The Verbmobil Treebanks[*]

Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, Heike Telljohann
Eberhard-Karls-Universität Tübingen • Germany

Abstract

The Verbmobil treebanks of spoken German, English, and Japanese are part of the Verbmobil project, which has the overriding goal to develop a speaker-independent system for the translation of spontaneous speech. In the framework of this language technology project, the treebanks provide training data for a variety of language technology modules. The treebanks consist of annotated syntactic tree structures based on transcribed dialogs in the scenarios of appointment negotiations, travel arrangements, and personal computer maintenance. The annotation schemes of the treebanks have been developed taking into account the specific characteristics of spoken language dialogs: repetitions, hesitations, "false starts", etc.

1 Introduction

The Verbmobil treebanks of spoken German, English, and Japanese are part of the Verbmobil project, which has the overriding goal to develop a speaker-independent system for the translation of spontaneous speech. In the framework of this language technology project, the treebanks provide training data for a variety of language technology modules, including the transfer component for machine translation and stochastic parsers.

The annotated trees are based on data transcribed from spoken language dialogs of the following scenarios: appointment negotiation, travel planning, hotel reservation, and personal computer maintenance. The linguistic annotations pertain to the levels of morphosyntax (part-of-speech tagging), syntactic phrase structure, and function-argument structure. The classificatory labels at each level of annotation are based upon a minimal set of assumptions concerning constituenthood, phrase attachment, and grammatical functions, which can be considered uncontroversial among major syntactic theories. In this respect, the annotations can be regarded as theory-neutral, which facilitates the reusability of the annotated data for empirical, linguistic investigations and language technology applications alike.

The linguistic annotation is performed semi-automatically with the help of the graphical annotation tool *Annotate* [1], [9], which was developed in the NEGRA project of the SFB 378 at the Universität des Saarlandes. Compared to entirely manual treebank construction, semi-automatic annotation can help to reduce the number of inconsistencies and annotation errors that will inevitably arise in any treebank of significant size. This semi-automatic method of annotation differs also from the one used in the Penn Treebank, for instance, where human correction succeeds the fully automatic parsing. Apart from providing a user-friendly graphical interface for annotating and editing trees, the *Annotate* tool offers database support, which is indispensible for maintaining large treebanks.

One of the major goals of our annotation efforts is to achieve the highest possible degree of consistency within the annotated corpora. Thus, the annotation scheme has to be made explicit, precise, and unambiguous to be easily assimilated by human annotators. To this end, detailed stylebooks [7], [8], [11] were developed at the outset of the project. Computational consistency checks were conducted at regular intervals throughout the annotation phase as a means of quality control. In addition, the semi-automatic annotation mode provided by the *Annotate* tool supports the annotation process.

2 Linguistic Annotation

2.1 Coping with Spontaneous Speech

The annotated trees are based on data transcribed from spoken language dialogs of the following scenarios: appointment negotiation, travel planning, hotel reservation, and personal computer maintenance. In contrast to written language, the segmentation of

[*] The work reported here was funded by the German Ministry of Education and Research (BMBF) in the framework of the Verbmobil project under grant FKZ:01 IV 701 M0.

spontaneous speech utterances into sentences provides interesting challenges. The specific characteristics of spoken language dialogs have to be taken into account: repetitions, hesitations, "false starts", etc. For this reason, the *dialog turn*, which consists of one or more sentences and/or phrases and which denotes an uninterrupted contribution by one dialog participant, has been defined as the primary domain of syntactic analysis and annotation.

Since the Tübingen treebanks are based exclusively on spontaneous speech data, a number of research questions have to be taken into consideration that are different from those concerning written data. In contrast to written language, in which the segmentation into sentences coincides with the domain of syntactic analysis, corpora of spontaneuous speech utterances pose interesting research questions in this regard. To be able to cope with the specific characteristics of spoken language (speech errors, fragmentary utterances, "false starts", repetitions, interruptions, and hesitation noises), the dialog turn has been defined as the primary segmentation domain of the Verbmobil dialogs. The dialog turns are preprocessed into syntactic units delimited by full stops and question marks, thus forming a secondary domain of analysis. These units themselves may consist of one or more sentences in the grammatical sense and/or phrases. It is the task of the human annotator to perform the segmentation of these units, to classify the segments, and to construct the syntactic trees, which capture syntactic and semantic dependency relations (predicate-argument structure).

