

Stylebook for the German Treebank in **VERBMOBIL**

Rosmary Stegmann
Heike Telljohann
Erhard W. Hinrichs

Seminar für Sprachwissenschaft
Universität Tübingen

September 2000

Rosmary Stegmann
Heike Telljohann
Erhard W. Hinrichs

Computerlinguistik
Seminar für Sprachwissenschaft
Eberhard-Karls-Universität Tübingen
Wilhelmstr. 113
72074 Tübingen

Tel.: 07071 - 29 74279

Fax: 07071 - 55 13 35

e-mail: eh@sfs.nphil.uni-tuebingen.de

Gehört zum Antragsabschnitt: 6.7 Baubanken

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 701 M0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Abstract

This report describes the design principles and the annotation scheme for the VERBMÖBIL treebank of German developed at the Eberhard-Karls-Universität Tübingen. It is intended as stylebook for the construction or use of treebank data for German. The guidelines focus on the syntactic annotation of spoken language data with its characteristics (e.g. repetitions, hesitations, 'false starts'). To ensure the reusability of the data, a theory-neutral and surface-oriented annotation scheme has been adopted that is inspired by the notion of topological fields enriched by a level of predicate-argument structure. The linguistic inventory used in the treebank is based on a minimal set of assumptions that are uncontroversial among major syntactic theories.

We would like to thank Manfred Sailer and Frank Richter for their helpful comments and support in form of encouraging and critical discussions from which we could strongly benefit for the challenging task of developing a data-oriented syntactic annotation scheme for a planned amount of 30,000¹ database entries.

We would also like to acknowledge the support of our Tübingen colleagues Dale Gerdemann, John Griffith, Valia Kordoni, Sandra Kübler, and Uli Schatz for their assistance with part-of-speech tagging of the data and with data conversion. In addition, Valia Kordoni deserves special thanks for her willingness to take responsibility for the German treebank during a crucial phase of the project, when the first author of this report left Tübingen to accept a position in industry and when the second author of this report had not yet come on board.

The development of the Tübingen VERBMOBIL treebanks was greatly facilitated by a number of VERBMOBIL partners whose contributions went well beyond the call of duty. Hans Uszkoreit and his colleagues at the Universität des Saarlandes kindly provided us with the graphical annotation tool *Annotate* which was developed as part of the research project (*Teilprojekt C3*; Principal investigators: Uszkoreit/Smolka) *Nebenläufige grammatische Verarbeitung* (NEGRA) in the Sonderforschungsbereich 378. The *Annotate* tool provides human annotators with a graphical, user-friendly interface for annotating and editing trees and also provides database support for maintaining large treebanks. We would like to express our special gratitude to Thorsten Brants, who has kindly and generously provided us with software support and user assistance for the *Annotate* tool from the very beginning of the Tübingen treebank project.

For assistance with part-of-speech tagging and data conversion with the transcribed VERBMOBIL data we are indebted to our VERBMOBIL colleagues at Siemens (Munich), particularly to Tobias Ruland, and at the IMS Stuttgart, particularly to Martin Emele.

¹In the meantime, we have reached 38,000 entries.

Contents

1	Introduction	5
2	Major Challenges and Design Decisions	7
3	The Theoretical Basis of the Annotation Scheme	10
3.1	Topological Fields	10
3.1.1	The Concept of Topological Fields	10
3.2	Constituent Analysis and Topological Fields	14
3.3	General Annotation Principles	15
3.3.1	Flat Clustering Principle	15
3.3.2	Longest Match Principle	16
3.3.3	High Attachment Principle	16
3.4	The Structure of an Annotated Tree	16
3.4.1	The Levels of Annotation	16
3.4.2	The Inventory of Labels	17
3.4.3	Where Does a Tree End?	22
3.4.4	Speech Errors and Repetitions	24
3.4.5	Isolated Phrases and Sentence Fragments	25
3.4.6	Long-Distance Dependencies	28
3.4.7	Empty Categories	29
4	The Annotation of the Internal Structure of Phrases	31
4.1	Noun Phrases	31
4.1.1	Prenominal Modification	31
4.1.2	Postnominal Modification	34
4.1.3	Appositions	37
4.1.4	Proper Name Phrases	39
4.1.5	Complex Cardinal Numbers	42
4.1.6	Spelling	42
4.1.7	Expletive <i>es</i>	43

4.2	Determiner Phrases	44
4.3	Prepositional Phrases	45
4.3.1	Prepositions	45
4.3.2	Circumpositions and Postpositions	49
4.4	Adjectival Phrases	50
4.5	Adverbial Phrases	52
4.6	Verb Phrases	53
4.6.1	Head of a Sentence and Verb Complex	53
4.6.2	Infinitives with <i>zu</i>	55
4.6.3	Imperatives	57
4.6.4	Particle Verbs	58
4.6.5	Verbs with Predicate	59
4.6.6	Verbs with OA Marking the Border Between MF and NF	61
4.6.7	Modal Verbs	62
5	Attachment Principles for Phrases	63
5.1	Attachment to Fields	63
5.2	Modifier Attachment	63
5.2.1	Ambiguous Modifiers Occuring with Isolated Phrases	65
5.2.2	Phrases Modified by Phrases	67
6	The Annotation of Sentences	69
6.1	The C-Field in Verb-Final Clauses	69
6.2	The KOORD-Field in all Clause Types	73
6.3	The PARORD-Field in Verb-Second Clauses	74
6.4	Resumptive Constructions: The LV-Field	76
6.5	Questions	77
6.5.1	W-Questions	77
6.5.2	Yes - No Questions	78
6.6	Relative Clauses	79
6.7	Constructions with <i>aber</i>	80
6.8	Constructions with <i>wenn (-dann)</i>	82
6.9	Elliptic Constructions	83
6.10	Coordination	85
6.10.1	Coordination of Phrases	85
6.10.2	Specific Coordination Phenomena	86
6.10.3	Coordination of Sentences	90
6.10.4	Coordination of Topological Fields	92
6.10.5	Coordinations with Unequal Conjuncts	93
6.10.6	Conjunctions Occurring with Isolated Phrases	96

6.10.7 Split-up Coordinations	97
6.11 Paratactic Constructions	98
6.12 Parentheses	100
7 The Annotation of Specific Syntactic Phenomena	101
7.1 <i>und zwar</i> -Constructions	101
7.2 Superlative and Comparative Forms	102
7.2.1 The Comparative Particles <i>wie</i> and <i>als</i>	102
7.3 Verbal and Adjectival Use of Participles	104
7.4 Adjectival Use of Verbal Particles	105
7.5 Discourse Markers	106
8 Problematic Issues	108
8.1 Problems with Grammatical Functions	108
8.1.1 Distinguishing FOPP, OPP, and V-MOD	108
8.1.2 Distinguishing MOD, MOD-MOD, and V-MOD	109
8.1.3 Problems with the Distinction of ON, PRED, and ON-MOD	110
8.1.4 Problems with the Distinction of ON-MOD, ON, and V- MOD within LV Constructions	112
References	114

Chapter 1

Introduction

This report describes the design principles and the annotation scheme for a treebank of German that was developed at the Eberhard-Karls-Universität Tübingen as part of the VERBMOBIL project.

VERBMOBIL is a joint research project that is conducted by a consortium of universities, research centers, and information technology companies and that is funded by the German Ministry for Education and Research (BMBF). The initial four-year phase of the project (VERBMOBIL-I) lasted from 1993–96. The second phase of the project (VERBMOBIL-II) commenced in 1997 and concluded in September 2000.

The overriding goal of the VERBMOBIL project was to develop a speaker-independent system for translating spontaneous speech. To this end, a number of scenarios have been defined as a testbed for the development of software prototypes. During the first phase of the project (VERBMOBIL-I) the scenario consisted of dialogs, in which two discourse participants negotiate business appointments. In VERBMOBIL-II this scenario is significantly extended along various dimensions. In order to obtain realistic and quantitatively significant data for the relevant scenarios, a major data collection initiative for spoken-language dialogs was launched. The dialogs were recorded in a variety of settings and were transcribed according to mutually agreed upon standards. The transcribed data were then further annotated for the purposes of signal processing and linguistic analysis.

The treebank project, carried out by the Division of Computational Linguistics at the Eberhard-Karls-Universität Tübingen (Lehrstuhl Prof. Hinrichs), constitutes part of the overall effort of linguistic annotation within the VERBMOBIL project. Treebanks for German, English, and Japanese have been developed. The

present report focuses on the Tübingen treebank for German only; the parallel development of the Tübingen treebank for English is described in (Kordoni 2000), the one for Japanese in (Kawata and Bartels 2000).

The size for the German treebank has reached more than 38,000 entries at the conclusion of the VERBMOBIL-II project phase. Coverage of the German treebank includes a set of 10,000 units arbitrarily extracted from the data collected during the VERBMOBIL-I project phase and the collection of the German VERBMOBIL-II dialogs. The overall annotation scheme for the German treebank was negotiated with all relevant partners in the VERBMOBIL-II consortium.

The purpose of this report is to describe the design principles and annotation scheme for the Tübingen treebank of German. It is intended as a guide for the treebank annotators in Tübingen and for interested parties who will work directly with the annotated treebank data or who are interested in the construction or use of treebank data for German.

The development of the Tübingen VERBMOBIL treebanks was greatly facilitated by a number of VERBMOBIL partners whose contributions went well beyond the call of duty. Hans Uszkoreit and his colleagues at the Universität des Saarlandes kindly provided us with the graphical annotation tool *Annotate* (Plaehn 1998). This tool was developed as part of the research project (*Teilprojekt C3*; Principal investigators: Uszkoreit/Smolka) *Nebenläufige grammatische Verarbeitung* (NEGRA) in the Sonderforschungsbereich 378. The *Annotate* tool provides human annotators with a graphical, user-friendly interface for annotating and editing trees and also provides database support for maintaining large treebanks ((Brants and Skut 1998) and (Skut et al. 1997)). We would like to express our special gratitude to Thorsten Brants, who has kindly and generously provided us with software support and user assistance for the *Annotate* tool from the very beginning of the Tübingen treebank project.

For assistance with part-of-speech tagging and data conversion with the transcribed VERBMOBIL data we are indebted to our VERBMOBIL colleagues at Siemens (Munich), particularly to Tobias Ruland, and at the IMS Stuttgart, particularly to Martin Emele.

Chapter 2

Major Challenges and Design Decisions

To the best of our knowledge, the Tübingen treebank for German is the first German treebank that is based exclusively on spontaneous speech data. The focus on spoken language immediately raises a number of research questions that do not arise if the input data are taken from newspaper corpora or other sources of written data.

Most syntactic theories consider individual sentences as the primary domain of linguistic theorizing and of syntactic annotation. For written language, the segmentation into sentences is largely unproblematic and coincides with the domain of syntactic analysis.

For corpora of spontaneous speech utterances, such an immediate fit does not exist. In the case at hand, the corpus of VERBMOBIL spoken language dialogs, the primary segmentation is that of the dialog turn, which consists of a single, typically uninterrupted contribution to the dialog by one of the dialog participants. These dialog turns exhibit all the properties characteristic of spontaneous speech, which include fragmentary utterances, false starts, repetitions, interruptions, and hesitation noises. Due to such factors, it is in many cases far from clear how to segment dialog turns into individual sentences and how to annotate and attach fragmentary utterances (cf. section 3.4 for more details).

The second main question that needed to be addressed at the outset of the project was the inventory of syntactic categories and grammatical functions to be used for syntactic annotation and specification of predicate-argument structure. Here our choices were guided by three considerations:

1. Linguistic adequacy: The inventory of grammatical categories has to be applicable not only to the particular sublanguage of VERBMOBIL dialogs, but to spoken language data in general.

This desideratum is of utmost importance to ensure the reusability of the annotated data beyond the VERBMOBIL project itself. Since the size of the treebanks is significant, it opens up the possibility of using these data for a variety of research projects in both computational and theoretical linguistics.

2. Processing considerations: The primary use of the treebank data within VERBMOBIL is to provide training data for machine translation modules and for stochastic parsers. Thus, the annotation scheme should be sensitive to processing considerations, as long as linguistic adequacy of the choice of annotations is not compromised. *Ceteris paribus*, processing considerations favor annotation schemes that pay close attention to properties of syntactic surface structure, particularly to word order regularities and distributional properties of words and phrases. At the same time the use of empty categories and data structures with crossing dependencies among phrases are to be avoided if the annotations are to be used for parsers that rely on the context-freeness of the underlying grammar.

3. Theory-neutrality: For the purposes of reusability of the treebank data, the annotation scheme should not reflect a commitment to a particular syntactic theory. Rather, the inventory of categories should be a reflection of the minimal assumptions that syntacticians share concerning questions of constituenthood, phrase attachment, and grammatical functions. In this sense, the annotations should be theory-neutral and minimal.

In order to satisfy the three criteria of linguistic adequacy, processing considerations and theory-neutrality an annotation scheme has been adopted that is inspired by the notion of *topological fields* in the sense of Herling (1821), Erdmann (1886), Drach (1937), and Höhle (1985). As the name already suggests, the framework of topological fields tries to capture fundamental word order regularities of German sentence structure, which any syntactic theory has to capture. In this respect, the descriptive inventory provided by the topological fields approach, is theory-neutral and surface-oriented. At the same time, the concept of topological fields favors tree-based annotations, i.e. bracketings that do not rely on crossing, discontinuous dependencies. Rather, such non-linear dependencies have to be expressed at the level of predicate-argument structure, which is constrained to the possible realizations of particular topological fields, but which constitutes

a second level of annotation with its own descriptive inventory of grammatical functions.

Since syntactic annotations that are informed by topological fields and that are enriched by a level of predicate-argument structure are surface-oriented and tree-based, such an annotation scheme supports computational applications that rely on context-freeness of the underlying grammar. Finally, the framework of topological fields is widely used in empirical and theoretical accounts of German syntax and has, thus, been well documented in the linguistics literature. This greatly facilitates thorough training of human annotators, since they can rely on the pre-existing body of literature. The purpose of this stylebook is to add to these reference materials.

Chapter 3

The Theoretical Basis of the Annotation Scheme

3.1 Topological Fields

The annotation scheme for the German treebank has been developed with special regard to the characteristics of the German language: the interaction of configurational and non-configurational syntactic properties, which arise from the partially free word order. On the one hand, there exist three different clause types 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)). On the other hand, there is a high degree of variability of complements and adjuncts. In order to treat the relatively high degree of word order freedom in German, the treebank adopts the notion of topological fields as the primary clustering principle of a sentence. The basic characteristics of the model of topological sequences within a German sentence were originally formulated by Herling (1821) and later by Erdmann (1886). Drach (1937) founded the traditional notion *field* for these sequences.

3.1.1 The Concept of Topological Fields

In a German sentence, the finite verb can appear in three different positions: verb-second, verb-initial, and verb-final. Only in verb-final clauses the verb complex consisting of the finite verb and non-finite verbal elements forms a unit. The discontinuous positioning of the verbal elements in verb-first and verb-second clauses is the traditional reason for structuring German sentences in fields. The positions of the verbal elements form the *Satzklammer* (sentence bracket) which divides the sentence into a *Vorfeld* (initial field), a *Mittelfeld* (middle field), and a *Nachfeld*

Table 3.1: Three clause types

E-Sätze	(KOORD) - (C) - X - VK - Y
F1-Sätze	(KOORD - (KL) - FINIT - X - VK - Y
F2-Sätze	(KOORD or PARORD) - (KL) - K - FINIT - X - VK - Y

(final field). The Vorfeld and the Mittelfeld are divided by the *linke Satzklammer* (left sentence bracket), which is the finite verb, the *rechte Satzklammer* (right sentence bracket) is the verb complex between the Mittelfeld and the Nachfeld.

Höhle (1985) denotes the three different clause types as E-Sätze (verb-final clauses), F1-Sätze (verb-initial clauses), and F2-Sätze (verb-second clauses). The topological schemata of these types are the following (cf. Table 3.1):

Abbreviations and explanations (Table 3.1):

VK: verb complex

FINIT: element denoting categories of finiteness

KOORD: coordinating particles (e.g. *und*, *oder*)

PARORD: non-coordinating particles (e.g. *denn*, *weil*)

X, Y: sequence of any number of constituents

C: complementizer

K: one constituent

KL: nominativus pendens, resumptive construction (*Linksversetzung*)

These schemes topologically analyse not only atomic sentences but also complex sentence constructions which contain embedded clauses. Such embedded clauses can occur in a *Linksversetzung* (resumptive construction), Vorfeld, Mittelfeld, or Nachfeld. Herling's theory of the coordination and embedding of sentences covers these phenomena in detail (Herling 1821).

According to Höhle (1985), we assume the existence of the following topological fields (cf. Table 3.2):

The following description of the topological fields does not claim completeness

Table 3.2: Topological fields

Field	Description
VF	Vorfeld (initial field/pre-field)
LK	Linke (Satz-)Klammer (left sentence bracket)
MF	Mittelfeld (middle field)
VC	Verbkomplex (verb complex)
NF	Nachfeld (final field/post-field)
LV	Linksversetzungsfeld (field for resumptive constructions)
C	C-Feld (field for complementizers), only in verb-final clauses; C is immediately followed by MF (i.e. there is no LK)
KOORD	Koordinationsfeld (field for coordinating particles), left-most element, in all clause types possible, (e.g. <i>und, oder</i>)
PARORD	Koordinationsfeld (field for coordinating particles) left-most element, only verb-second (e.g. <i>denn, weil</i>)

regarding all descriptive details but rather mentions their main characteristics.¹

VF: The Vorfeld consists of only one constituent. Usually it is the subject. But because of the high degree of non-configurationality in German, in most cases the subject can occur in the Mittelfeld, thus allowing almost every other constituent to occupy the Vorfeld.

LK: This is the position of the finite verb in verb-second and verb-first clauses or a conjunction in verb-final clauses. The Linke Klammer consists of exactly one element.

MF: Apart from those units which are optionally located in other fields, any non-verbal constituent may occur in the Mittelfeld. The linear order of the constituents depends on the specific word order principles for German and their interaction.

VC: In verb-second and verb-first clauses the verb complex consists of one or more non-finite elements or - depending on the verb - of a separable prefix. In verb-final clauses the verb complex also contains the finite verb. The rule for the linear order in general is: right determines left. If there is a finite verb in the verb complex, it is always the right-most element (exception: *daß er getroffen hätte*

¹In the following the abbreviations for the fields listed in Table 3.2 above are used.

werden können).

NF: For some clause types (e.g. *so daß*-Sätze), the Nachfeld is the obligatory position. Embedded complement clauses, relative clauses, and single constituents can optionally occur in the Nachfeld. In contrast to the Vorfeld it may be occupied by more than one constituent.

LV: A Linkversetzung is a pendent constituent. It can be regarded as a syntactic anticipation of a part of a sentence. There are many restrictions which apply for this position (cf. section 6.4).

C: The C-position only occurs in verb-final clauses. It is obligatorily occupied in finite German verb-final clauses. In non-finite verb-final clauses the C-position may be empty. The C-field can only be occupied not only by conjunctions of sentential objects (e.g. *dass, ob*) or sentence initial conjunctions like *um, obwohl, wenn* but also by complex interrogative or relative phrases, e.g. ..., '*um wieviel Uhr der Zug ankommt*'. (cf. section 6.1).

KOORD: This is an alternative field for coordinative particles. In contrast to the PARORD-field, it can optionally occur as the left-most element of all clause types (cf. section 6.2).

PARORD: This is the field for coordinative particles which optionally belong to the sentence they introduce. It is always the left-most element of a verb-second clause (cf. section 6.3).

Concerning the distribution of constituents to topological fields see also the chapter *Deskriptive Generalisierungen* in Grewendorf (1991).

The combination and ordering of these fields in order to constitute verb-first, verb-second, or verb-final clauses is described in Höhle (1985). The role of individual topological fields within the German treebank will be explained in the course of chapter 5.

The topological model, which is the basis of most traditional German grammars, only provides descriptive parameters concerning the sentence structure without making any statement about the regularities within the fields and the hierarchical constituent structure of the sentence. For more complicated phenomena, it offers only a catalog of detailed descriptions.

3.2 Constituent Analysis and Topological Fields

The main weakness of the concept of topological fields is the above mentioned fact that the hierarchical constituent structure of a sentence cannot be described. The aim is to find a form of representation which combines the topological model with a constituent analysis in order to describe the hierarchy of the linguistic units within the fields. In our annotation scheme, the integration of a constituent analysis was achieved by a second level of annotation strictly within the bounds of topological fields: a predicate-argument structure with its own descriptive inventory of syntactic categories and grammatical functions. The constituent structure is represented by phrase structure trees (phrase markers) whose node labels carry this information.

In order to analyse syntactic constructions, it is necessary to define the number and types of constituents within the fields.

1. Number of constituents within the fields:

In general, VF, C, PARORD, and KOORD only contain **one** constituent. More than one constituent is allowed within MF and NF.

2. Types of constituents within the fields:

Embedded clauses either belong to NF, VF, or LV. Usually, outside the spoken language context, verb-final clauses do not occur isolated. They need to be attached if possible.

Sometimes the decision to which field a constituent belongs is difficult. The following tests and criteria help to assign the constituents to the proper field:

Field borders:

LK constitutes the border between VF and MF, VC constitutes the border between MF and NF. Whenever a constituent can occur both in MF and NF, it is attached to MF:

ich glaube [MF so drei Stunden].

sagen [MF wir von vierzehn bis siebzehn Uhr].

MF Test:

A test for the MF is to place the constituent in question between LK and

VC: **laß** *uns von vierzehn bis siebzehn Uhr sagen*

The predicate complement of verbs like *recht haben* and the accusative object of certain verbs like *Zeit haben* also indicate the beginning of the NF: In verb-second clauses, the NF begins after the predicate or the accusative object of the verb, similar to the function of a verb particle:

da haben [MF *Sie auch wieder recht*] [NF *natürlich*].
wann hast [MF *Du denn Zeit*] [NF *in den nächsten Tagen*]?

A test is to form a sentence in which the verb occurs at the end (verb-final) and to check, whether the phrase in question (*natürlich*) may occur in MF position without changing the meaning of the sentence: *recht natürlich (haben)* means something different than *natürlich recht haben*.

This indicates that *natürlich* needs to be attached at a “higher” node rather than close to the VC:

* *weil* [MF *Sie da auch wieder recht natürlich*] [VC *haben*]

3.3 General Annotation Principles

Our annotation scheme tries to strike a balance between pragmatic requirements on the one hand and linguistic reality on the other hand. The following three common annotation principles are adopted to group the constituents within a syntactic tree: the *flat clustering principle*, the *longest match principle*, and the *high attachment principle*.

3.3.1 Flat Clustering Principle

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, but are not typically connected to surrounding constituents as a whole.

3.3.2 Longest Match Principle

The *longest match principle* demands that as many daughter nodes as possible are combined into a single mother node, provided that the resulting construction is syntactically as well as semantically well-formed.

3.3.3 High Attachment Principle

The *high attachment principle* prescribes that in case of syntactic and semantic ambiguity in the attachment of modifiers such ambiguous modifiers are attached to the highest possible level in a tree structure. Premodifiers and postmodifiers are treated in a different way. First both kinds of modifiers are projected to their phrase level. 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 was taken to avoid the problematic distinction whether a postmodifier is a free adjunct or a complement of the modified phrase.

