkrete Möglichkeit ab.
It emerges the concrete possibility *verb part.*
'The concrete possibility emerges.'

(10) [Es] wird bis zum Morgen getanzt.
There is until the morning danced.
'People are dancing until morning.'

(11) [Es] steht schlecht um ihn.
It stands bad for him.
'He is in a bad way.'

The annotation of such relations is performed manually with the annotation tool MMAX [8]. Its graphical user interface allows for easy selection of the relevant markables and the accompanying relation between the contextually dependent expression and its antecedent. In a first step, the relevant markables receive an attribute value: coreferential, anaphoric, cataphoric, bound, part-of, instance, or expletive. Second, the relation between a contextually dependent expression and its antecedent is established, except for the attribute "expletive", which is not related to an antecedent in the text. MMAX distinguishes between to kinds of relations: a set relation is defined as a transitive undirected relation. A pointer relation, in contrast, is intransitive and directed. Expressions marked by the attribute "coreferential", "anaphoric", "cataphoric" or "bound" share a set relation with their antecedent. Expressions marked by the attribute "part-of" and "instance" share a pointer relation with their antecedent.

The resulting annotation is converted into the Annotate export format [1] and the XML format, in which the treebank is available[1].

[1]For licensing information please visit the webpage `http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml`.

5 Conclusion

This paper presents three recent additions to the previous layers of annotation in the TüBa-D/Z, which significantly enhance the usability of the treebank for NLP applications. While each addition is independently motivated, it is important to note that the new information could be incorporated into the existing annotation scheme attesting to the flexibility and open architecture of the annotation scheme.

References

[1] Thorsten Brants. *The NeGra Export Format for Annotated Corpora*. Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany, 1997.

[2] Sarah Davies, Massimo Poesio, Florence Bruneseaux, and Laurent Romary. *Annotating Coreference in Dialogues: Proposal for a Scheme for MATE*. MATE, 1998.

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

[4] 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. Springer, Berlin, 2000.

[5] Erhard W. Hinrichs and Julia Trushkina. Getting a grip an morphological disambiguation. In *Proceedings of KONVENS 2002, 6. Konferenz zur Verarbeitung natürlicher Sprache*, pages 59–66, Saarbrücken, Germany, 2002.

[6] 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.

[7] Claudia Kunze and Andreas Wagner. Integrating GermaNet into EuroWordNet, a multilingual lexical-semantic database. *Sprache und Datenverarbeitung. International Journal for Language Data Processing*, 23(2):5–19, 1999.

[8] Christoph Müller and Michael Strube. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 2003.

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

[10] Rosmary Stegmann, Heike Telljohann, and Erhard W. Hinrichs. Stylebook for the German Treebank in VERBMOBIL. Technical Report 239, Verbmobil, 2000.

[11] 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 2004)*, Lisbon, Portugal, 2004.

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

[13] Julia Trushkina and Erhard W. Hinrichs. A hybrid model for morpho-syntactic annotation of German with a large tagset. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 238–246, Barcelona, Spain, 2004.

[14] Kees van Deemter and Rodger Kibble. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(2):629–637, 2000.