2.2 Annotation Format and Annotation Principles

The treebanks for each of the three languages consist of sets of tree diagrams that represent the syntactic structure of a transcribed utterance. Each tree consists of a set of terminal symbols (words), a set of preterminal symbols (parts of speech), and a set of nonterminal symbols drawn from a fixed inventory of syntactic categories. Each local syntactic tree is further annotated by an edge label that indicates its grammatical function.

Constituents are grouped by three overriding principles of annotation. The *longest match principle* demands that as many daughter nodes as possible are combined into a single mother node, provided that the resulting constituent is syntactically as well as semantically well-formed. The *flat clustering principle* keeps the number of hierarchy levels in a syntactic structure as small as possible. As a consequence, any degree of branching is allowed. Speech errors, repetitions, corrections, and hesitations are structured as much as possible (mostly up to the level of phrasal categories), but are not typically connected to surrounding constituents as a whole. In cases of attachment ambiguities (e.g. of prepositional phrases for which more than one constituent can serve as a semantically plausible attachment host), the *high attachment principle* prescribes that such ambiguous modifiers are attached at the highest possible level in a tree structure.

3 The German Treebank

The size of the German treebank consists of more than 38,000 fully annotated syntactic units in the sense of section 2.1. The annotated data covers all dialogs collected during the Verbmobil-II (1997-2000) phase of the project and 10,000 units from the Verbmobil-I (1993-96) dialog corpus.

3.1 The Theoretical Basis of the Annotation Scheme

For the development of the annotation scheme for the German treebank, the characteristics of the German language have been taken into account. The partially free word order in German is responsible for the interaction of configurational and nonconfigurational syntactic properties. Three different clause types are distinguished with respect to the fixed position of the finite verb in a sentence: verb-second (V-2), verb-initial (V-1), and verb-final (V-end). At the same time, there is a high degree of variability concerning the positions of complements and adjuncts.

3.1.1 Topological Fields and Constituent Structure

In order to capture the fundamental word order regularities of German sentence structure, the annotation scheme for the German treebank adopts the notion of topological fields in the sense of Herling [4], Erdmann [3], Drach [2], and Höhle [6] as the primary clustering principle of a German sentence. The topological model provides only descriptive parameters concerning sentence structure without making any statement about the regularities within the fields and the hierarchical constituent structure of the sentence.

To integrate a constituent analysis, we established a second level of annotation strictly within the bounds of topological fields: a phrase level of predicate-argument structure with its own descriptive inventory of syntactic categories and grammatical functions. This scheme facilitates a theory-neutral and surfaceoriented representation of syntactic trees without crossing branches and traces. Long-distance dependencies are denoted by special naming conventions for edge labels. In general, we distinguish four levels of annotation within a German annotated syntactic tree, which are listed in **Table 1**:

Level	**Inventory**
sentence level	root node labels for different types of sentences
field level	node labels for topological fields
phrase level	node labels for syntactic categories and edge labels for grammatical functions
lexical level	lexical entries tagged with the part-of-speech (POS) tags taken from the STTS tagset [10]

Table 1 Four levels of annotation

3.1.2 The Inventory of Labels

Table 2 and **Table 3** list the complete descriptive inventory of labels denoting syntactic categories and grammatical functions used in the German treebank. Node labels (cf. Table 2) indicate the syntactic category of phrases or sentences as well as topological fields and combinations of topological fields within coordinations.

Edge labels (cf. Table 3) denote the grammatical function of lexical entries, phrases, topological fields, and sentences. Since case information is given and a distinction of unambiguous modifiers is made for some of these labels, German tree structures are also enriched by semantic roles.

Node Labels	**Description**
Root Node Labels	
SIMPX	simplex clause
R-SIMPX	relative clause
P-SIMPX	paratactic construction of simplex clauses
DM	discourse marker
Field Conjunct Node Labels	
LKM, LKMVC, LKMVCN, LKMN, LKVCN, LKN, MVC, MVCN, MN, VCN, CM, CMVC	combinations of fields, node labels are derived by concatenation of conjunct field labels (V = VF, M = MF, N = NF) e.g. LKM = LK + MF
Topological Field Node Labels	
LV	resumptive construction (Linksversetzung)
VF	initial field, contains only one constituent (Vorfeld)
LK	left sentence bracket, (Linke (Satz-)Klammer)
MF	middle field, may contain almost any constituent (Mittelfeld)
VC	verb complex (Verbkomplex)
NF	final field, one or more constituents (Nachfeld)
C	complementizer field, only verb-final clauses (C-Feld)
KOORD	field for coordinating particles, left-most element, in all sentence types possible
PARORD	field for coordinating particles, left-most element, only in V-2 (e.g. *denn, weil*)
FKOORD	coordination consisting of conjuncts of fields
Phrase Node Labels	
NX	noun phrase
PX	prepositional phrase
ADVX	adverbial phrase
ADJX	adjectival phrase
VXFIN	finite verb phrase
VXINF	infinite verb phrase
DP	determiner phrase (e.g. *gar keine*)
KONX	conjunction phrase/complex (*und zwar* in VF)