3.4 The Structure of an Annotated Tree

3.4.1 The Levels of Annotation

A syntactic tree consists of nodes and edges. Nodes represent constituents on different levels of annotation. Edges always link daughter nodes to a mother node. The root node of a tree is assumed as the sentence node of a construction. One level below the sentence node, the nodes of the topological fields are located. This is the reason why topological fields can be regarded as the top-level ordering principle for sentences in the German treebank. The sequence of the fields in the three clause types never violates the topological schemes given by Höhle. Within each sentence structure, in general at least two topological fields are occupied (exception infinitive constructions, cf. section 4.6.2). Others may be left empty (cf. section 6.9, elliptic constructions).

Table 3.3 lists the four levels of annotation, that we distinguish within the structure of an annotated syntactic tree:

Node labels denote the syntactic category of a phrase or sentence, a topological field, or a grammatical property.

Table 3.3: Levels of annotation

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 (Schiller et al. 1995)

Edges always link daughter nodes to a mother node. Edge labels denote the grammatical function of lexical entries, phrases, topological fields, and clauses.

3.4.2 The Inventory of Labels

The **part-of-speech labels** used for the annotation are taken from the Stuttgart-Tübingen Tagset (STTS) (Schiller et al. 1995).² The STTS is a guideline for the annotation of German text corpora on the lexical level. Every single part-of-speech of a text is assigned one specific tag. The tagset consists of the tags listed in Table 3.4.2 (cf. (Schiller et al. 1995)):

The tagging of the data was done with the help of the Brill-tagger (Brill 1992) and then corrected manually.

Node labels indicate the syntactic category of a phrase or sentence, but they are also used to label topological fields and sequences of topological fields within coordinations or to indicate specific grammatical properties of constituents (e.g. P-SIMPX). Table 3.4 lists all node labels, that are used in the German treebank.

Edge labels indicate the grammatical function of lexical entries, phrases, topological fields, and clauses. Since case information is given and a distinction of different modifiers is made by these labels, the syntactic tree structures also contain semantic roles. The specific set of edge labels for the German treebank is listed in Table 3.5.

²PAV has been changed into a new tag called PROP (pronominal form of a prepositional phrase) in order to justify PX as the syntactic category of its mother. The tag BS has been introduced to handle single characters (spelling etc.).

Table 3.4.2: The STTS tagset

POS =	description	examples
ADJA	attributive adjective	<i>[das] große [Haus]</i>
ADJD	adverbial or predicative adjective	<i>[er f"ahrt] schnell, [er ist] schnell</i>
ADV	adverb	<i>schon, bald, doch</i>
APPR	preposition; left circumposition	<i>in [der Stadt], ohne [mich]</i>
APPRART	preposition + article	<i>im [Haus], zur [Sache]</i>
APPO	postposition	<i>[ihm] zufolge, [der Sache] wegen</i>
APZR	right circumposition	<i>[von jetzt] an</i>
ART	definite or indefinite article	<i>der, die, das, ein, eine</i>
BS	letter	<i>\$A, \$I</i>
CARD	cardinal number	<i>zwei [M"anner], [im Jahre] 1994</i>
FM	foreign language material	<i>[Er hat das mit "] A big fish [" "ubersetzt]</i>
ITJ	interjection	<i>mhm, ach, tja</i>
KOUI	subordinating conjunction with <i>zu</i> + infinitive	<i>um [zu leben], anstatt [zu fragen]</i>
KOUS	subordinating conjunction with clause	<i>weil, da, damit, wenn, ob</i>
KON	coordinative conjunction	<i>und, oder, aber</i>
KOKOM	particle of comparison, no clause	<i>als, wie</i>
NN	noun	<i>Tisch, Herr, [das] Reisen</i>
NE	proper noun	<i>Hans, Hamburg, HSV</i>
PDS	substituting demonstrative pronoun	<i>dieser, jener</i>
PDAT	attributive demonstrative pronoun	<i>jener [Mensch]</i>
PIS	substituting indefinite pronoun	<i>keiner, viele, man, niemand</i>
PIAT	attributive indefinite pronoun without determiner	<i>kein [Mensch], irgendein [Glas]</i>
PIDAT	attributive indefinite pronoun with determiner	<i>[ein] wenig [Wasser], [die] beiden [Br"uder]</i>
PPER	irreflexive personal pronoun	<i>ich, er, ihm, mich, dir</i>
PPOSS	substituting possessive pronoun	<i>meins, deiner</i>
PPOSAT	attributive possessive pronoun	<i>mein [Buch], deine [Mutter]</i>

Stylebook for the German Treebank

POS =	description	examples
PRELS PRELAT PRF	relative pronoun substituting attributive reflexive personal pronoun	<i>[der Hund,] der</i> <i>[der Mann ,] dessen [Hund]</i> <i>sich, einander, dich, mir</i>
PWS PWAT PWAV	substituting interrogative pronoun attributive interrogative pronoun adverbial interrogative or relative pronoun	<i>wer, was</i> <i>welche [Farbe], wessen [Hut]</i> <i>warum, wo, wann, wor"uber, wobei</i>
PROP	pronominal adverb	<i>daf"ur, dabei, deswegen, trotzdem</i>
PTKZU PTKNEG PTKVZ PTKANT PTKA	<i>zu</i> + infinitive negation particle separated verb particle answer particle particle with adjective or adverb	<i>zu [gehen]</i> <i>nicht</i> <i>[er kommt] an, [er f"ahrt] rad</i> <i>ja, nein, danke, bitte</i> <i>am [sch"onsten], zu [schnell]</i>
TRUNC	truncated word - first part	<i>An- [und Abreise]</i>
VVFIN VVIMP VVINFINF VVIZU VVPP VAFIN VAIMP VAINFINF VAPP VMFIN VMINFINF VMPP	finite main verb imperative, main verb infinitive, main infinitive + <i>zu</i> , main past participle, main finite verb, aux imperative, aux infinitive, aux past participle, aux finite verb, modal infinitive, modal past participle, modal	<i>[du] gehst, [wir] kommen [an]</i> <i>komm [!]</i> <i>gehen, ankommen</i> <i>anzukommen, loszulassen</i> <i>gegangen, angekommen</i> <i>[du] bist, [wir] werden</i> <i>sei [ruhig !]</i> <i>werden, sein</i> <i>gewesen</i> <i>d"urfen</i> <i>wollen</i> <i>[er hat] gekonnt</i>
XY	non-word containing special characters	<i>D2XW3</i>
\$, \$.	comma sentence-final punctuation	<i>,</i> <i>. ? ! ; :</i>

Table 3.4: Node labels

Node Labels	Description
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. <i>gar keine</i>)
Topological Field Node Labels	
LV	resumptive construction (Linksversetzung)
VF	initial field (Vorfeld)
LK	left sentence bracket (Linke (Satz-)Klammer)
MF	middle field (Mittelfeld)
VC	verb complex (Verbkomplex)
NF	final field (Nachfeld)
C	complementizer field (C-Feld)
KOORD	field for coordinative particles
PARORD	field for coordinative particles
FKOORD	coordination consisting of conjuncts of fields
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
Root Node Labels	
SIMPX	simplex clause
R-SIMPX	relative clause
P-SIMPX	paratactic construction of simplex clauses
DM	discourse marker

Table 3.5: Edge labels

Edge Labels	Description
Edge Labels denoting Head	
HD	head
-	non-head
Complement Edge Labels	
ON	nominative object
OD	dative object
OA	accusative object
OS	sentential object
OPP	prepositional object
OADV	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	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	
refl	first verbal object in VC selected by a verbal object

Within phrases, the main grammatical functions are HD and “-”, indicating the head (HD) of a phrase and definite modifiers or specifiers (“-”).

Edge labels are empty (“-”), for instance, for determiners, attached modifiers within phrases, discourse markers, date expressions consisting of intervals, below C, sentence nodes, and mother nodes of coordinations.

3.4.3 Where Does a Tree End?

In order 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 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.

A SIMPX tree ends where a complete syntactically well-formed sentence ends according to the “longest match” strategy. The model of topological fields does not prescribe that all fields have to be occupied. The fact that fields can be left empty also allows us to cope with elliptic sentences (cf. section 6.9).

Note that in the VERBMOBIL dialogs punctuation is not always a reliable help to detect sentence borders. It has to be considered carefully, whether utterances are also units of meaning and whether every constituent can be assigned a proper grammatical function within the sentence.

The following turn, for instance, consists of two trees:

[dann fange ich einfach mal an und wollte Sie mal fragen, wie das aussieht] ähm [wir müssten also insgesamt drei Arbeitssitzungen festlegen].

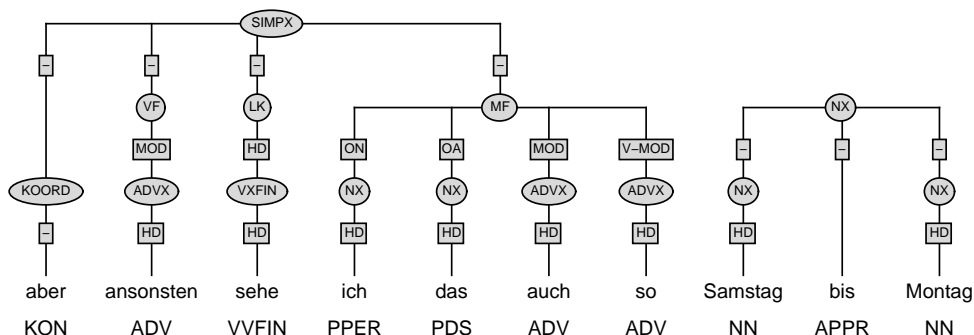
In some cases, it can be difficult to decide whether a constituent should be included in the tree or not:

[LV die Woche davor], [VF wie] [LK sähe] [MF das] [VC aus], [NF erste Novemberwoche]?

In this case, *erste Novemberwoche* is assumed to modify the subject of the sentence (*das*). Thus, its grammatical function in the sentence is well-motivated and, as a consequence, it can be included in the NF of the tree.

In the following examples, there are constituents that cannot be embedded be-

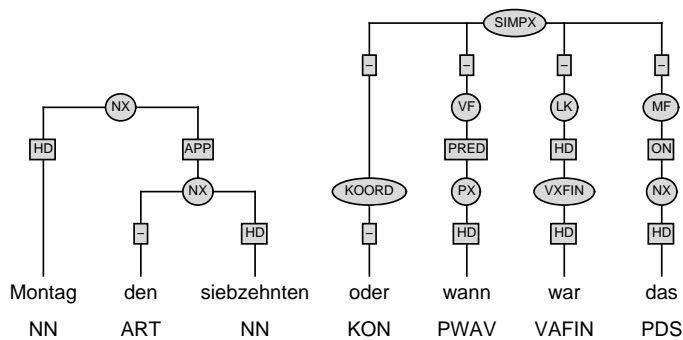
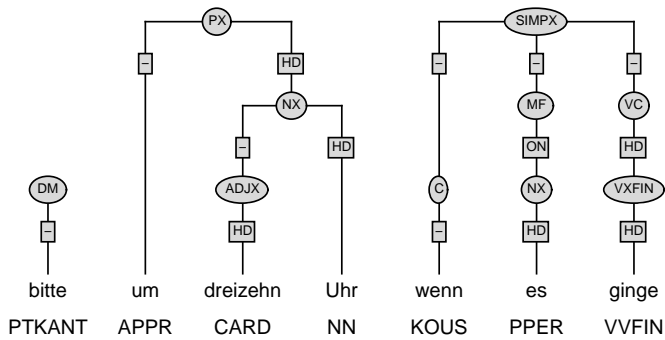
cause they cannot be assigned reasonable grammatical functions and any potential attachment structure would cause problems to existing attachment strategies.³



The tree above 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. This unit consists of a grammatically well-formed sentence and an isolated phrase. In accordance with the four annotation levels shown in Table 3.3, 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 lacking a verbal constituent.

Discourse markers are another type of isolated units. They can occur on the left, right, or in the middle (e.g. as parenthetical utterances) of a well-formed sentence without being connected to the tree at all (cf. section 7.5):

³This tree diagram and all following tree diagrams in this report were generated with the aid of the Negra *Annotate* tool.

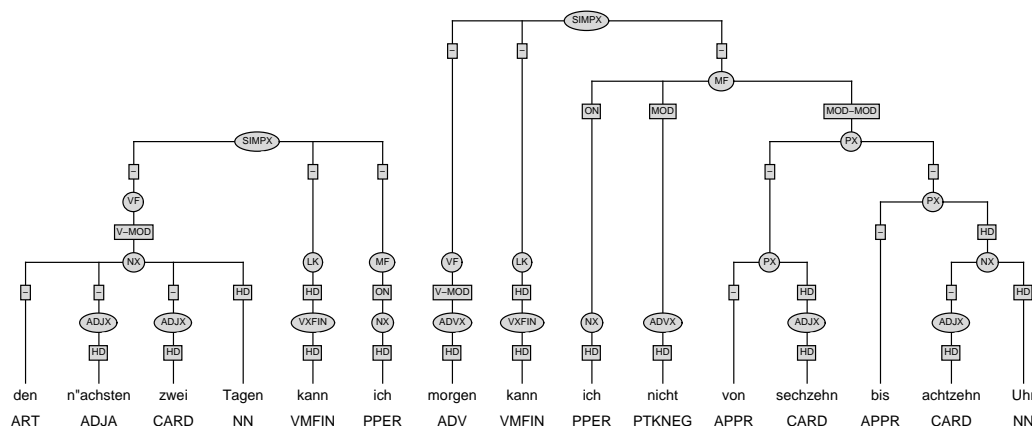


3.4.4 Speech Errors and Repetitions

Similar to discourse markers, speech errors and repetitions are treated as isolated elements in the tree, i.e. they are structured as much as possible (mostly up to the level of phrasal categories), but they are not typically connected to surrounding constituents as a whole. As a consequence, the well-formed part of the sentence is not affected by repeated or ungrammatical elements. This decision was taken, first, because in most cases these elements cannot be assigned a specific grammatical function within the well-formed sentence that was not already assigned to some other constituent, and second, because the attachment of speech errors or repetitions would conflict with the topological field analysis in many cases:

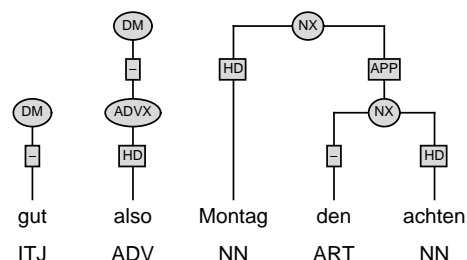
1. *[das] dann haben wir doch die Sachen.*
(*das* conflicts with *dann* and cannot be included, since the VF allows only one constituent)
2. *der [ist] paßt mir ausgezeichnet.*
(the LK is filled by *paßt* and may only contain one constituent, so there is no space for *ist* in any field)

The following example shows a complex speech error that is not attached to the well-formed sentence:



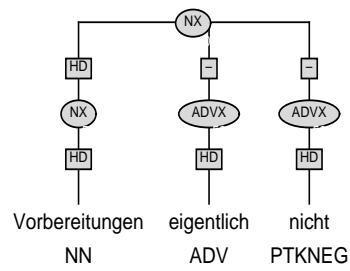
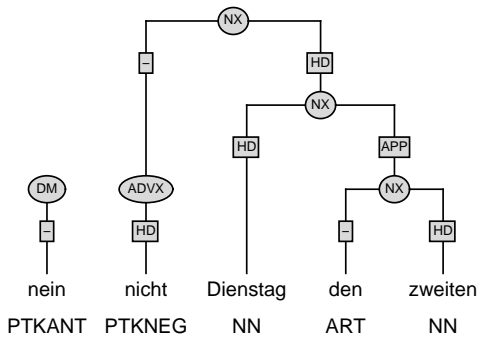
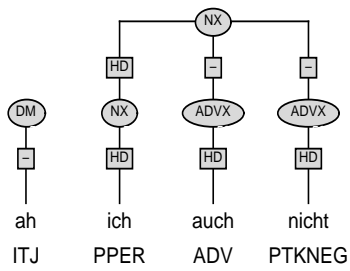
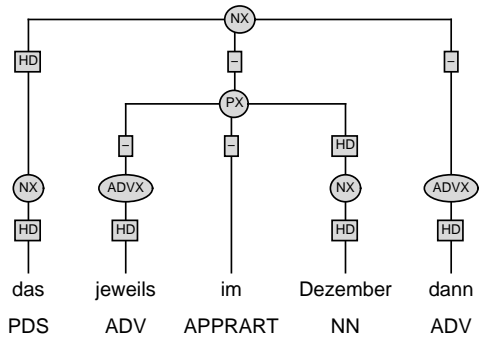
3.4.5 Isolated Phrases and Sentence Fragments

There are utterances that cannot be analysed as SIMPX because they are missing a verbal constituent. However, they are not discourse markers either. These utterances are annotated as isolated phrases or sentence fragments. Their root node is a phrasal category of their lexical head (NX, PX, ADVX, etc.), e.g. *Montag, den achten*:

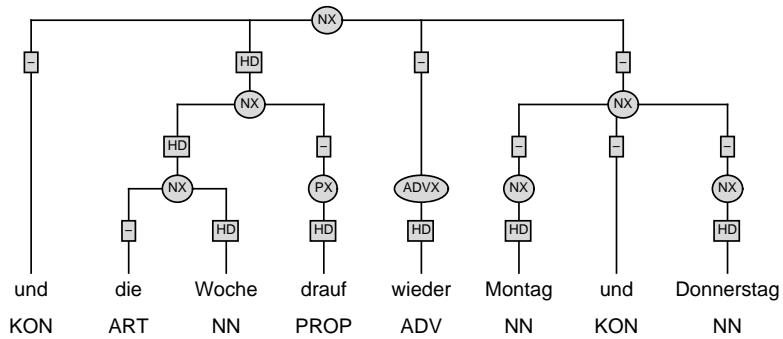
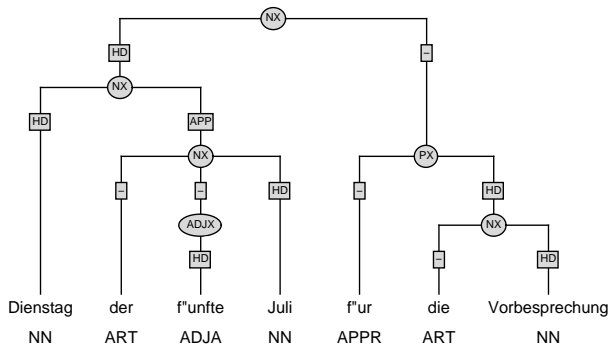
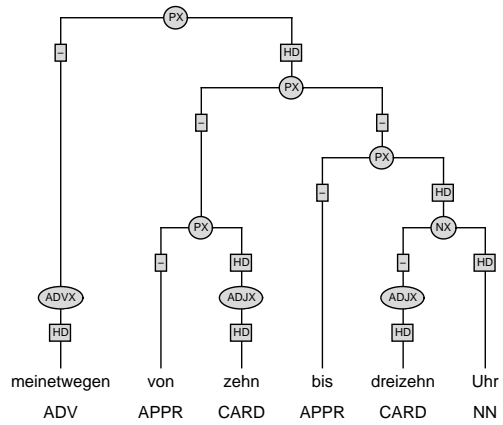


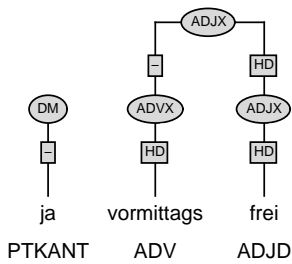
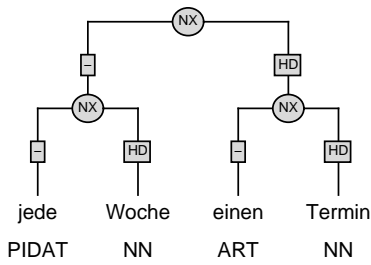
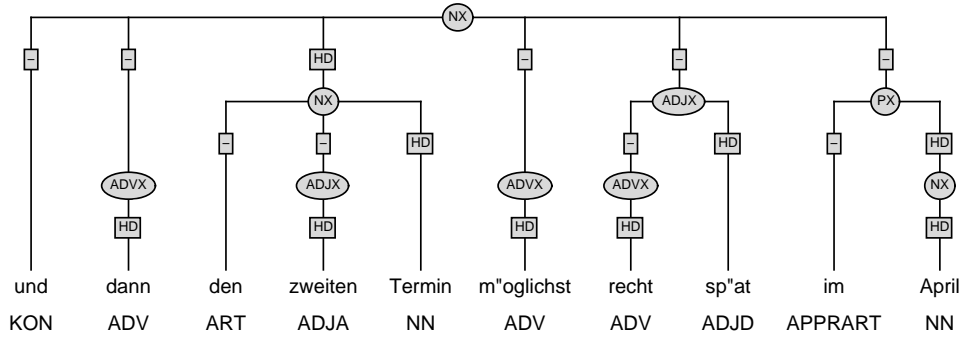
In accordance with the *longest match principle*, as many parts of the fragment as possible have to be projected up to the phrase level and included into a tree structure. It has to be decided which part of the whole construction is the head and which parts are modifying this head. In cases of more complex fragments, it can be difficult to determine the *head* of the entire construction. Note that none of the possible solutions for a given construction provided is unproblematic and that a compromise has to be made in order to keep the different parts of the fragment together as a unit and at the same time to be able to assign a head to it:

Stylebook for the German Treebank



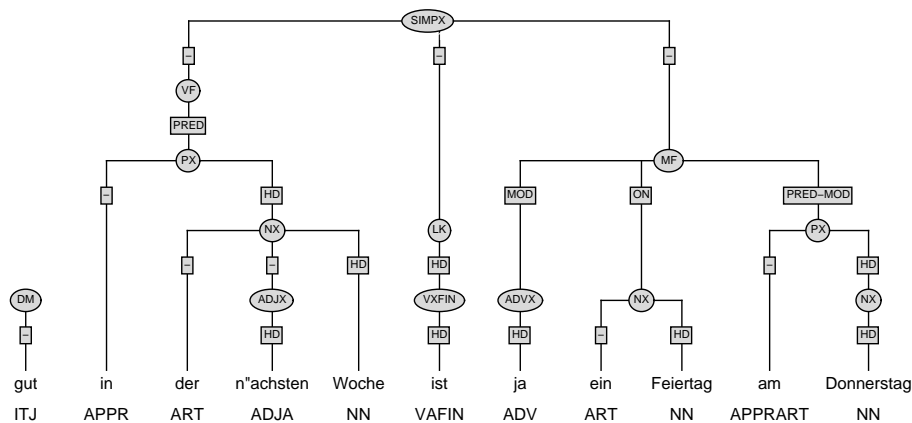
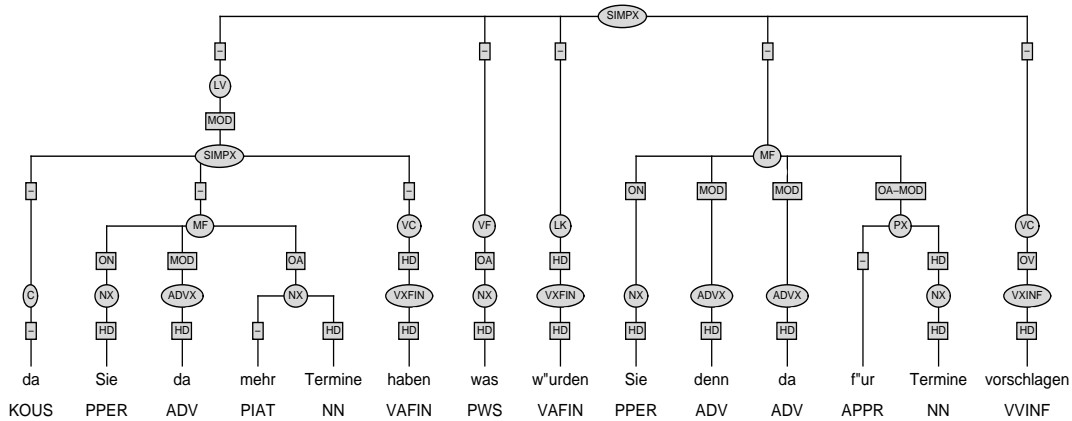
Stylebook for the German Treebank





3.4.6 Long-Distance Dependencies

Our annotation scheme facilitates a theory-neutral and surface-oriented representation of syntactic trees without crossing branches and traces. Where dependency relations cross the border between constituents or topological fields the references between these discontinuous constituents are encoded by special naming conventions for edge labels. We use unambiguous edge labels such as OA-MOD (referring to OA) or PRED-MOD (referring to PRED) etc.:

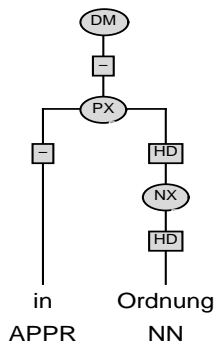


3.4.7 Empty Categories

In general an empty category analysis, e.g. for phrases without heads, is being avoided in the German treebank. Specifiers, prepositions⁴, complementizers, discourse markers, KOORD and PARORD constituents, conjunctions, and unambiguous modifiers (that are attached to phrases immediately rather than to topological fields) are not labelled with grammatical functions. Furthermore, the edges below the sentence node are empty. They are not labelled in order to speed up annotation where the information is unnecessary or self-evident.