Table 2 Node labels

Edge Labels	Description
Edge Labels denoting Head	
HD	head
-	non-head
Complement Edge Labels	
ON	nominative object (subject)
OD	dative object
OA	accusative object
OS	sentential object
OPP	prepositional object
OADVP	adverbial object
OADJP	adjectival object
PRED	predicate
OV	verbal object
FOPP	optional prepositional object
VPT	separable verb prefix
APP	apposition
Modifier Edge Labels	
MOD	ambiguous modifier
ON-MOD, OA-MOD, OD-MOD, MOD-MOD, V-MOD, OPP-MOD, PRED-MOD, FOPP-MOD	unambiguous modifiers modifying complements or modifiers e.g. V-MOD = modifier of the verb
Edge Labels in Split-up Coordinations	
ONK, ODK, OAK, OPPK, FOPPK, OADJPK, PREDK, MODK, OA-MODK, V-MODK, OPP-MODK, PREDMODK, MOD-MODK	second conjunct in split-up coordinations e.g. ONK = second conjunct of a nominative object (subject)
Secondary Edge Labels	
ref1	first verbal object in VC selected by a verbal object

Table 3 Edge labels

3.1.3 Annotation Examples

Figure 1 shows an example of an annotated tree. The leaves of the tree consist of pairs of non-terminal symbols and part-of-speech tags. Non-terminal symbols are represented by spherical nodes, edge labels by rectangular nodes.

Figure 1 is an example of a grammatically well-formed sentence and an isolated phrase. In accordance with the four annotation levels shown in Table 1, the sentence is annotated top-down by the root node (SIMPX), the field nodes (KOORD, VF, LK, and MF), the phrase nodes (ADVX, VXFIN, and NX), and finally the tagged lexical entries. The edge labels between the field level and the phrase level indicate that the syntactic structure contains a subject (ON), an accusative object (OA), two ambiguous modifiers (MOD), and one unambiguous modifier (V-MOD) modifying the finite verb, which itself is the head (HD) of the entire syntactic construction.

The noun phrase *Samstag bis Montag* is not attached to the sentence structure, because otherwise the well-formedness of the construction would be violated. Thus, it has to be annotated as an isolated phrase lakking a verbal constituent.

As Figure 1 demonstrates, the topological model favors the *flat clustering principle* inasmuch the MF (and NF if applicable) allow for n-ary branching structures.

The annotation of complex phrases is also carried out following the *flat clustering principle* in order to keep the number of hierarchy levels in a syntactic structure as small as possible. **Figure 2** shows an example of a complex prepositional phrase (PX) including a premodifier (*ganz*) as well as a postmodifier (*vom Hauptbahnhof*). Both modifiers are projected to their phrase levels (ADVX and PX). Since the modification scope of premodifiers is unambiguous, they are directly attached to the head of the phrase which they are modifying. By contrast, postmodifiers are always attached on a higher level to preserve ambiguity. This decision, referred to in section 2.2 as the *high attachment principle*, was taken to avoid the problematic distinction whether a postmodifier is a free adjunct or a complement of the modified phrase.

If a modifying constituent is not adjacent to the modified constituent, their dependency relation, which can even go beyond the border of topological fields, is indicated by specific edge labels expressing the non-ambiguity of the modifier. Thus, the use of crossing branches is not necessary. In **Figure 3**, for instance, the noun phrase in the NF (OA-MOD) modifies the accusative object (OA) in the MF.

4 The English and Japanese Treebanks

The English treebank consists of appr. 30,000 annotated trees. This data covers all English dialogs collected during the Verbmobil-I (1993-96) and the Verbmobil-II (1997-2000) phase of the project. The size of the Japanese treebank is about 20,000 fully annotated trees. This is, to the best of our knowledge, the first large-scale treebank, in which the string level is transcribed into Roman characters. This representation will make the data internationally available to a wider group of researchers in the NLP community.

The syntactic annotation scheme and the constituent structures for the English and Japanese treebanks are inspired by the syntactic theory of Head-Driven Phrase Structure Grammar.

Space limitations prohibit more detailed descriptions of the English and Japanese treebank in the present paper. For more detailed information we refer interested readers to Hinrichs et al. [5] and the annotation stylebooks [7], [8] that were prepared for each language.

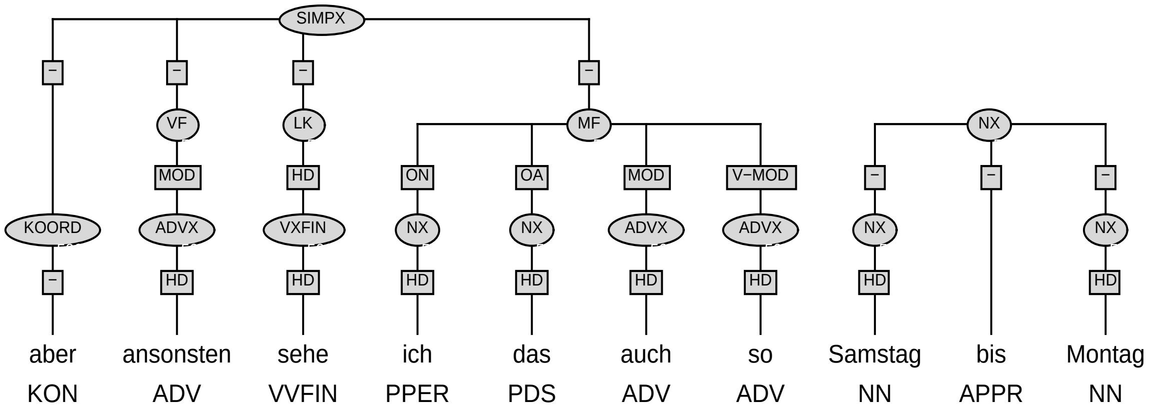
SIMPX
VF
LK
MF
NX
MOD
HD
ON
OA
MOD
V−MOD
KOORD
ADVX
VXFIN
NX
NX
ADVX
ADVX
NX
NX
HD
−
aber ansonsten sehe ich das auch so Samstag bis Montag
KON ADV VVFIN PPER PDS ADV ADV NN APPR NN

Figure 1 Annotated syntactic unit of the German treebank

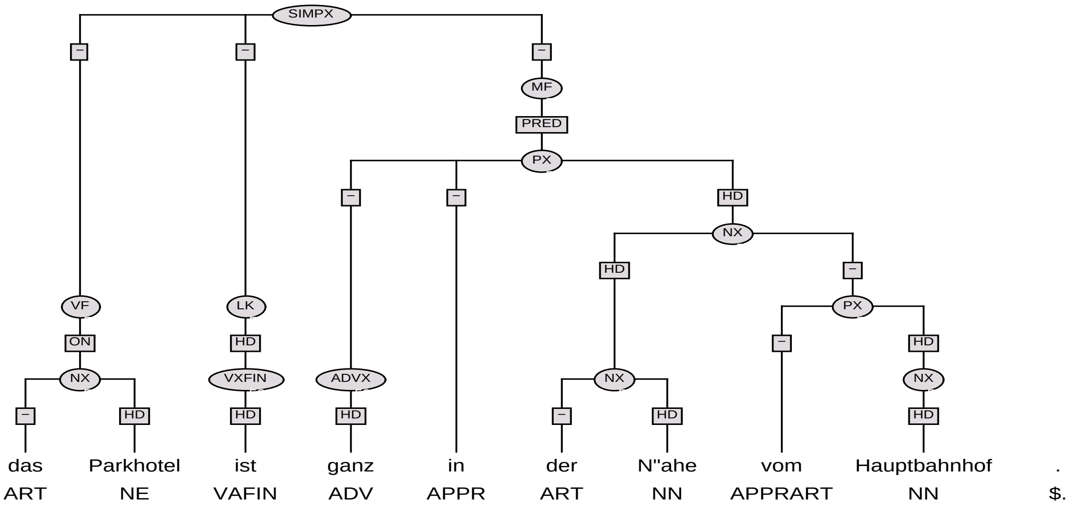
SIMPX
MF
PRED
PX
HD
NX
HD
PX
VF
ON
NX
LK
HD
VXFIN
ADVX
NX
HD
−
das Parkhotel ist ganz in der N"ahe vom Hauptbahnhof .
ART NE VAFIN ADV APPR ART NN APPRART NN $.