The following example shows empty edge labels below the PX node and the DM node:

⁴In order to facilitate the identification of dependencies between verbs and their nominal complements and adjuncts and in keeping with basic assumptions in Dependency Grammar, the annotated head of a prepositional phrase is the NX (or complement) rather than the preposition itself. Therefore, prepositions carry no edge label.



The treatment of phrases with apparently “empty heads” will be explained in section 4.3.1 where specific date constructions are presented.

Chapter 4

The Annotation of the Internal Structure of Phrases

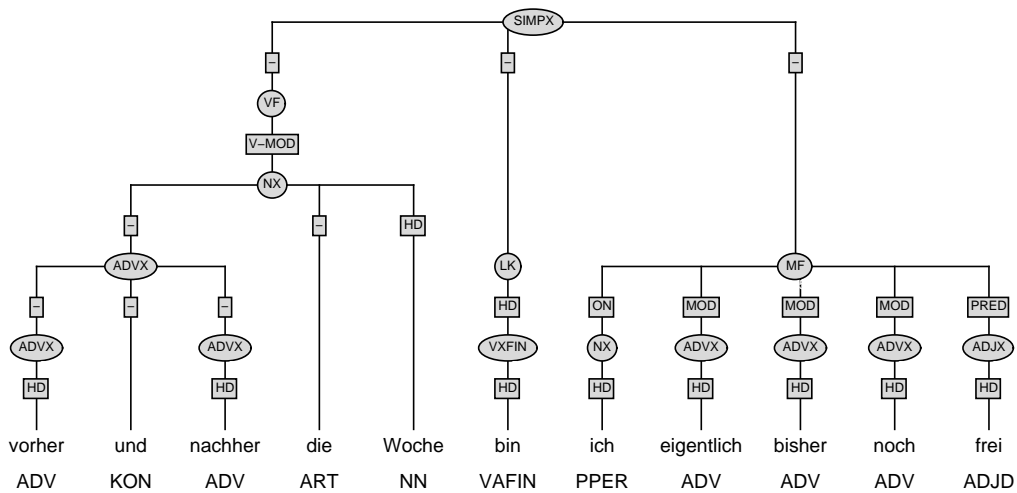
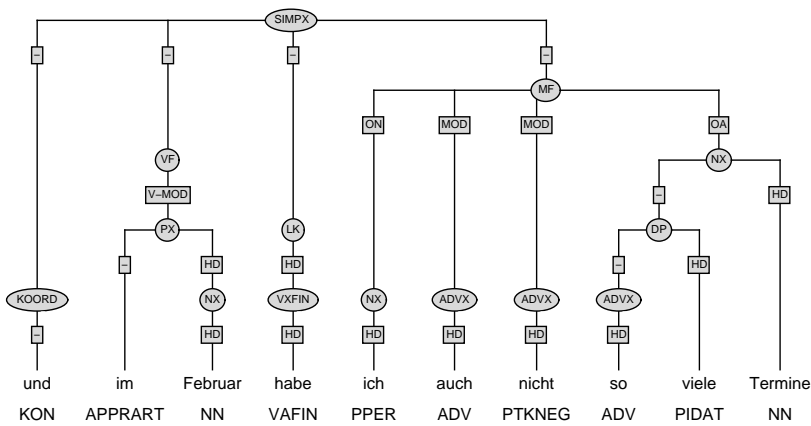
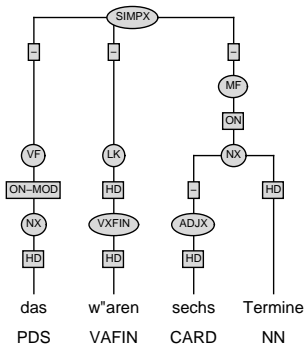
4.1 Noun Phrases

A simple noun phrase NX without any modifier consists of a head noun (noun, proper noun, or a pronoun) and (optionally) a determiner. Both are directly attached to the same level so that the edge label of the head noun carries the label HD and the edge label of the determiner is empty. In case of a more complex NX, the annotation differs depending whether the modifiers are prenominal or postnominal modifiers of this phrase have to be attached to the NX either prenominally or postnominally. According to their distribution, ordinal numbers either occur as a premodifying attributive adjective (e.g. *der dritte/ADJA Juni* or as a head noun (e.g. *wir treffen uns am vierten/NN*). In the first case the premodifier is projected to adjectival phrase, in the latter case to a noun phrase.

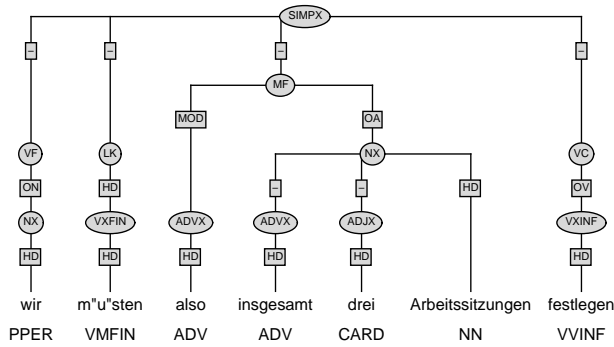
4.1.1 Prenominal Modification

All prenominal modifiers are first projected to their phrase level. In a second step they are directly attached to the head noun on the same level because their scope of modification is unambiguous. Thus, their edge labels are empty (whereas the edge labels of modifiers that are attached to topological fields are non-empty). A few examples of prenominal modification:

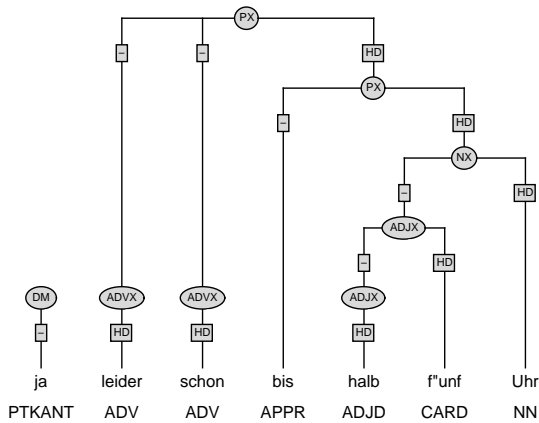
Stylebook for the German Treebank



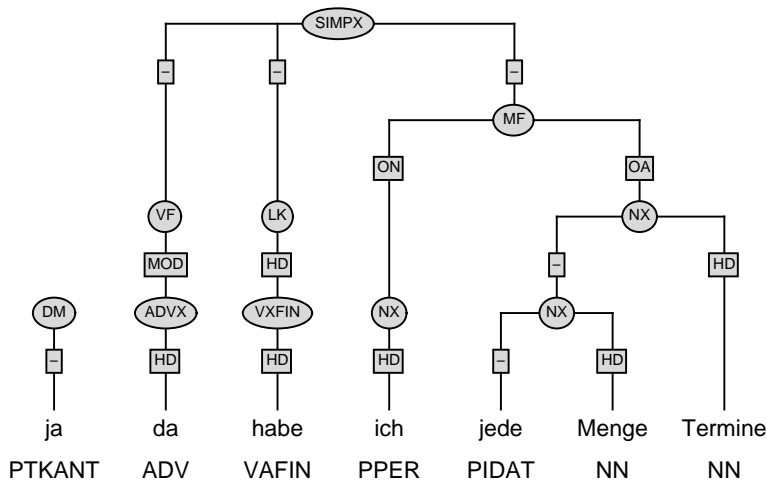
If there is more than one prenominal modifier, all of these modifiers are attached to the head noun on the same level which yields a rather flat NX structure:



Whenever a modifier is modified by another modifier, the modification strategy is the same. The complex adjunct phrase as a whole is the modifier of the noun phrase. For instance:

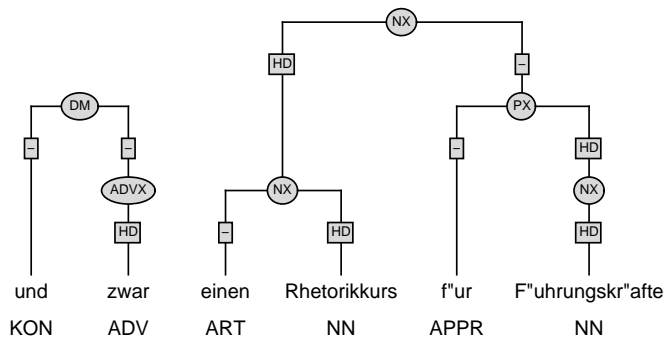


Note that determiners always have to be attached on a “low level”, i.e. as immediate sisters of the lexical head, except in the case of complex proper names (cf. section 4.1.4, e.g. *das Hotel Cristal Hannover*) and head noun coordination (cf. section 6.10.2 *den siebzehnten bis neunzehnten und den vierundzwanzigsten bis sechsundzwanzigsten*):

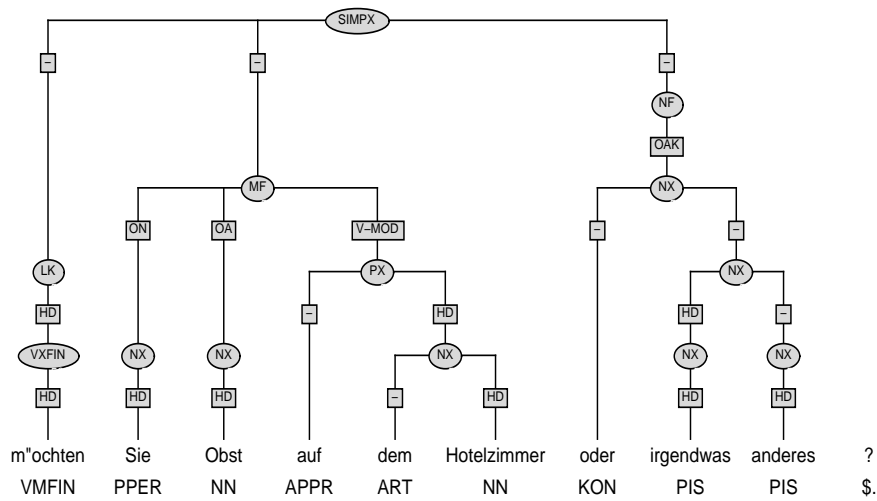
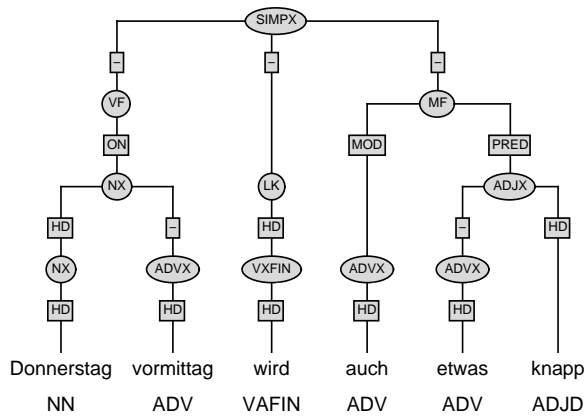
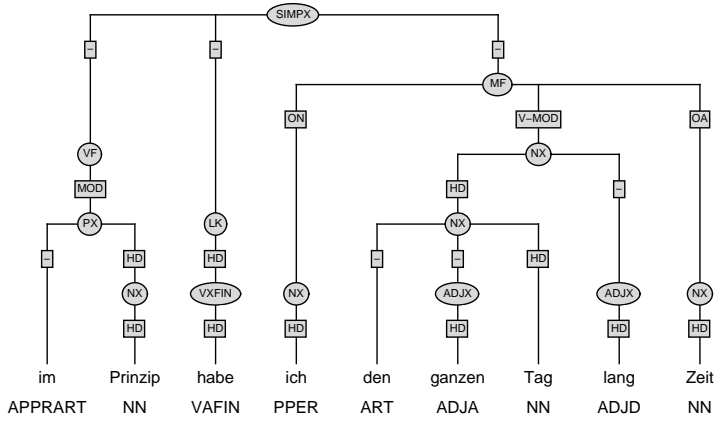


4.1.2 Postnominal Modification

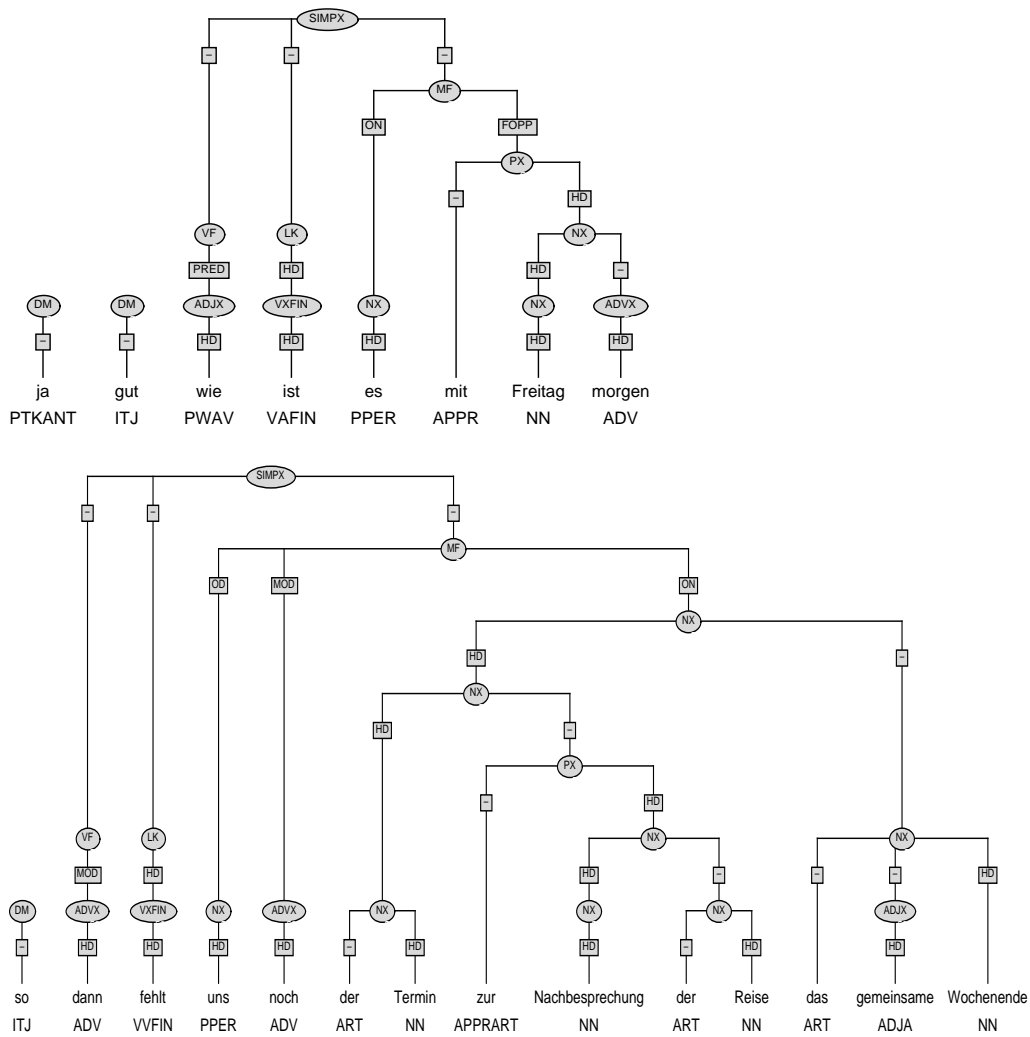
Whereas prenominal modifiers are always directly attached to the head noun on the same level, postnominal modifiers are attached to the NX on a higher level to avoid the problematic distinction whether the modifier is a free adjunct or a complement of the head noun. For this reason, head noun and postnominal modifiers are always first projected to the phrase level and then the postnominal modifier is attached to the head noun on a higher level:



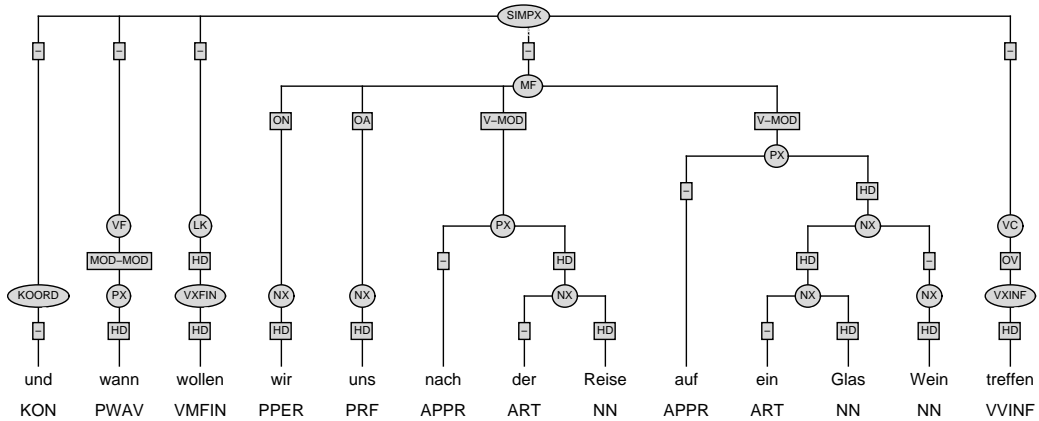
Stylebook for the German Treebank



Stylebook for the German Treebank

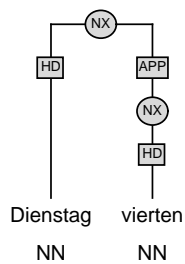
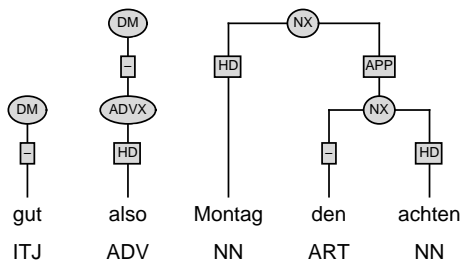


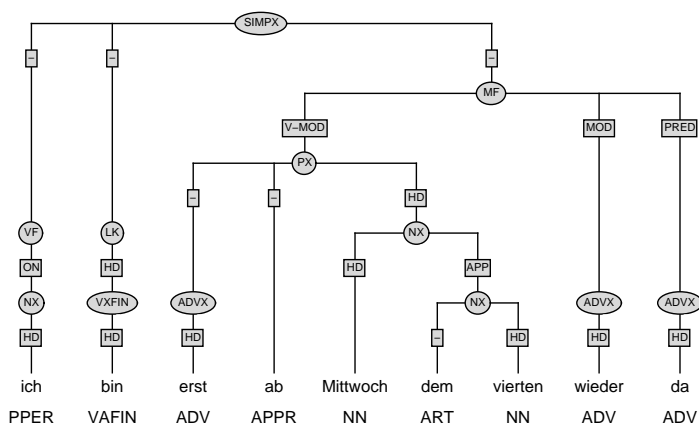
In the following case, *Wein* specifies the noun phrase *ein Glas*. For this reason it is treated as a postnominal modifier:



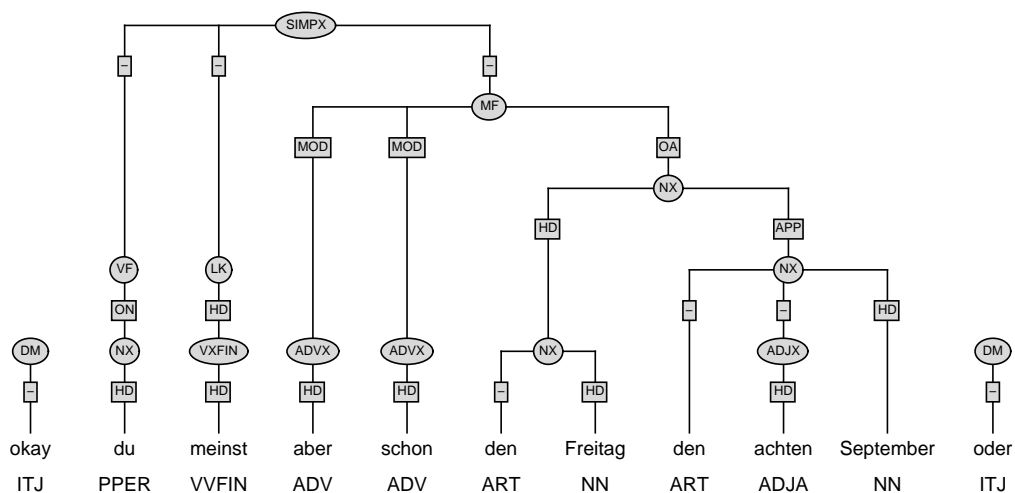
4.1.3 Appositions

Appositions such as in *Montag, der zweite* or *am Montag, dem zweiten* are labelled as APP and are immediately attached to the head noun on a “low level” because their scope of modification is unambiguous. An apposition in German is a special form of an attribute, which always follows the head noun and agrees in case with the head noun:

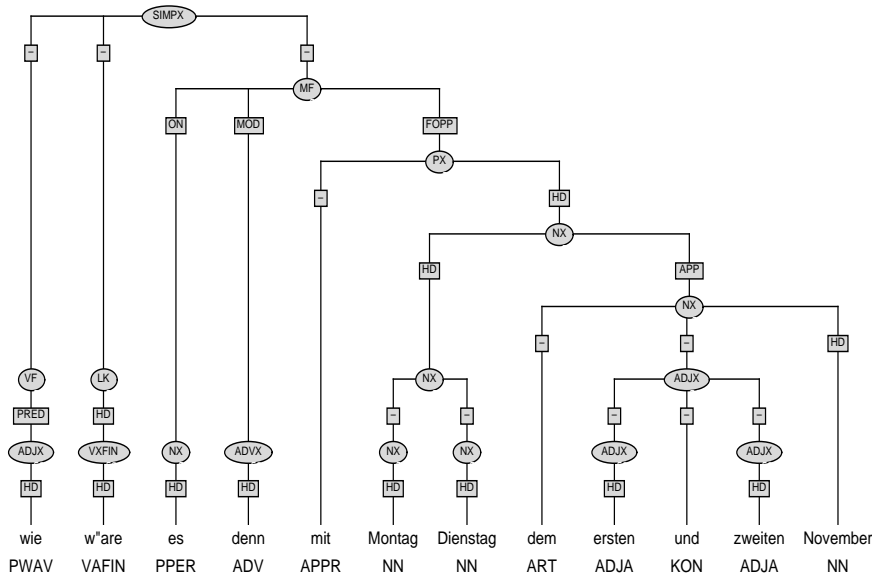




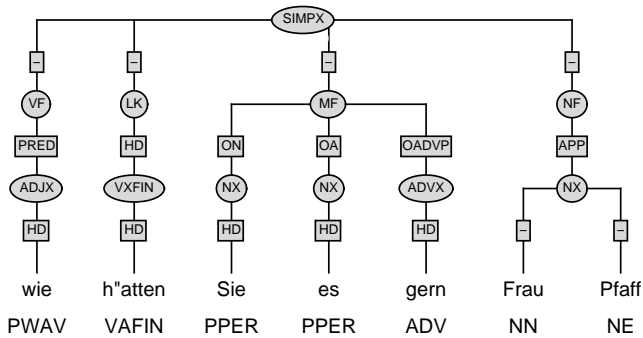
If the head noun (*Freitag*) occurs with a determiner, first the NX is projected and the apposition will be attached to the NX in a second step because head noun and apposition are two separate noun phrases:



The same strategy applies to coordinations or enumerations of nouns to which the apposition is attached:



A special case of apposition are names which constitute appositions of personal pronouns that are not adjacent. Appositions of this kind can only occur in the NF:

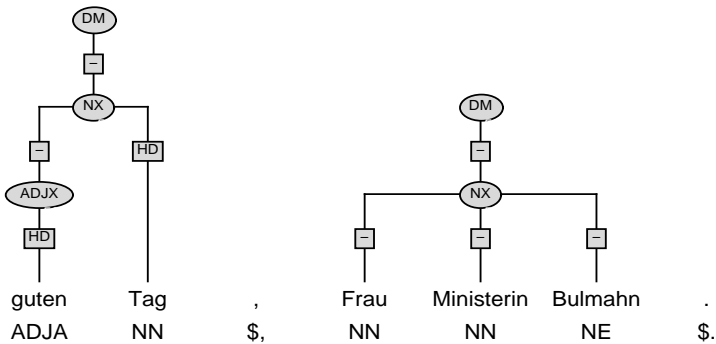
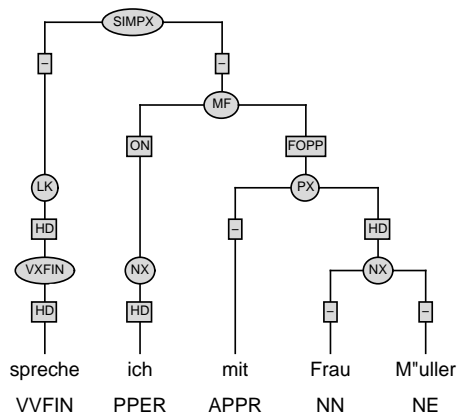
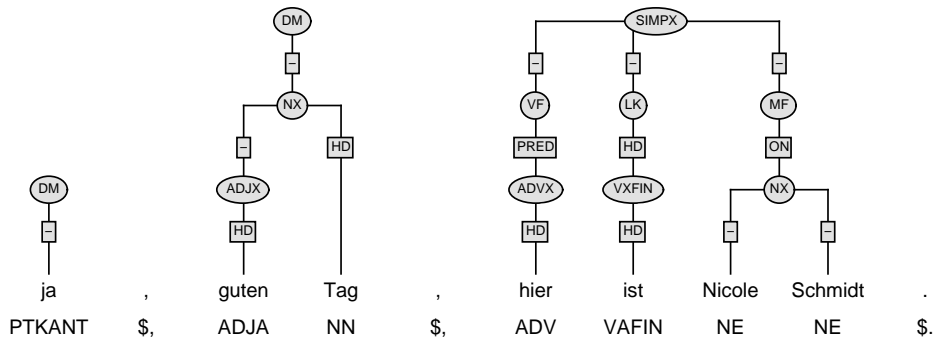


4.1.4 Proper Name Phrases

Proper names including titles etc. are attached on the same level to indicate that there is no obvious dependency relation between them:¹

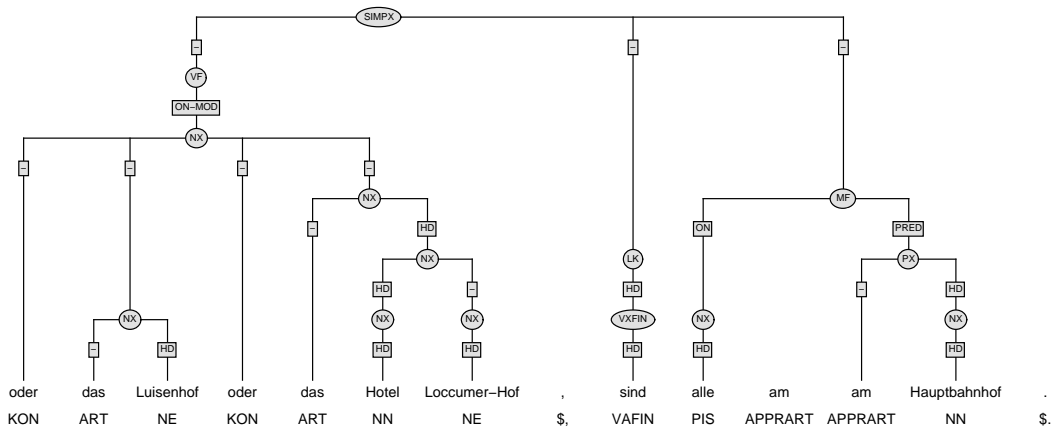
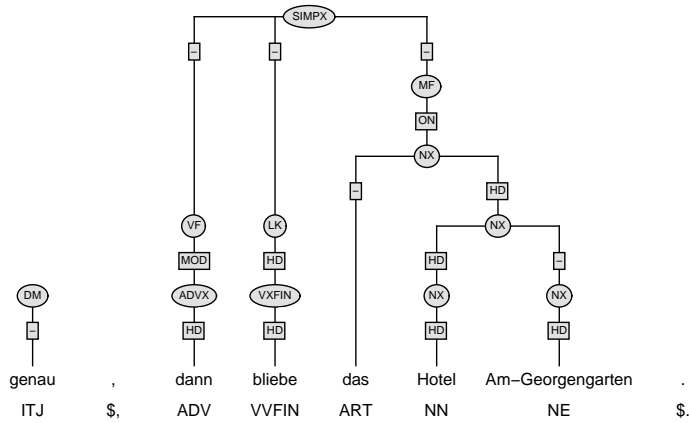
¹Commas in VERBMOBIL do not serve as markers for the internal sentence structure, they rather serve as hesitation and pause markers.

Stylebook for the German Treebank

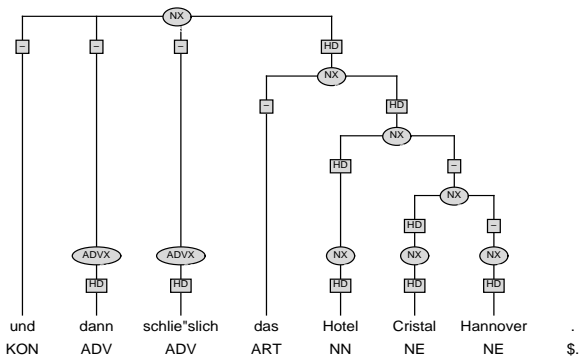


These cases have to be distinguished from cases in which the proper name modifies a word postnominally. In the latter case, both nouns are projected and then combined on the next higher phrase level. If there is an article, it always agrees with the head noun and is attached on the highest level. Single proper names with an article are treated like all other nouns (e.g. *das Luisenhof*).

Stylebook for the German Treebank

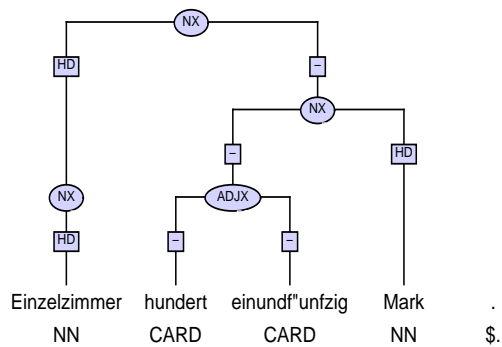
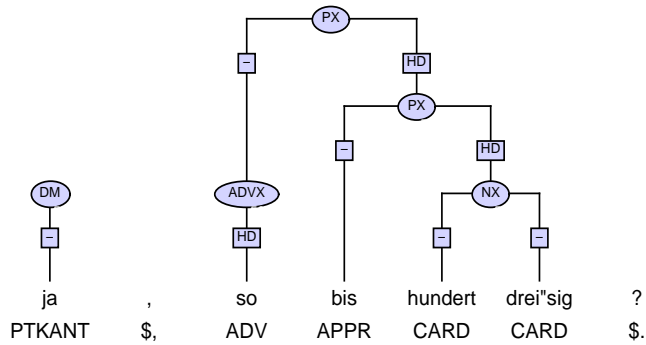


More complex nominal constructions like in the following example are structured according to their internal dependency relations:



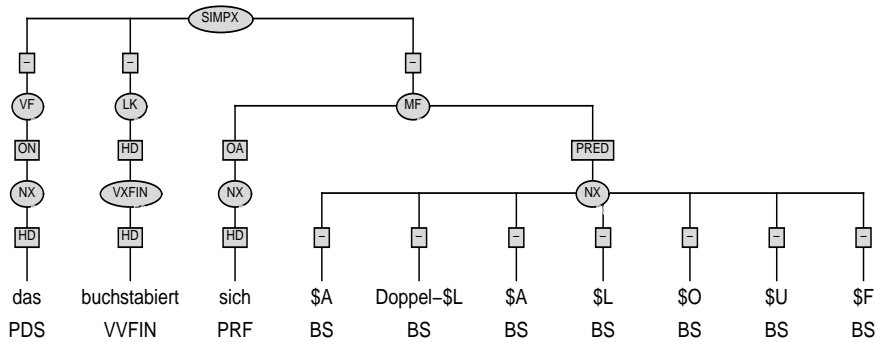
4.1.5 Complex Cardinal Numbers

Since in VERBMOBIL cardinal numbers, tagged as CARD, are written separately, e.g. *ehundert dreiundsiebzig*, they are attached on the same level like proper name phrases. They can either be projected to an NX or an ADJX:



4.1.6 Spelling

The spelling of words and expressions such as *Doppel-l* are very rare in VERBMOBIL. In accordance with the strategy for proper names, they are annotated as follows:



4.1.7 Expletive *es*

The first position of a sentence can be occupied by an obligatory element with purely morphosyntactic function: the expletive *es*. As the expletive *es* has no semantic content, it cannot be regarded as a proper argument of the verb, although its syntactic role is that of a purely formal grammatical subject. Two different kinds of the expletive *es* need to be distinguished:

1. *es* as a formal subject

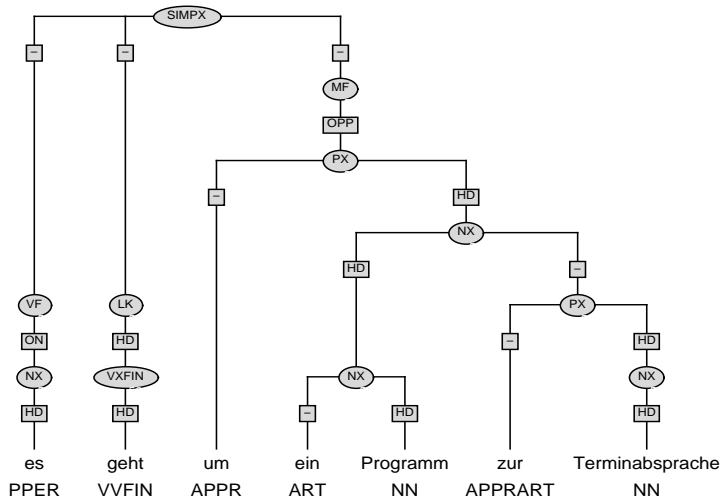
The expletive *es* is the subject of weather verbs (e.g. *regnen*) or other verbs that can denote agentless events. It can also occur in the MF (*morgen geht es um die Terminabsprachen*). It is important that there is an agreement relation between the expletive element and the finite verb. The subject position is filled with the formal subject *es*, which is used because of syntactic reasons.

2. *es* as a subject expletive

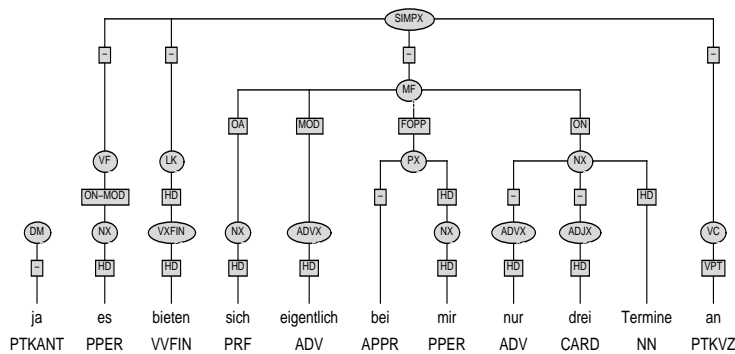
The expletive *es* can only occur in the first position (**nur der zwölfte und der neunzehnte würde es in Frage kommen*.) while the subject of the sentence is a constituent in the MF. There is no agreement relation between *es* and the finite verb but between the subject and the finite verb.

The differences mentioned above are the reason for distinguishing the two kinds of the expletive *es* by different edge labels:

- The verb in the following example is lacking a subject, therefore the subject position is filled with the formal subject *es*, having ON as edge label.



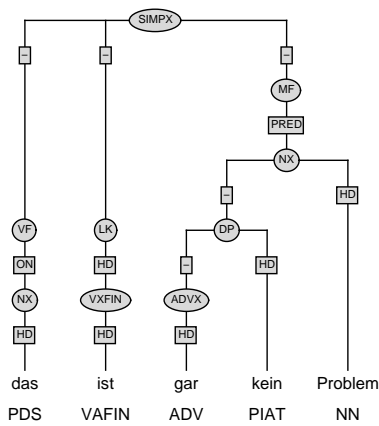
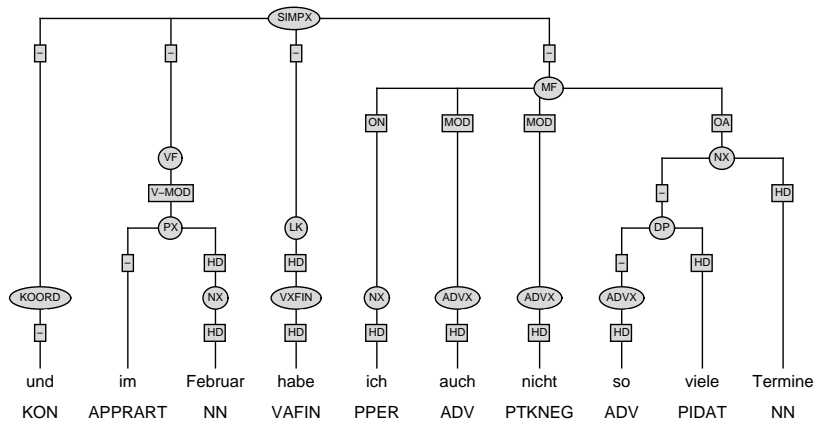
- In the example below, the expletive *es* occupies the subject position but the subject of the sentence is in the MF. That is the reason why the expletive *es* is labelled as an element (ON-MOD) modifying the subject.



4.2 Determiner Phrases

Certain pronouns serving as determiners in noun phrases may be modified, for instance, by degree adverbs such as in *so viele Termine*, *gar kein Problem*, etc.

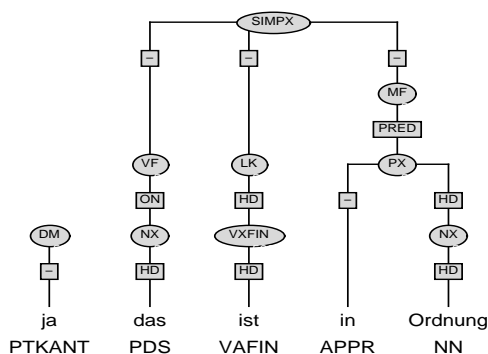
In the case of *so viele Termine*, the premodifying [ADVX *so*] is attached to [PIDAT *viele*]. Together, they constitute a determiner phrase (DP), which is then attached to the head noun *Termine* on the same level:



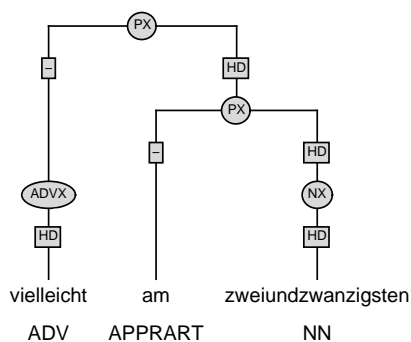
4.3 Prepositional Phrases

4.3.1 Prepositions

In order to facilitate the identification of dependencies between verbs and their nominal complements and adjuncts and in keeping with basic assumptions in Dependency Grammar, not the preposition itself but the complement in prepositional phrases is annotated as the head of the phrase:



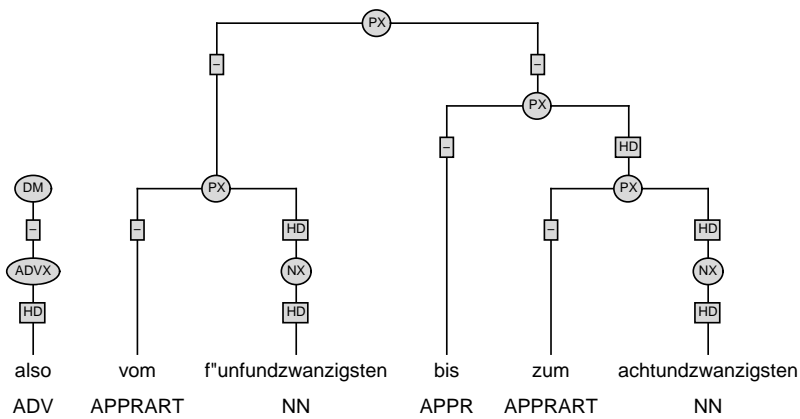
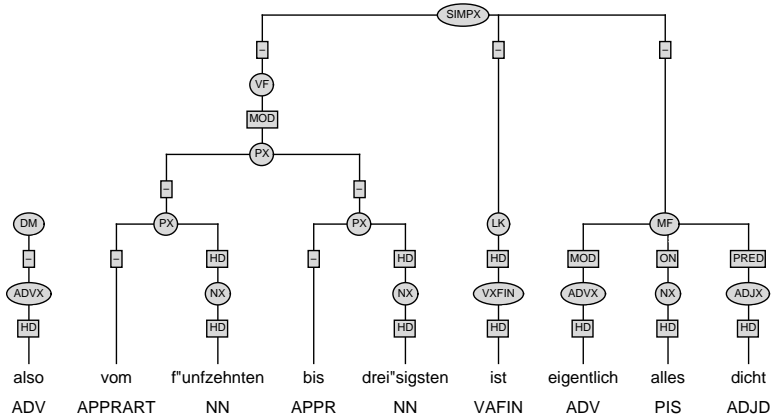
In German, there are so-called *Verschmelzungsformen*, merged forms of a preposition and a determiner, e.g. *zu dem Beispiel* merges to *zum Beispiel*, *an dem zweiundzwanzigsten* merges to *am zweiundzwanzigsten*, etc. The merged form is assigned the POS tag APPRART. These prepositional phrases are annotated in a form, that is consistent with basic prepositional phrases, the merged form acting as preposition:



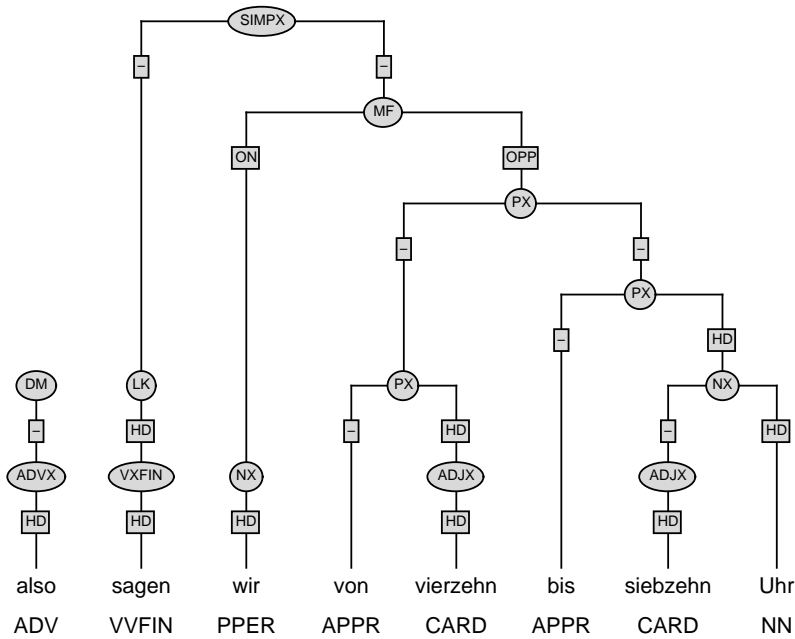
In some syntactic theories, **empty categories** or **traces** are used in order to explain specific phenomena such as the date expressions discussed in the following. In the German treebank, this approach was not feasible because such an annotation would not go conform with the processing considerations (cf. chapter 2), which favor syntactic surface structure. Furthermore, following this strategy would mean to take a theory-driven rather than a data-driven point of view towards annotation.

Closely related and similarly problematic are intervals with *bis* or *zwischen*.

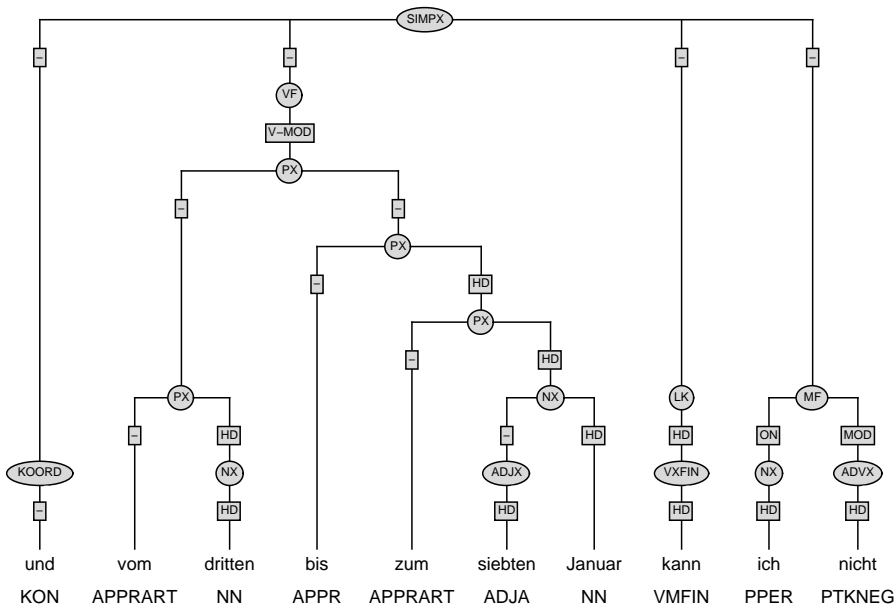
This problem occurs, for example, with intervals with *von/bis*. There is no head since the two phrases are considered to be equal in status:



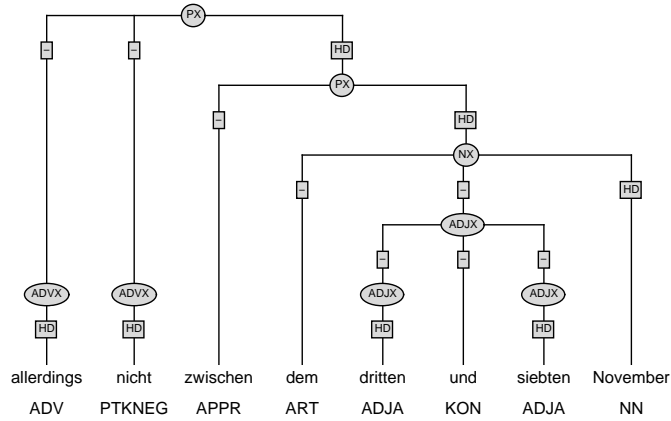
In the following example, it is assumed that *Uhr* serves as the head of both *vierzehn* and *siebzehn*. Therefore, *vierzehn* is tagged as ADJX rather than as NX. There is no need to insert a trace here:



In contrast to time expressions including *Uhr*, date expressions including days and months are annotated differently. In the following example, *dritten* has to be projected to NX, not to ADJX, because it might be the case that it refers to some month other than *Januar*, perhaps mentioned in a previous sentence.

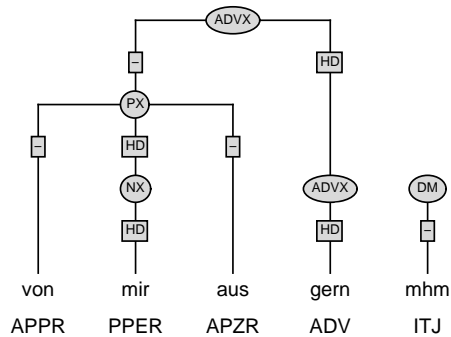


Zwischen requires a complement consisting of a coordination (e.g. adjectives, specifying days of months):



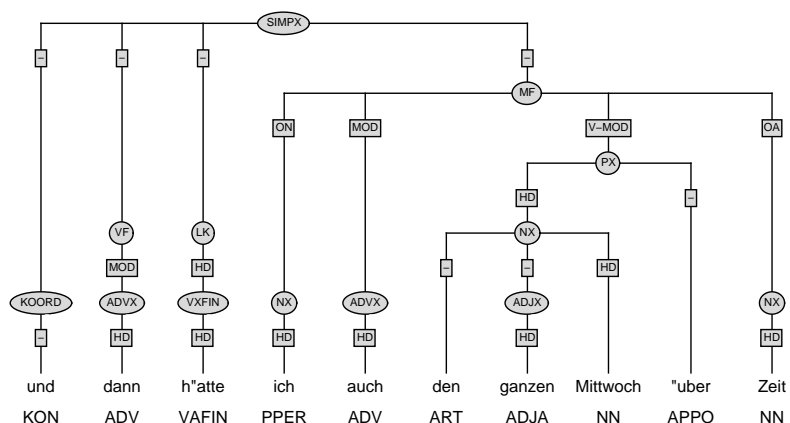
4.3.2 Circumpositions and Postpositions

Circumpositions are treated as ternary branching PXs. The preposition on the left is tagged as APPR and the postposition on the right as APZR. Again the complement is the head of the PX:



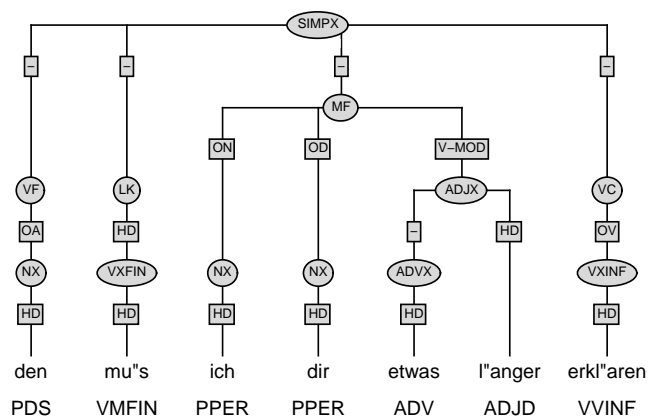
Other circumpositions are treated in the same way: *vom sechsten, siebten Februar an*

Postpositions are tagged as APPO. The complement of the postposition occurs on the left and constitutes the head of the prepositional phrase:

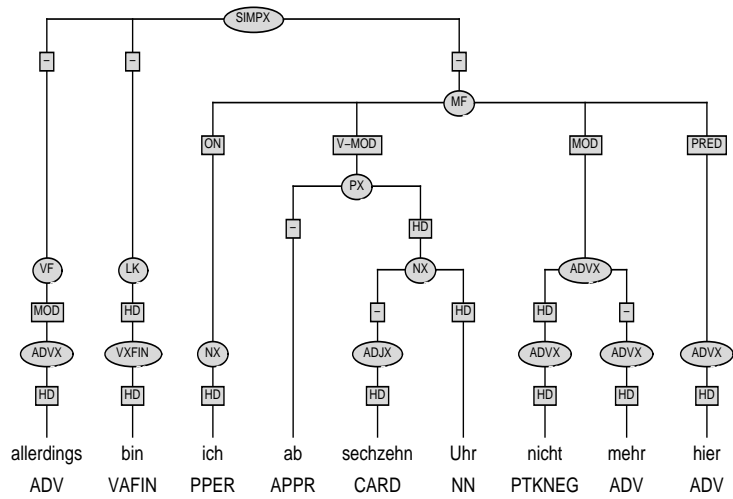
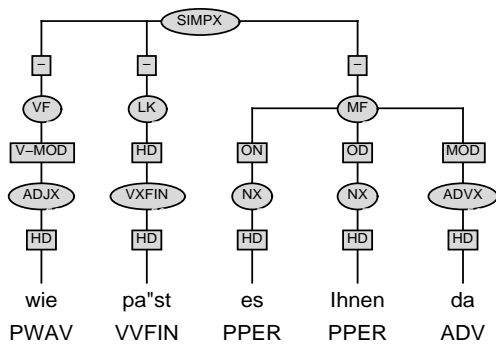
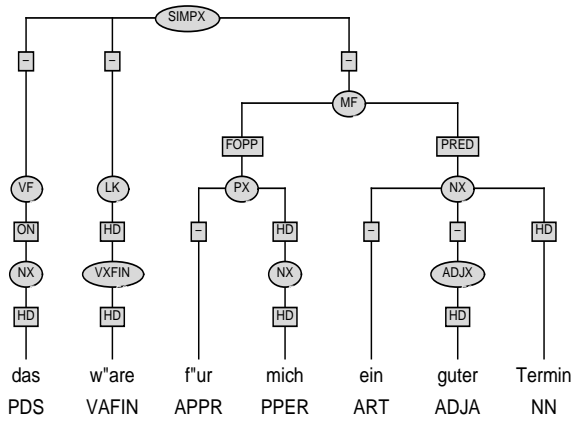


4.4 Adjectival Phrases

Part-of-speech tags of adjectival phrases are ADJD (*das ist gut*), ADJA (*der frühere Termin*), PWAV (*wie paßt Ihnen Donnerstag*), or CARD (*fünf Uhr*). The internal structure of attributive adjectival phrases are annotated in the same way. A few examples of ADJXs:

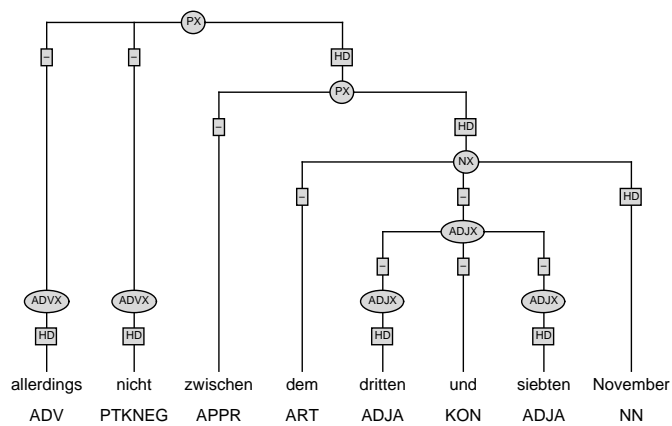
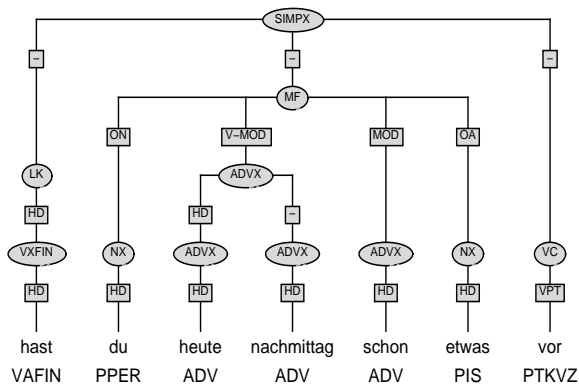
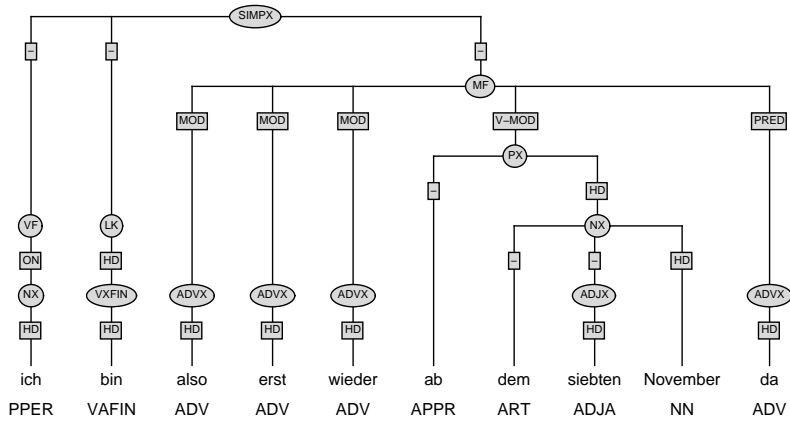


Stylebook for the German Treebank



4.5 Adverbial Phrases

Note that not only ADV projects to ADVX, but also PTKNEG. A few examples of ADVXs:



4.6 Verb Phrases

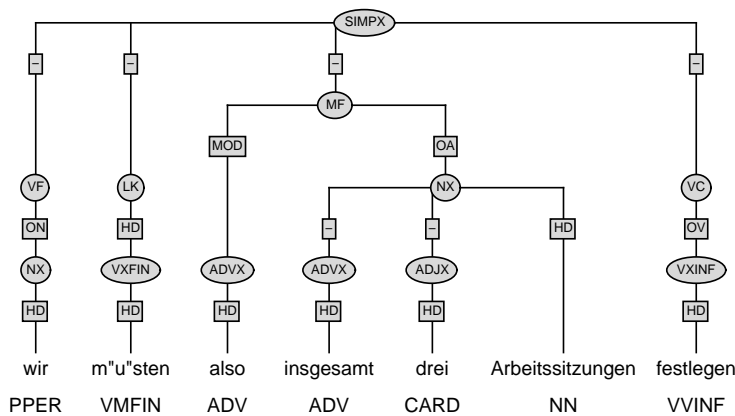
Within verb phrases, VXFİN labels finite verb phrases, VXINF labels non-finite verb phrases and participle phrases.

Since infinitives and participles share certain properties (e.g. exchangeability in *Er hat ihn kommen hören/gehört.*), they are assumed to carry the same phrase label (VXINF).

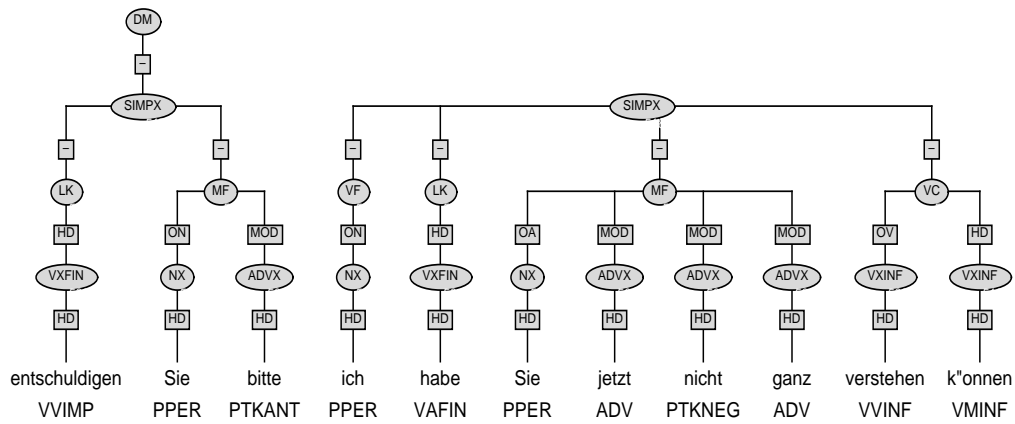
4.6.1 Head of a Sentence and Verb Complex

The finite verb, which can either appear in LK (verb-first clauses and verb-second clauses) or in VC (verb-final clauses), is always the head of the entire sentence. Non-finite verbal elements belong to VC. If the finite verb is located in LK and if there are more than one non-finite element in VC, the non-finite element which is selected by the finite verb is denoted as the head of VC. All other elements of VC are verbal objects. To denote the dependency relations within the verb complex, we distinguish the verbal objects by attaching different *secondary edge labels* to them. The head of VC selects the verbal object OV. This verbal object may select a further verbal object OV carrying the secondary edge label *ref1*, which itself may select a further verbal object OV with the secondary edge label *ref2*, and so on.

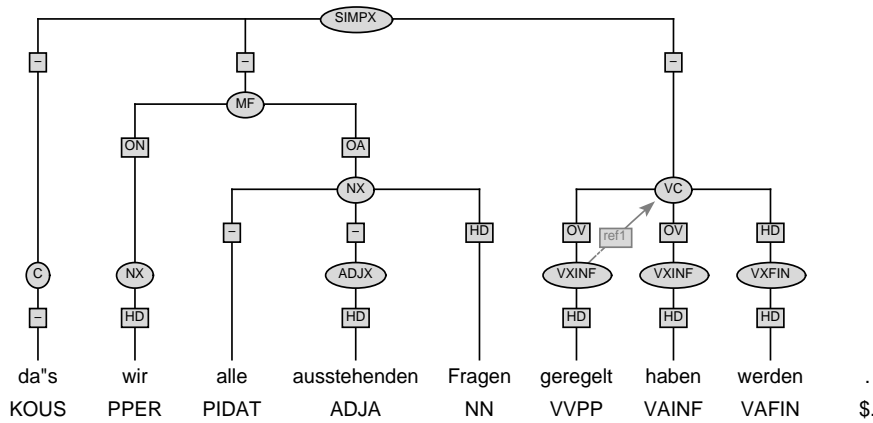
The following example shows a verb-second clause with the head of the sentence in LK and a verb complex consisting of a single non-finite element.



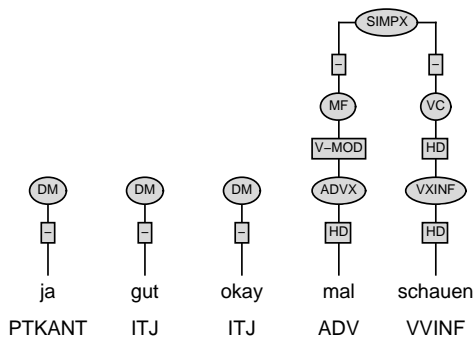
If the verb complex consists of more than one immediate daughter, the one that is selected by the finite verb is the head of VC.



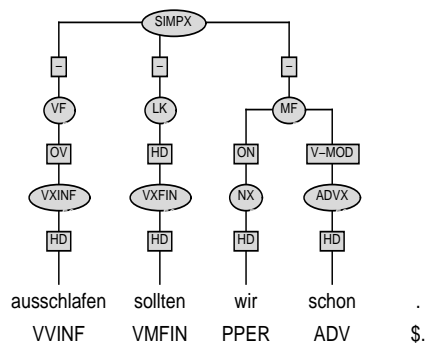
If more than one verbal object is present in VC, the first verbal object that is selected by a verbal object carries a secondary edge label ref1:



If there is no finite verb at all, the rightmost element of the verb complex (if there is more than one element) is annotated as the head of the sentence:

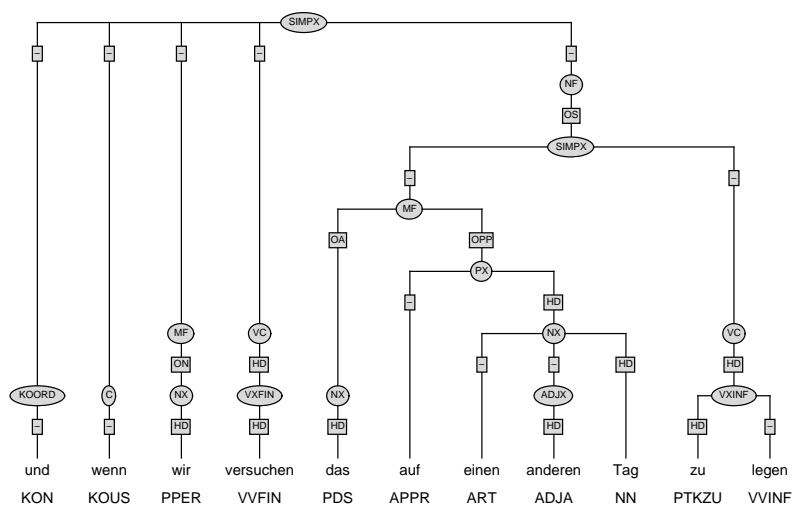


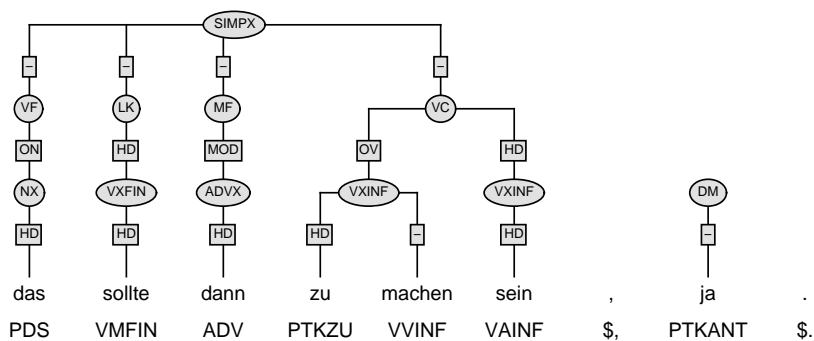
If the verb complex is topicalized, the non-finite verb phrase is not projected to the NF but to the VF.



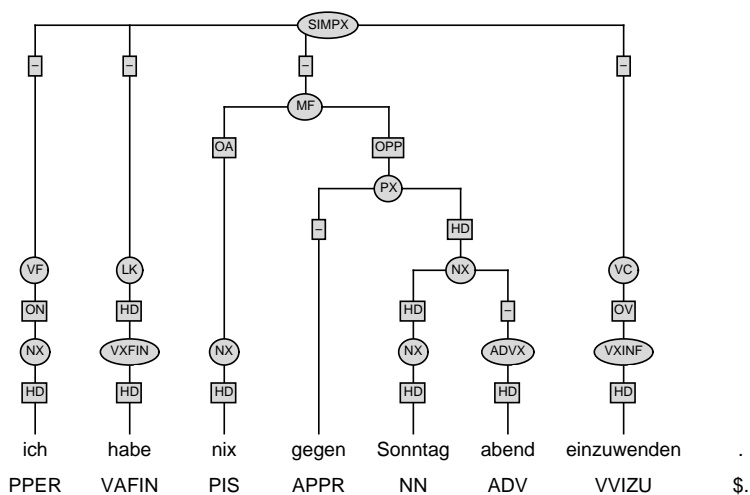
4.6.2 Infinitives with *zu*

For infinitives with *zu*, *zu* is considered the head, the infinitive is considered the complement since *zu* determines the infiniteness of the verb on its right:

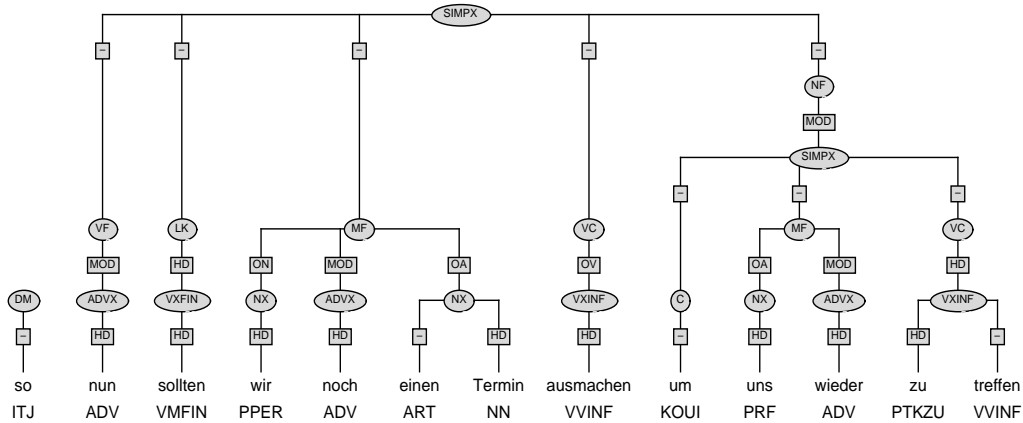




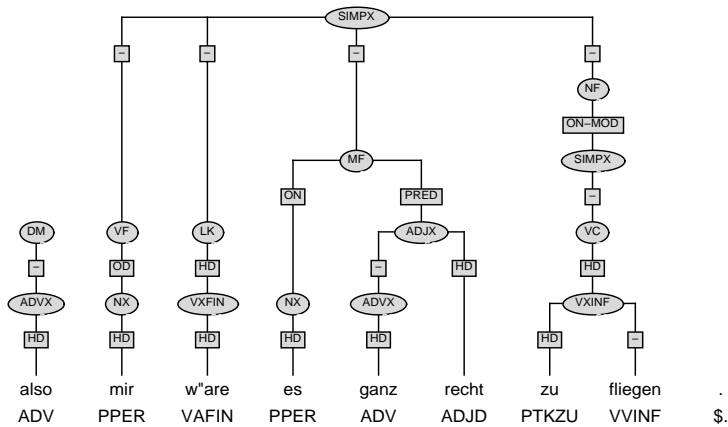
The infinitive with *zu* can also be realized as an infix of the verb. In this case, the verb is tagged as VVIZU and directly grouped as VXINF with the function OV:



Optional infinitive clauses with *um zu* for example can be treated in the same way (see also section 6.1) :



Infinitive clauses consisting only of *zu*-infinitive are annotated as SIMPX with just one field (VC):



4.6.3 Imperatives

Only **second person singular** verbs are tagged as **VVIMP** or **VAIMP** (e.g. *schlage (du) mal einen Termin vor/entschuldigen Sie bitte*), first person plural verbs (*nehmen wir doch den Montag*) are no imperatives in the German treebank.