Figure 2 Premodification and postmodification in a complex phrase

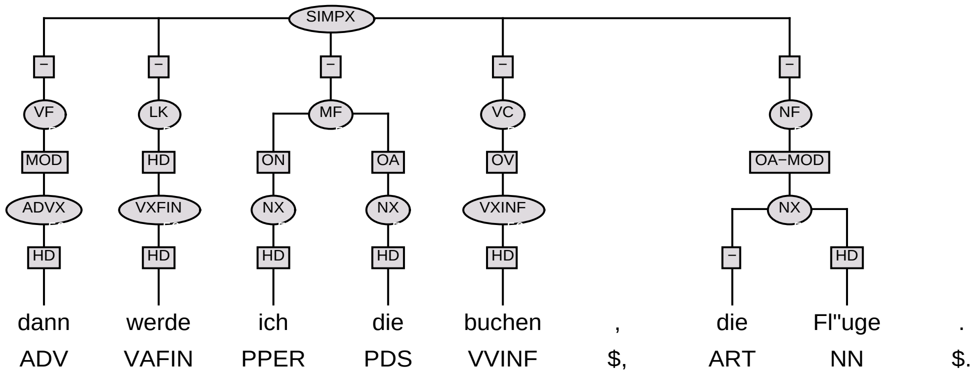

Figure 3 Long-distance dependency

5 Conclusion

To the best of our knowledge, the Verbmobil treebanks of German, English, and Japanese constitute the largest collection of annotated spoken language data currently available for these three languages. While the subject domain is limited to the scenarios of appointment negotiations, travel arrangements, and personal computer maintenance, great care has been taken to define general annotation schemes that are domain independent and theory-neutral. This generality of the annotation schemes should greatly facilitate the reusability of the treebank data for a wide variety of applications in theoretical and computational linguistics.

6 Bibliography

[1] Brants, T.; Skut, W.: Automation of treebank annotation. In: Proceedings of the Conference on New Methods in Language Processing (NeMLaP-3/CoNLL98), Sydney, Australia, 1998, pp. 49-57

[2] Drach, E.: Grundgedanken der Deutschen Satzlehre. Frankfurt/M., 1937

[3] Erdmann, O.: Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt. Stuttgart. Erste Abteilung, 1886

[4] Herling, S. H. A.: Über die Topik der deutschen Sprache. In: Abhandlungen des frankfurterischen Gelehrtenvereins für deutsche Sprache. Frankfurt/M., Drittes Stück, 1821, S. 296-362, 394

[5] Hinrichs, E. W.; Bartels, J.; Kawata, Y.; Kordoni, V.; Telljohann, H.: The Tübingen Treebanks for Spoken German, English, and Japanese. In: Verbmobil: Foundations of Speech-to-Speech Translations, Berlin: Springer Verlag, to appear, 2000

[6] Höhle, T. N.: Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder. In: Schöne, A. (Hrsg.): Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen, 1985, S. 329-340

[7] Kawata, Y.; Bartels, J.: Stylebook for the Japanese Treebank in VERBMOBIL. Technical report. Eberhard-Karls-Universität Tübingen, to appear, 2000

[8] Kordoni, V.: Stylebook for the English Treebank in VERBMOBIL. Technical report. Eberhard-Karls-Universität Tübingen, 1998

[9] Plaehn, O.: *Annotate* – Bedienungsanleitung, Universität des Saarlandes, FR 8.7 Computerlinguistik, Projekt C3 Nebenläufige Grammatische Verarbeitung, Sonderforschungsbereich 378, Ressourcenadaptive Kognitive Prozesse, 1998

[10] Schiller, A.; Teufel, S.; Thielen, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report. Universitäten Stuttgart und Tübingen, 1995

[11] Stegmann, R.; Telljohann, H.; Hinrichs, E. W.: Stylebook for the German Treebank in VERBMOBIL. Technical report. Eberhard-Karls-Universität Tübingen, 1998