In verb-second clauses that could be both imperatives or questions, for instance, imperative is not annotated, since it would exclude the normal, “unmarked” reading of the sentence. Note that for most utterances in VERBMOBIL dialogs, imperative will not be annotated because of this kind of ambiguity:

schlagen Sie etwas vor!
schlagen Sie etwas vor?

Second person imperatives can be identified, for instance, by different morphological verb forms:

schlag(e) (du) mal etwas vor!

instead of

schlägst du mal etwas vor?

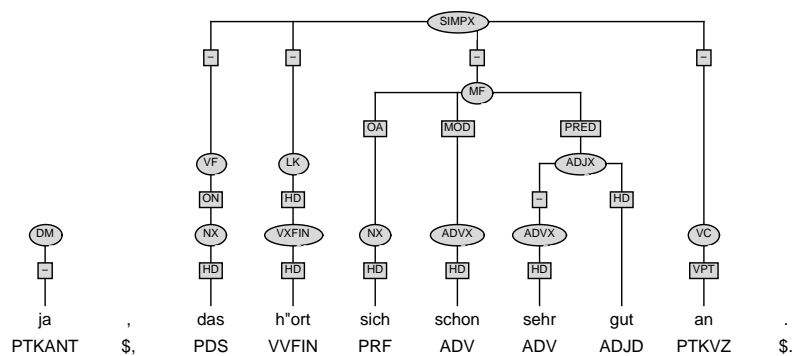
or by the co-occurrence of particles that exclude a non-imperative meaning of the sentence. The following sentence cannot be interpreted as a question:

entschuldigen Sie bitte!

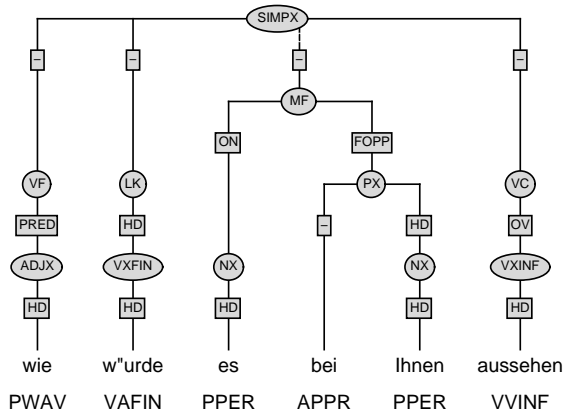
But note that the presence or absence of a subject cannot be used as a criterion in order to detect imperatives, since in the VERBMOBIL dialogs the polite form *Sie* is used in most cases instead of *du*.

4.6.4 Particle Verbs

Separable verb particles are annotated with the edge label VPT:



The particle verb may also occur unseparated within the verb complex:

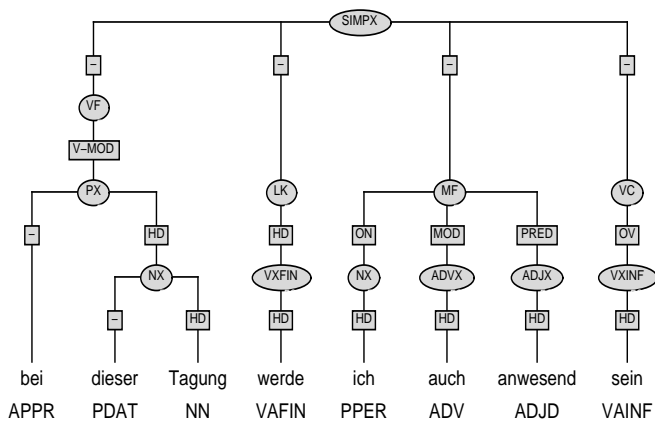


Some examples for particle verbs: *anfangen*, *festhalten*, *vorhaben*, *aussehen*, *ausmachen* etc.

4.6.5 Verbs with Predicate

Typically, the complement type PRED (predicate) occurs with verbs like *sein*, *haben*, *scheinen*, *aussehen*, *sich anhören*, *sich buchstabieren* etc. PRED is annotated, if the following conditions apply:

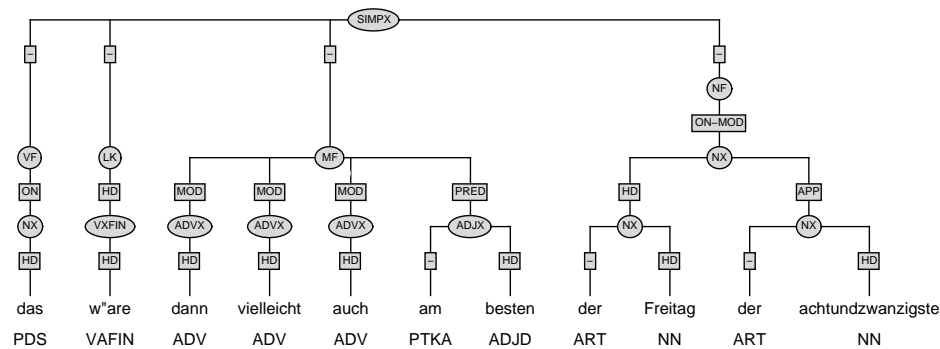
- if it is not possible to ask for the case of the constituent in question properly (e.g. *gut* in *Das ist gut*.)
- if the constituent in question actually predicates the subject, i.e. the subject is characterized as having the property expressed by PRED (e.g. in *Montag ist schön*. *Montag* is characterized by the property of being nice)
- many PRED verbs are raising-verbs (subject without theta-role)



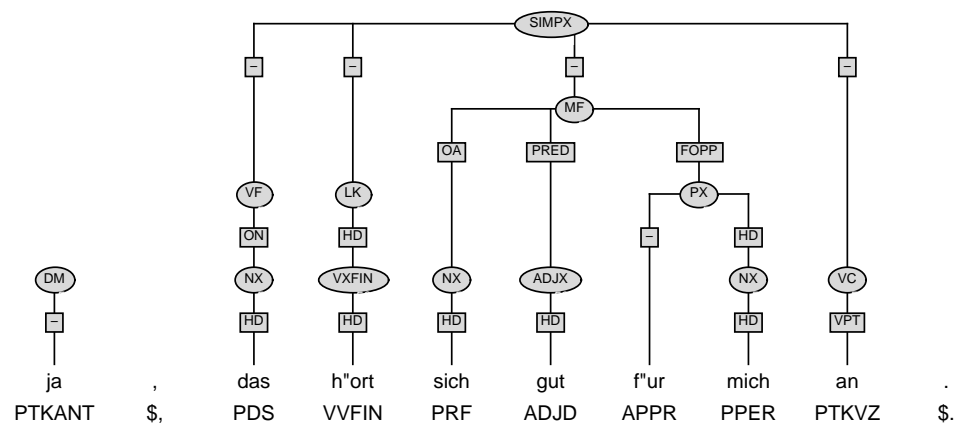
Some examples for verbs that take predicates: *recht sein*, *recht haben*, *leid tun*, *frei sein*, *fertig sein*, *sich gut/schlecht treffen*, *gut finden* etc.

PRED verbs have to be distinguished carefully from verbs occurring with ordinary modifiers (V-MOD) such as *gut passen*.

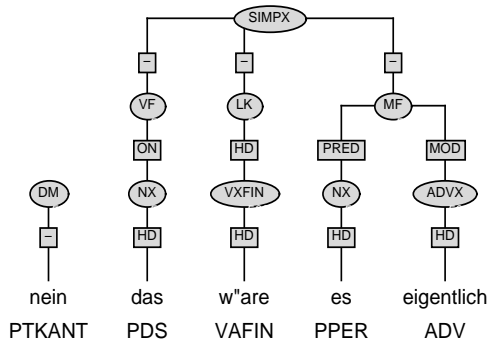
With respect to topological fields, note that PRED marks the border between MF and NF, i.e. whatever constituent occurs **on the right** of PRED necessarily belongs to the NF:



But depending on the order of the elements within the MF, PRED does not necessarily constitute the border between the MF and the NF. Another constituent, for example, can occur between PRED and VC:

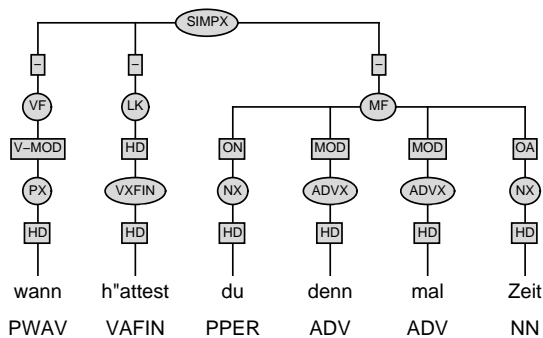


If PRED is a pronoun we have to follow the word order regularity that pronouns in the MF have to precede other constituents:



4.6.6 Verbs with OA Marking the Border Between MF and NF

Some of the transitive verbs with accusative object (OA) almost seem to form a lexical unit inasmuch the meaning of the verb seems to be determined by the OA to a very high degree, e.g. *Zeit haben*:



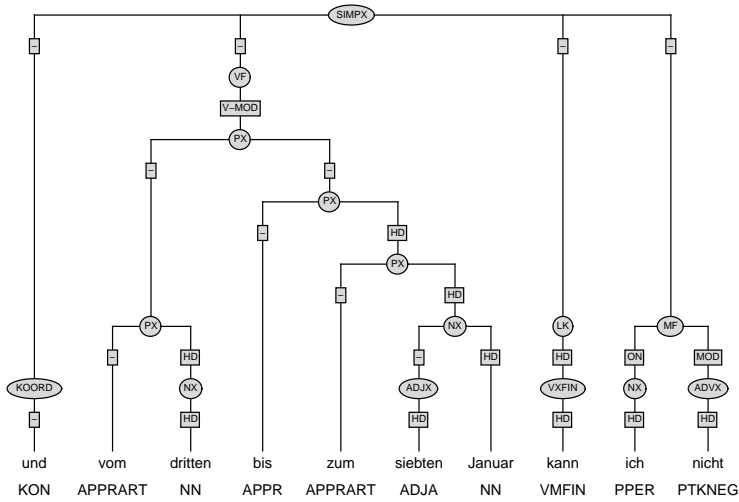
Examples for such verbs with accusative object: *Pech haben* etc.

For these “almost lexicalized” forms, the OA marks the border between MF and NF (as PRED does for the predicative verbs). I.e. the OA occurs within MF, but further constituents on its right have to be located in the NF. But note that not all verbs with OA have this field-separating function. For instance, in the sentence

1. *ich habe den Termin vielleicht/schon.*
vielleicht/schon is located in the MF, whereas in the sentence
1. *ich habe Zeit vielleicht/am Montag.*
vielleicht/am Montag is located in the NF.

4.6.7 Modal Verbs

If the modal verb is the main verb of a sentences, verbal modifiers refer to the modal verb the same way they refer to other main verbs:



Chapter 5

Attachment Principles for Phrases

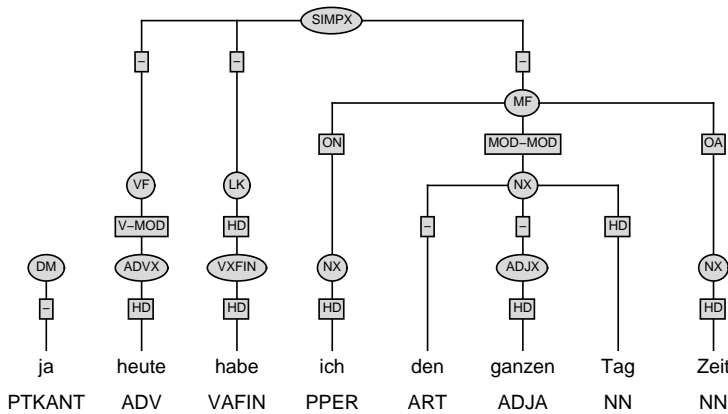
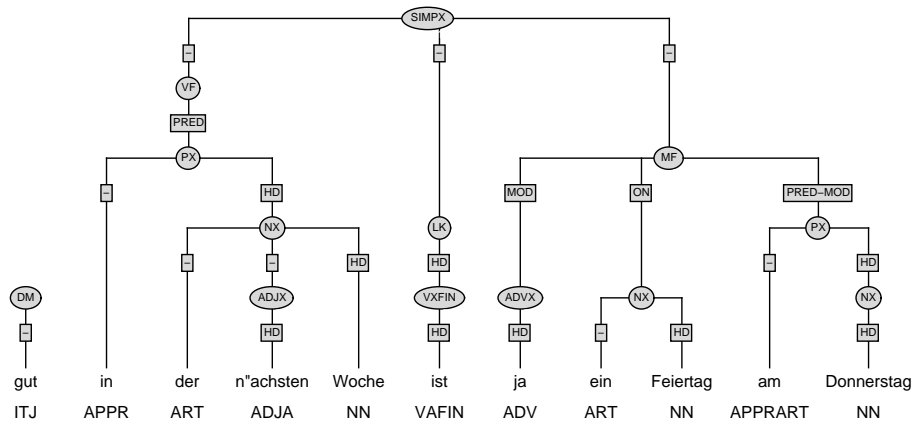
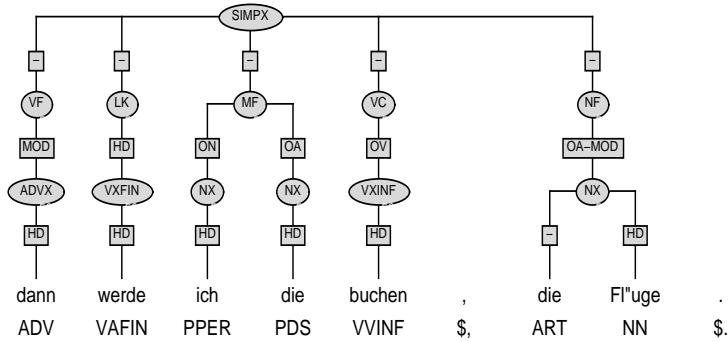
5.1 Attachment to Fields

Phrases are attached to the topological field in which they occur. Their edge labels denote the respective grammatical function within the sentence structure. Note that in LK and VC there can only occur verb forms, separable verbal prefixes, or infinitive particles. LK and VC mark the beginning and the end of the MF.¹

5.2 Modifier Attachment

A constituent that modifies another constituent within a tree structure is either adjacent or discontinuous. In the first case it is immediately attached to the constituent that it modifies, concerning the attachment rules for phrases. In the second case, the dependency, which can even go beyond the border of topological fields, is indicated by edge labels, that express the non-ambiguity of the modifier (e.g. OA-MOD is the modifier of OA). Thus, edge labels like OA-MOD, V-MOD, OPP-MOD, MOD-MOD, etc. express that the respective constituent modifies **only one** other constituent in the sentence (OA, V, OPP, a modifier, etc.) which is not adjacent:

¹See also section 3.2



If such a modifying constituent is ambiguous (i.e. it modifies more than one constituent, the entire sentence, or a constituent that occurred in previous sentence external utterances), it is attached to its topological field and given the general edge label MOD to preserve ambiguity.

Definition of MOD: a constituent is called MOD, if it cannot be assigned a more specific label, either because it is ambiguous or because there is no more specific label (e.g. for sentence modifiers or for constituents that refer to some sentence external expression).

Definition of X-MOD: X is a variable that can be substituted by OA, OPP, MOD, V, etc. X-MOD labels a modifier which can *either* modify a complete X-constituent *or* a constituent **within** an X-constituent. The latter condition is important in order to avoid the necessity of further specifying that the modifier only specifies the head noun within a PP rather than the entire PP or a constituent within the PP. Furthermore, this strategy keeps syntactic categories and grammatical functions separate.

Typical MODs

Temporal or local expressions like *da*, *dann*, but also certain adverbials or particles like *eigentlich*, *ja*, *vielleicht*, *auch*, *natürlich* show attachment ambiguity and therefore are annotated as MOD. Generally, extended forms with *da* and other pronominal adverbs like *trotzdem*, *deswegen*, *hierauf*, etc. are also MOD.

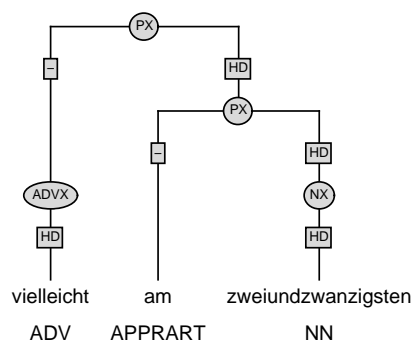
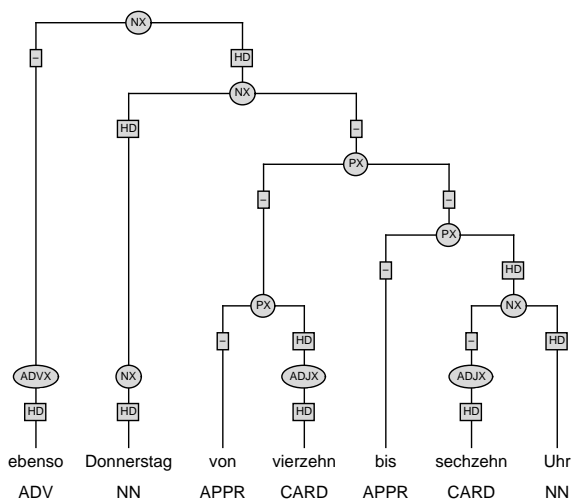
However, non-pronominal adverb expressions such as *vorher*, *später* etc. are V-MOD rather than MOD.

5.2.1 Ambiguous Modifiers Occuring with Isolated Phrases

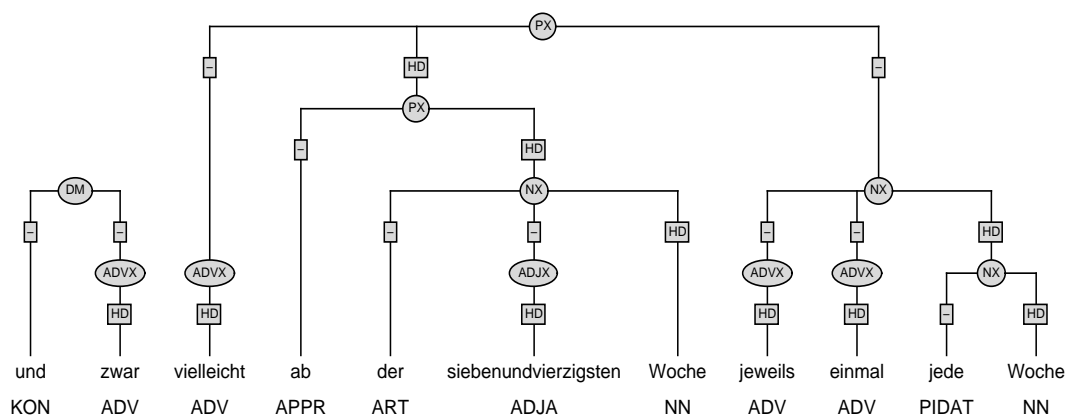
Since isolated phrases do not consist of topological fields, ambiguous modifier (MOD) have to be attached to the phrase itself. The isolated phrase is projected one level higher and the modifier is attached on this level. Thus, the ambiguity information can be preserved even without topological fields or explicit MOD labelling, just by the existence of yet another projection level of the phrase.

In the following examples, *ebenso* and *vielleicht* might refer to something that is implicit or has been mentioned before:

Stylebook for the German Treebank

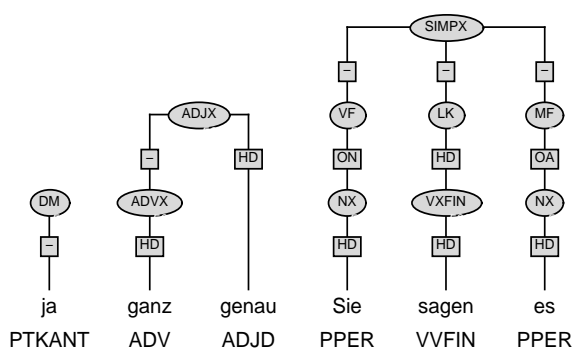


If there is more than one ambiguous modifier in an isolated phrase, all of them are attached on the same level:



The overall attachment strategy described above has been chosen in order to keep syntactic structure flat and to be able to preserve attachment ambiguities where necessary.

Sometimes it is difficult to determine whether a modifier is definite or not. It is often helpful to test whether the two constituents in question can be topicalized together without causing a change in meaning to the entire sentence. If the topicalization is possible under this condition, the two constituents can be considered a syntactic unit, i.e. the modifier is attached immediately to the phrase which it modifies, for instance, *ganz genau* in the following example:



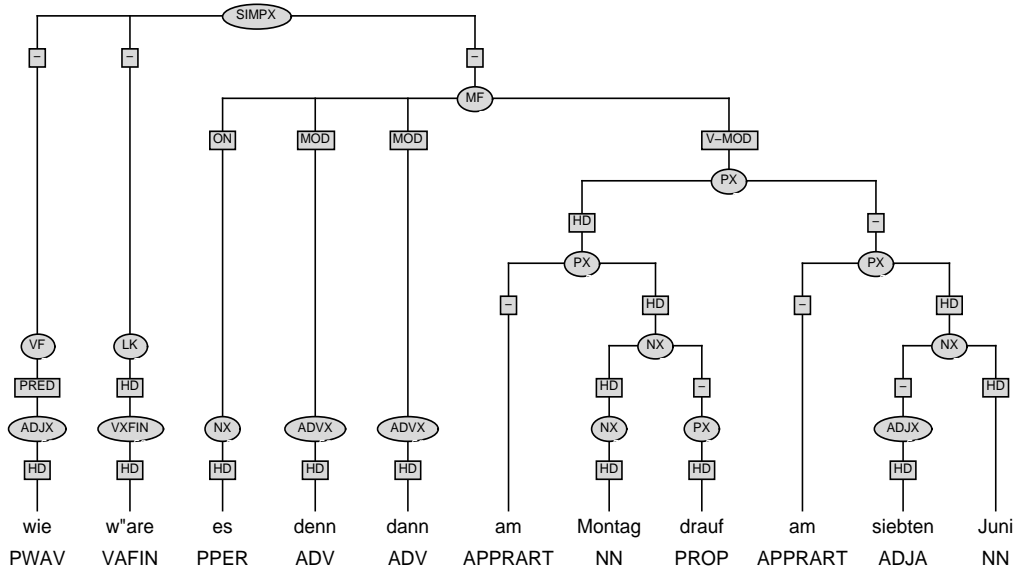
But in cases of doubt, modifiers are marked as ambiguous (MOD) rather than as definite modifiers.

For details about pre- and post-head attachment principles for specific syntactic categories see section 4.1 above.

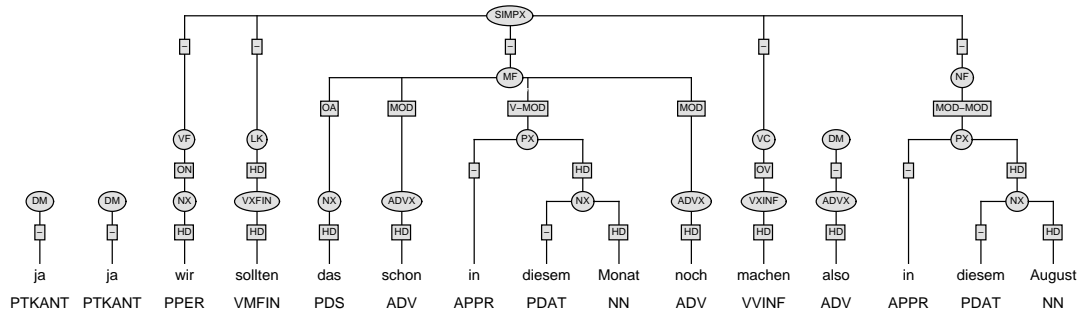
5.2.2 Phrases Modified by Phrases

Sometimes two phrases in a sentence express almost the same information. But one of them is more specific than the other. They can be adjacent like in the following example (*am Montag drauf am siebten Juni*). The phrase giving the more specific information always modifies the more general one. In this respect they are similar to appositions. Both phrases are projected to a complex phrase on the next higher node.

Stylebook for the German Treebank



If the two phrases occur in different fields like in the following construction, the reference of modification is denoted by edge labels (*in diesem Monat* (V-MOD) ... *in diesem August* (MOD-MOD)).



Chapter 6

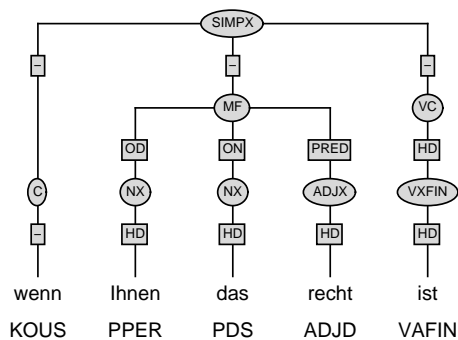
The Annotation of Sentences

The approach of topological fields supports the *flat clustering principle* inasmuch the MF and the NF allow for more than one constituent being attached to the same field node. The field nodes form a level of annotation between the phrase level and the sentence level. The last step to complete a sentence structure is to attach the field nodes to the highest annotation level of the whole structure: the root node.

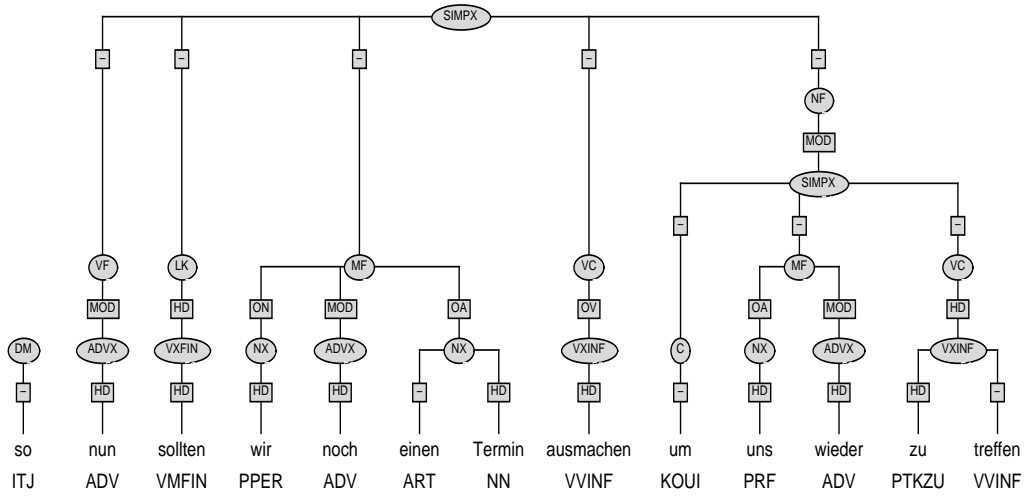
In the following sections the annotation of the different sentence structures that occur in the German treebank will be demonstrated.

6.1 The C-Field in Verb-Final Clauses

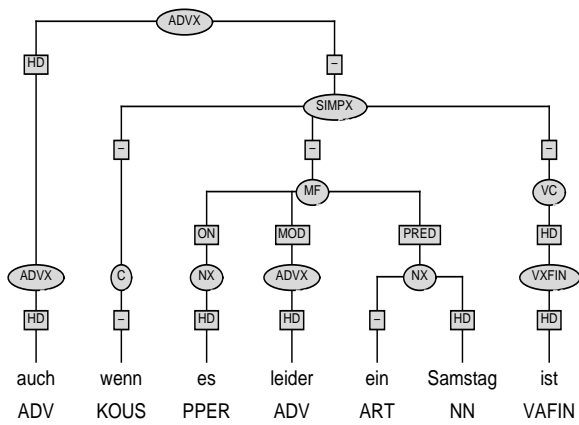
The C-field (complementizer field) is the field for subordinating conjunctions KOUS (e.g. *daß, wenn, da, weil, ob*), KOUI (e.g. *um (+zu)*), relative (PRELS), and interrogative (PWAV) pronouns and (complex) interrogative or relative phrases. Thus, it only occurs in verb-final clauses.



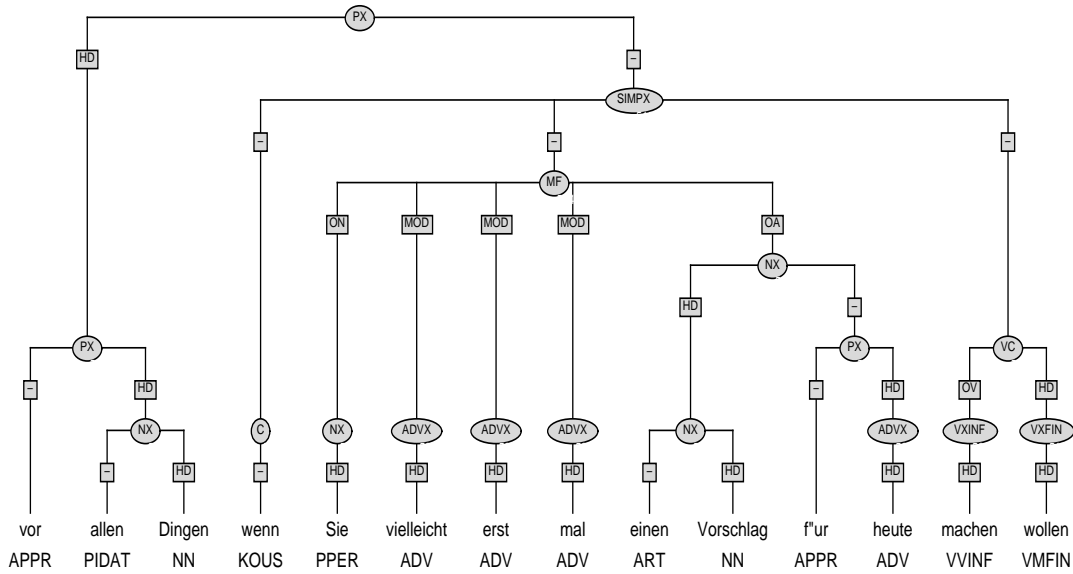
Stylebook for the German Treebank



Since C generally does not contain more than one constituent, *auch* in the following example is not allowed to occur in the C-field together with *wenn*. The *wenn*-clause is rather the modifier of the adverbial phrase *auch*:

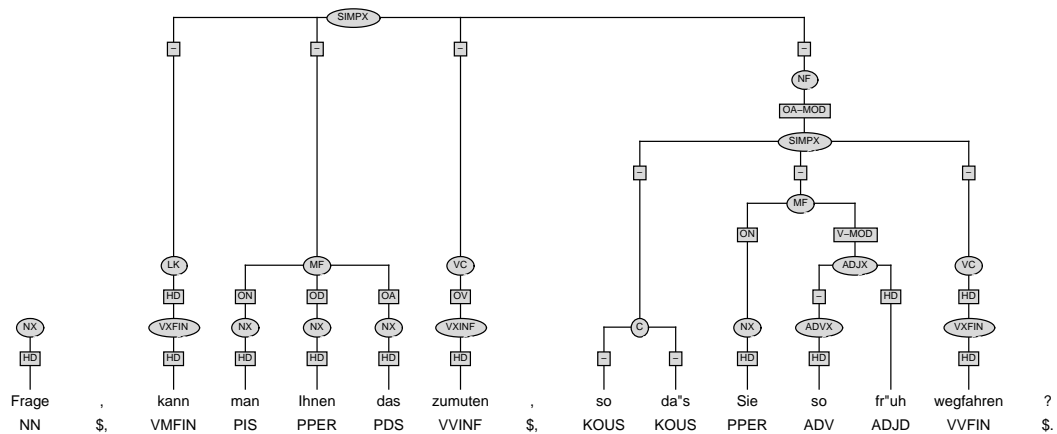


A similar case is *vor allen Dingen, wenn ...*:



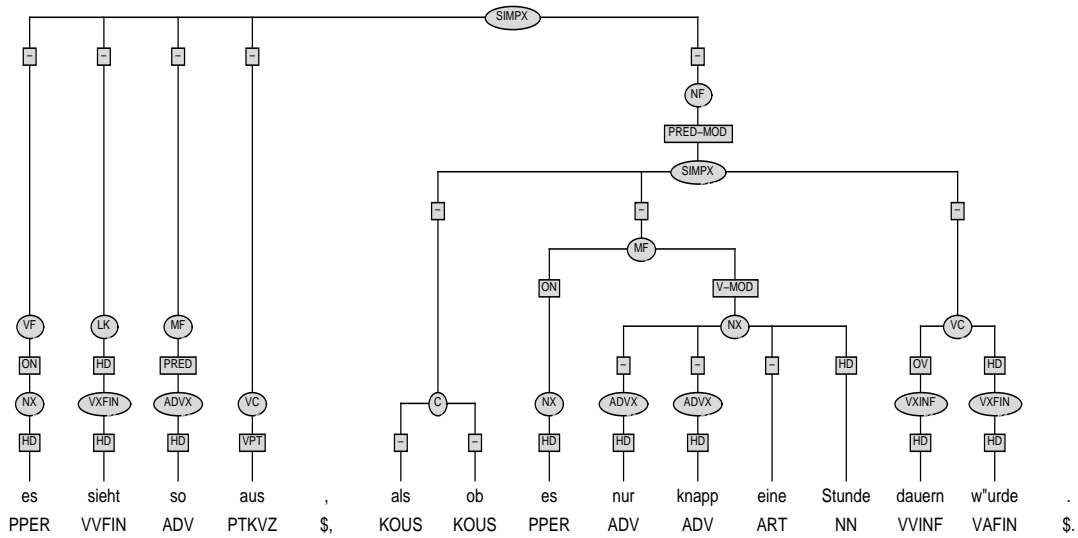
According to Höhle¹, the most plausible analysis is that ADVX or PX respectively subcategorizes for the verb-final clause.

There are conjunctions in German which consist of two elements (e.g. *so daß* and *als ob*). Both of them are directly attached to the C-field, while none of them carries a head label.

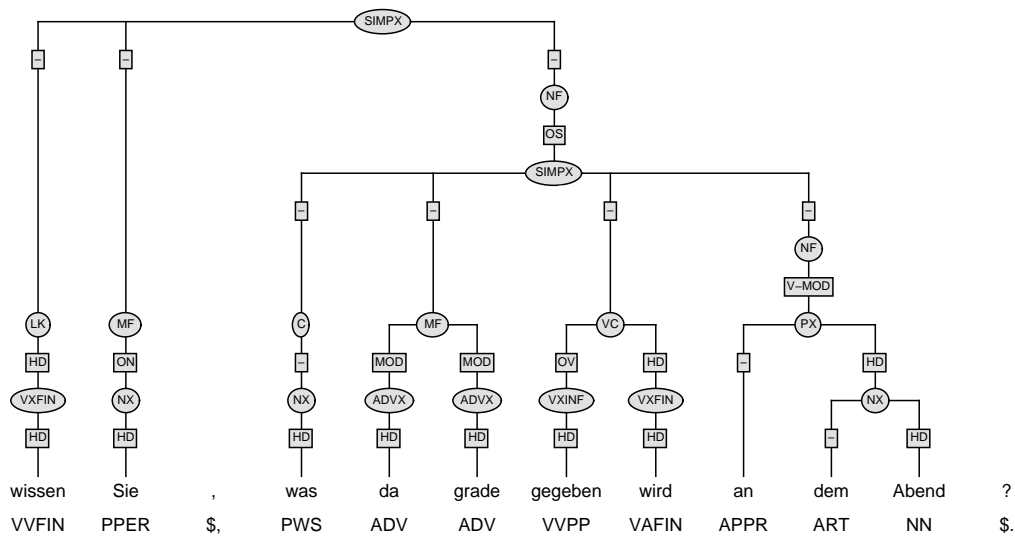


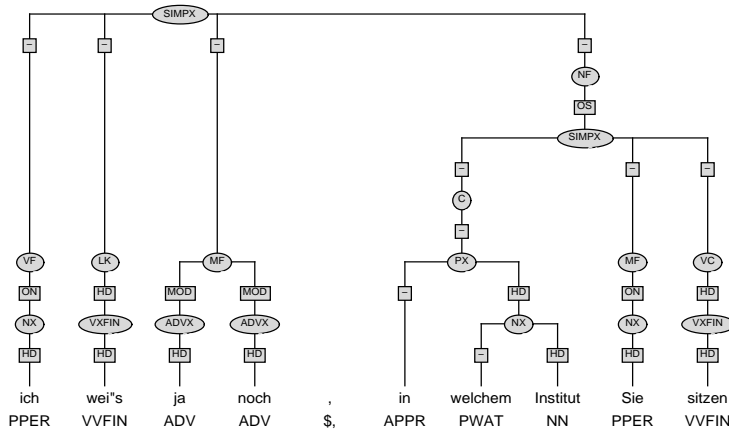
¹personal discussion Stegmann - Höhle, December 1997

Stylebook for the German Treebank



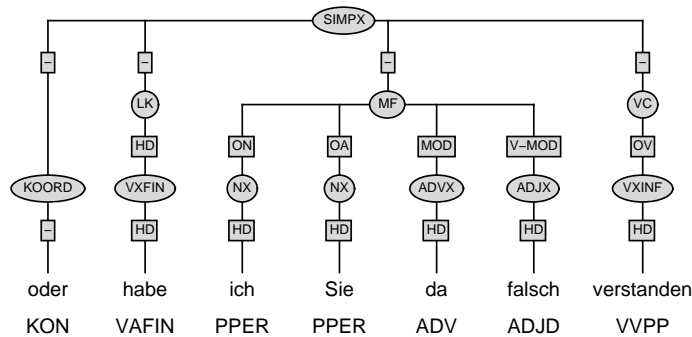
If the constituent in the C-field is a pronoun or a complex phrase it is first projected to the phrase level and then to the C-field.

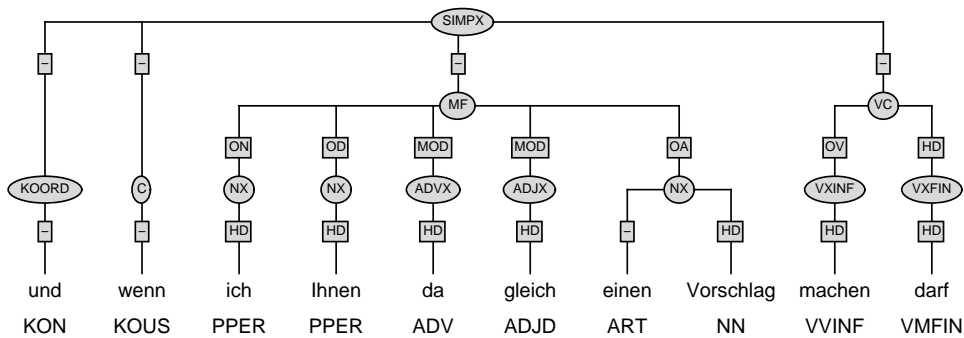
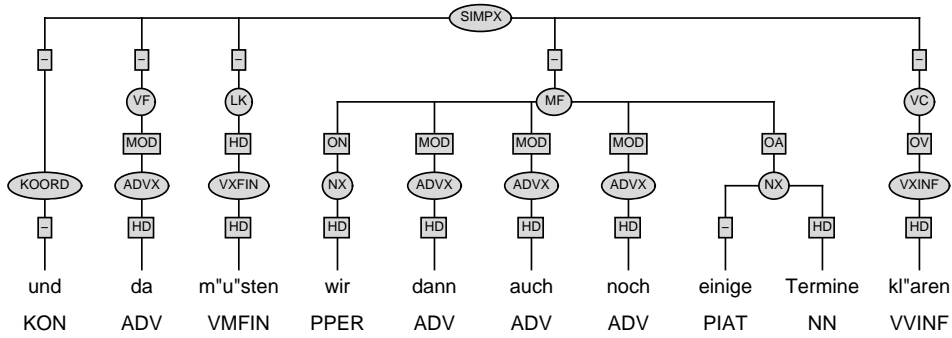




6.2 The KOORD-Field in all Clause Types

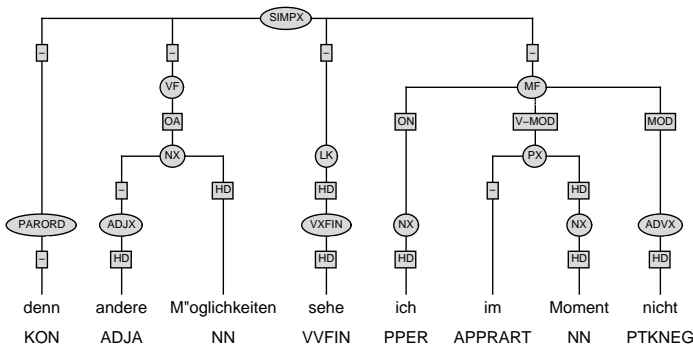
The KOORD-field can optionally occur as the left-most field of all clause types (V-1, V-2, V-end). For verb-second clauses it can be regarded as an alternative field to the PARORD. The KOORD-field contains coordinative particles like *und*, *oder*, *aber*, etc. (cf. Höhle (1985)). Here are a few examples of different clause types:



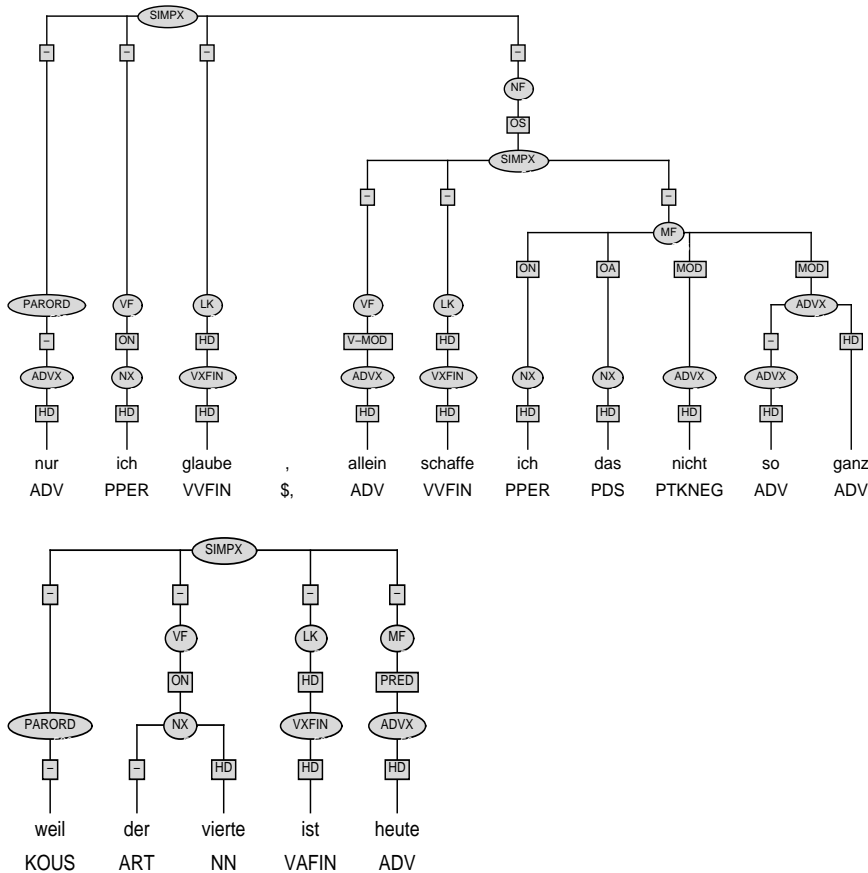


6.3 The PARORD-Field in Verb-Second Clauses

PARORD is an alternative field to KOORD for verb-second clauses only. Typical PARORD expressions are *denn*, *nur*, *weil*²:



² *weil* can occur in verb-second and in verb-final clauses. In the first case, it is grouped in the PARORD-field, in the latter case, it belongs to the C-field.



According to Höhle³, *also*, *allerdings*, *und zwar*, *nun* do **not** occur in PARORD, but in other fields (KOORD, VF, etc.) or have to be analysed as sentence external discourse markers. Examples for all of these expressions will be given below, since they occur very often in VERBMOBIL dialogs:

1. [DM also] [SIMPX mir paßt es sehr gut am Donnerstag.]
2. [DM allerdings] [SIMPX wir müssen ja noch ein Treffen ausmachen.]
Note: *allerdings* is only DM if it does not occur in the VF, like in the following cases, for example:
[VF allerdings schon am achten nachmittags] habe ich Zeit.
[VF allerdings] geht es bei mir nur nachmittags.
3. [DM und zwar] [PX am ersten und zweiten]
4. [DM nun] [SIMPX nach der Reise bin ich erst mal wieder eine Woche lang nicht zur Verfügung.]

³personal discussion Stegmann - Höhle, December 1997

6.4 Resumptive Constructions: The LV-Field

Resumptive constructions are analysed as suggested by Höhle (1985) and Kathol (1995), by using the field LV⁴ (“Linksversetzung”) which is located on the left of the VF:

[LV die Woche davor], [VF wie] sähe es da aus?
[LV Montag der siebzehnte], [VF das] würde mir gut passen.

The typical feature of a resumptive construction is that there is a pronominal constituent somewhere in the sentence, on the right of the LV-field, which refers back to the expression within the LV-field.

In general, the LV-field is not restricted to one constituent:

[LV [die Woche davor] [erste Novemberwoche]] [VF wie] sähe das aus?

Specific LV-Constructions

Grammatical functions within a LV-construction are assigned according to the following principle:

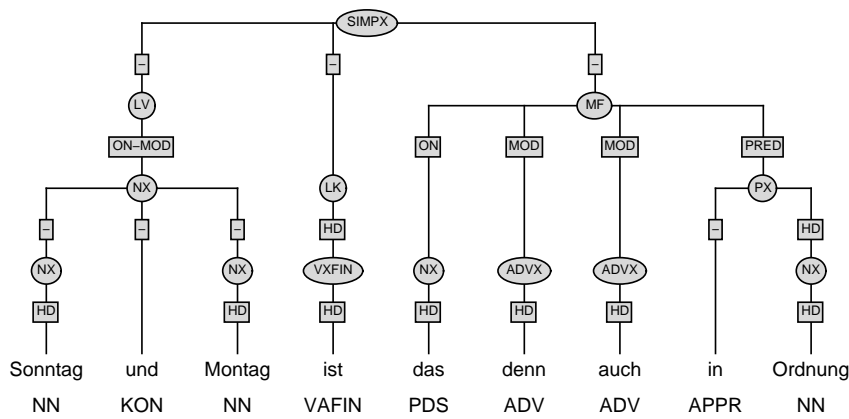
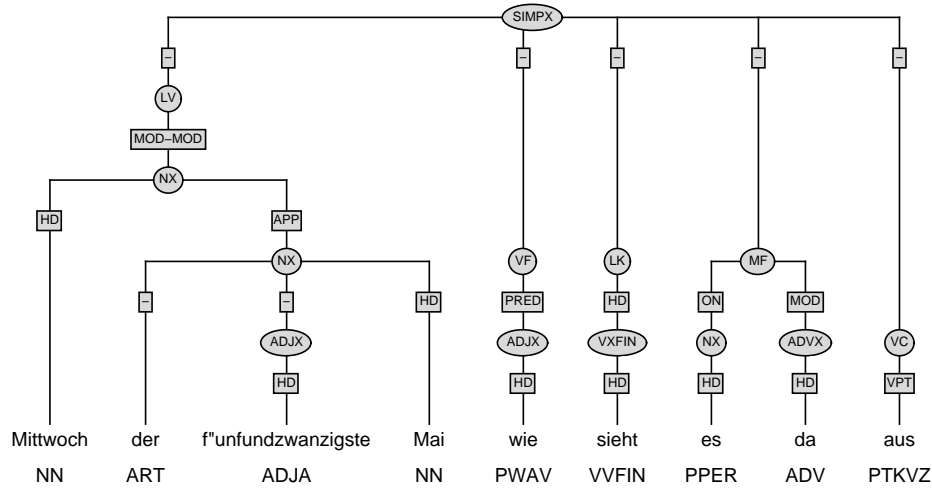
- The LV-constituent is licensed by some (pronominal) constituent within the core sentence. The core sentence exceeds from VF to NF. Therefore, the licensing constituent is considered to be modified by the constituent within the LV-field.

For instance, ON-MOD is licensed by ON, which is also in strong accordance with the assumption that the original position of the subject in verb-second clauses is the VF:

[LV(ON-MOD) Montag der siebzehnte] [VF(ON) das] würde mir gut passen.

However, the licensing constituent does not need to be in VF position. In the following cases, the licensing expressions occur within the MF:

⁴The name of the field has been changed by the authors.



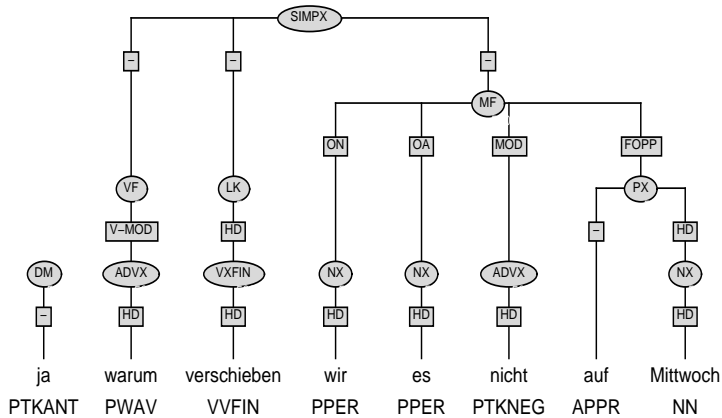
6.5 Questions

6.5.1 W-Questions

In general, w-questions are verb-second clauses with interrogative pronouns in the VF. The problem here is to decide on the syntactic category of the interrogative phrase. We follow the strategy to assign the category of the phrase that would be given as the answer to the interrogative phrase:

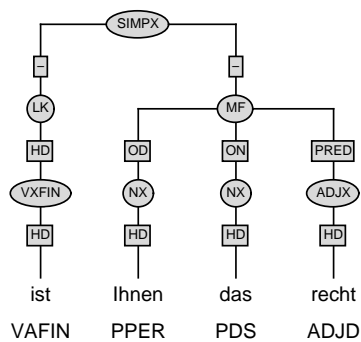
[PX wann] hätten Sie Zeit? → [PX am Donnerstag]
[ADJX wie] wäre es um neun Uhr? → [ADJX gut], [ADJX schlecht]
[NX welche Termine] wären günstig? → [NX der zweite und vierte Januar]

warum is an exception in so far that it is an adverbial interrogative pronoun on the part-of-speech tag level, but the answer to it would be given in form of a sentence. We decided to annotate it as ADVX rather than OS in order to be consistent with its part-of-speech tag information (PWAV):

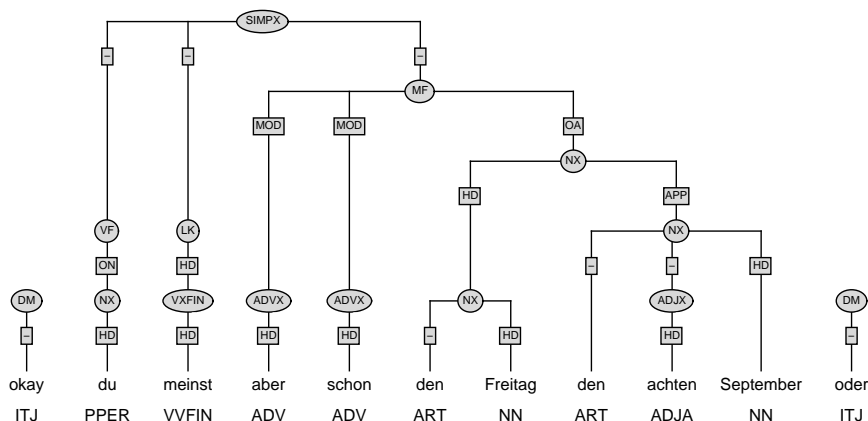


6.5.2 Yes - No Questions

Yes - no questions may occur in various forms, but the most typical form is the verb-first clause:

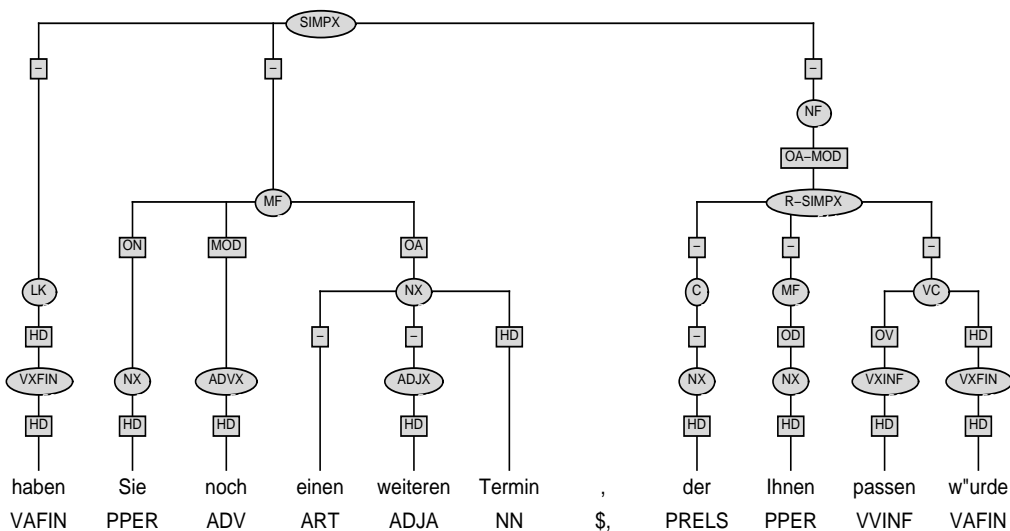


Otherwise, interjections such as *oder*, *ne*, *ja* etc. at the end of a verb-second clause indicate that it is actually meant as a question:



6.6 Relative Clauses

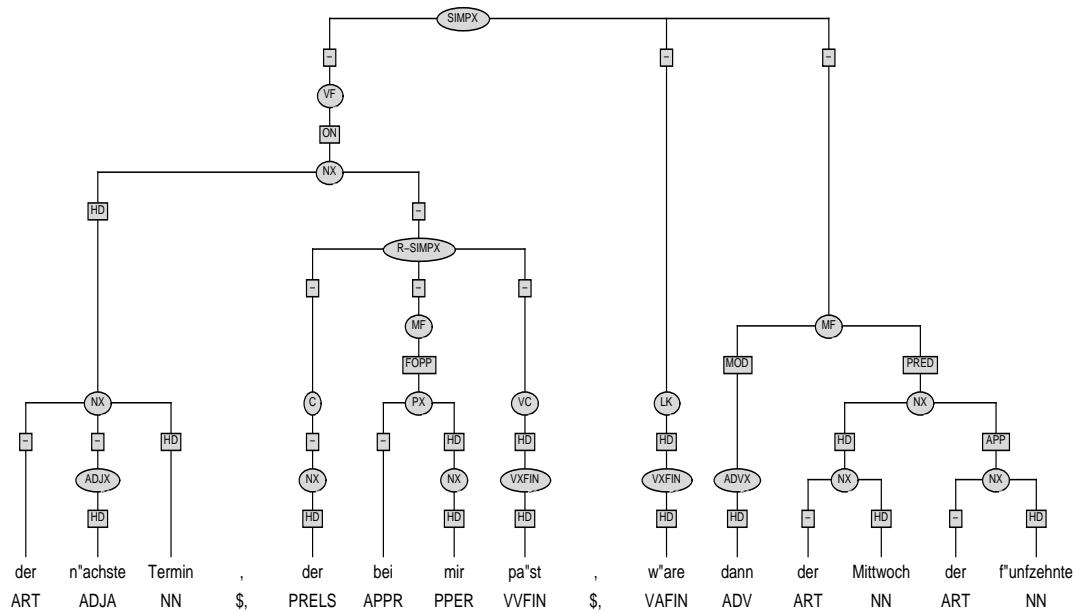
For relative clauses (R-SIMPX), the relative pronoun occurs in the C-field. It is first projected to the phrase level before being attached to the C node. The relative clause itself is located in the NF like in the following example if no other constituent follows. Its edge label shows to which constituent of the matrix clause it is related. OA-MOD, for example, suggests that the relative clause refers to OA:



The position of the relative clause in the NF is justified by the fact that it does not necessarily occur as an immediate constituent located on the right side of the noun phrase it refers to. For example, a verb complex can occur between

the noun phrase and the relative clause (*er hat das Hotel gebucht, das so teuer ist*). In sentences like this the complexity of the noun phrase (NP + relative clause) is important. This so called *heavyness* follows Behaghel's first physical law (Behaghel 1932): complex noun phrases tend to find a position at the end of the sentence even if they deviate from their basic order. If the relative clause does not follow the noun phrase immediately, its unmarked position is in the NF. Unless there is strong evidence for a position in the MF, the relative clause is located in the NF.

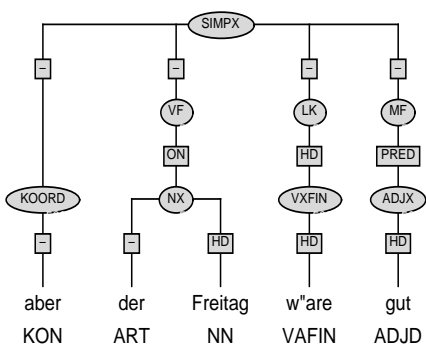
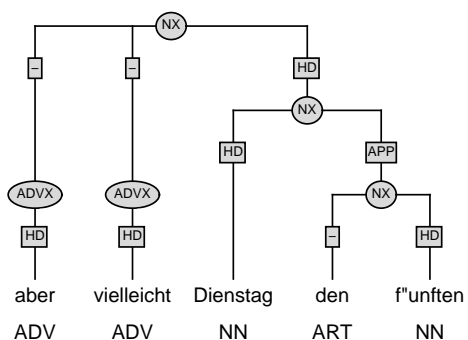
In the following sentence, the relative clause and the noun phrase it refers to are adjacent constituents. As they are located in the VF, the relative clause modifies the noun phrase *der nächste Termin* directly. Here the noun phrase is the head of the complex noun phrase.



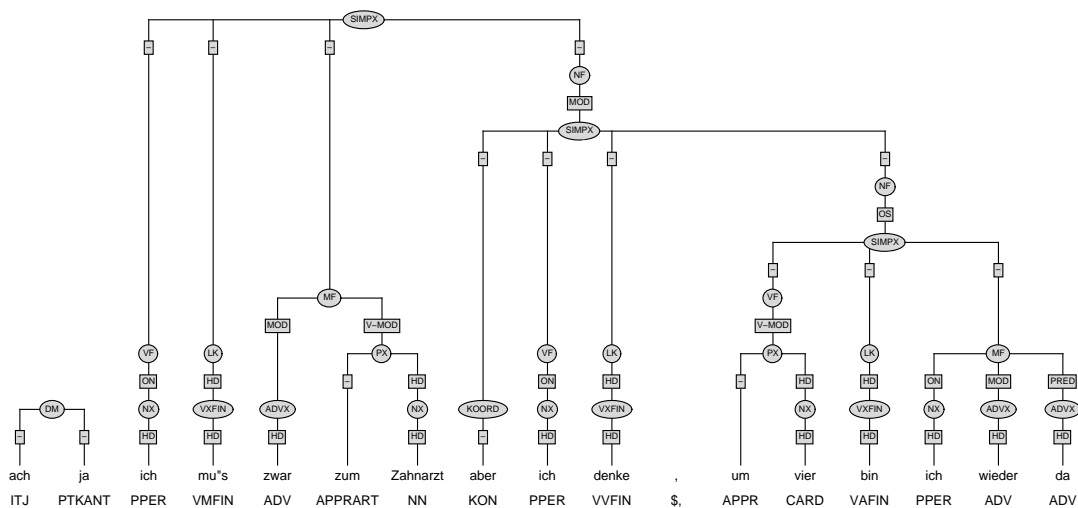
6.7 Constructions with *aber*

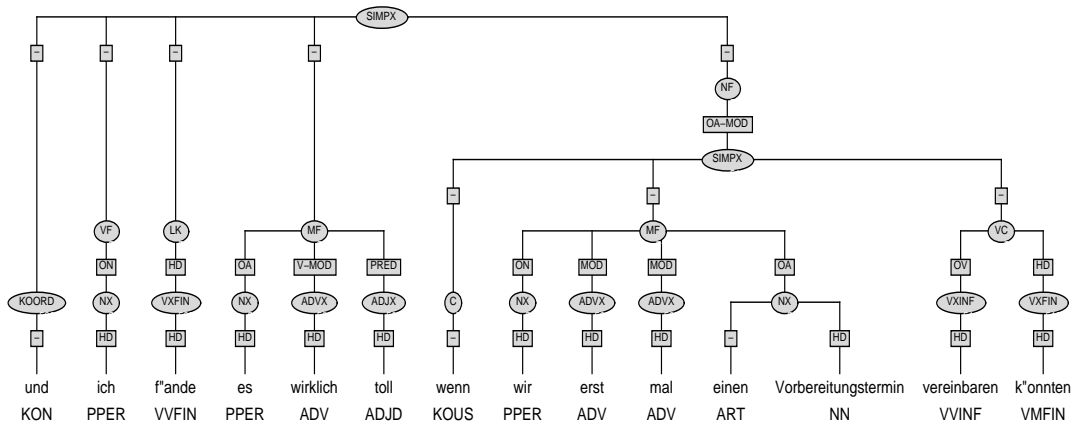
aber either occurs as an adverbial expression (ADV) within sentences, or the beginning of an isolated phrase, or as a sentence initial conjunction in KOORD (part-of-speech tag KON):

Stylebook for the German Treebank



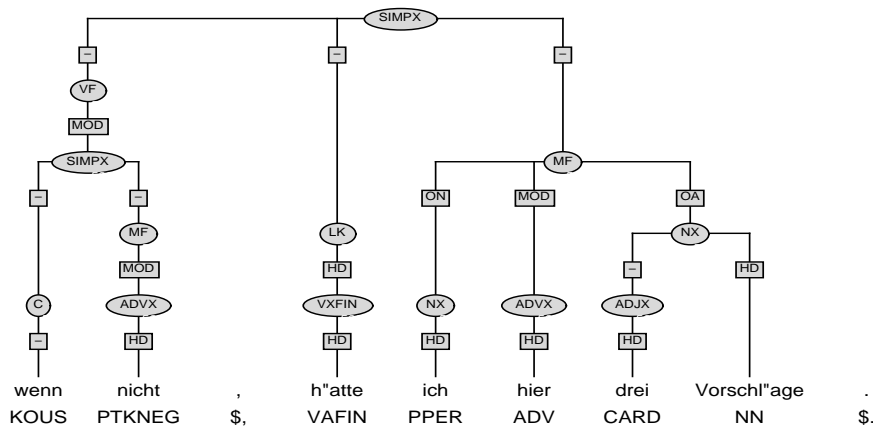
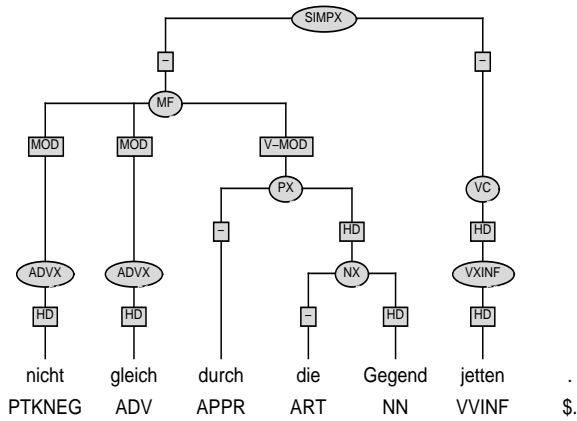
aber also occurs in *zwar ... aber ...*-constructions. The clause with *zwar* and the clause with *aber* have to be attached to each other. In the resulting construction the *aber*-clause is modifying the *zwar*-clause:



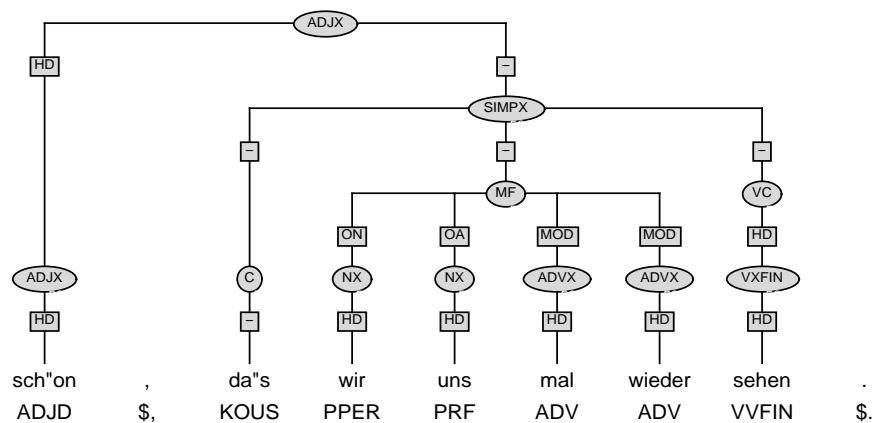


6.9 Elliptic Constructions

In elliptic constructions, syntactically necessary linguistic elements are missing which can be reconstructed from the context or the speech situation. The model of topological fields does not make any assumptions about dependency relations, but it allows that topological fields may be left empty. For the description of elliptical constructions, it is an appropriate model because neither crossing branches nor traces have to be used to annotate the surface structure of a sentence. In the German treebank, elliptic constructions are treated like complete sentences if they either contain a verb or a C-field. In the first example, the subject as well as the finite verb is missing. But since this utterance contains a main verb, it can be annotated as a SIMPX-clause. The *wenn*-clause in the second example is lacking a verbal constituent, but the occupied C-field indicates that there is a modifying elliptic subclause. Thus, its lexical elements are projected up to the sentence level:



In contrast to incomplete sentence structures like above, complex phrasal utterances lacking a verbal constituent but containing a complete sentence are projected up to their phrase level and are modified by the SIMPX-clause:



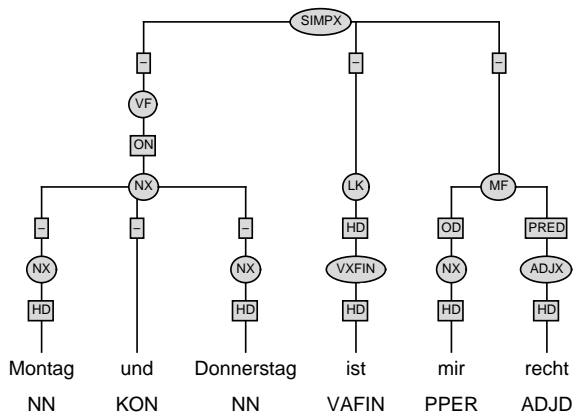
6.10 Coordination

Within coordinations, the conjuncts are first projected to the phrase level, then they are attached to the mother node which is ternary branching (conjunction in the middle). The edge labels below the mother node of the coordination are empty. This scheme is the same for all syntactic categories. In contrast to conjunctions in the KOORD-field, the conjunction in coordinations (*und*, *oder*, etc.) is directly attached to the mother node of the conjuncts. It is tagged as KON.

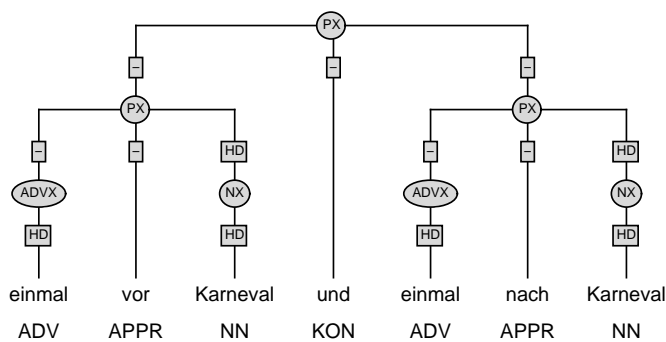
In the following, the coordination of phrases, sentences, combination of fields as well as special cases of coordination will be demonstrated.

6.10.1 Coordination of Phrases

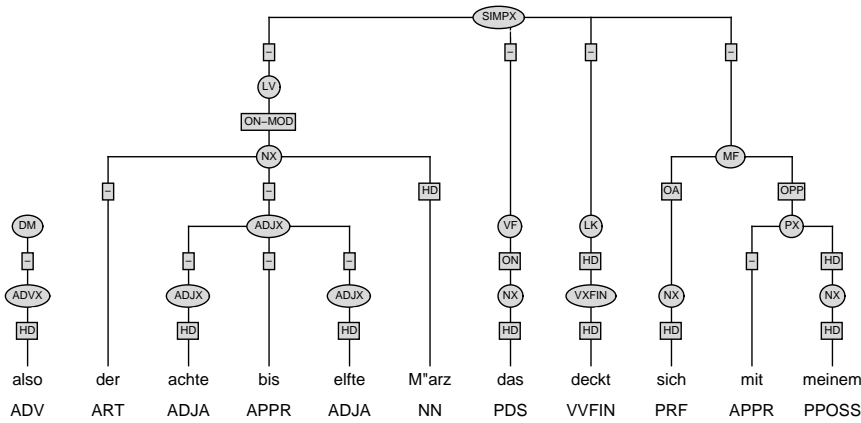
Noun Phrases



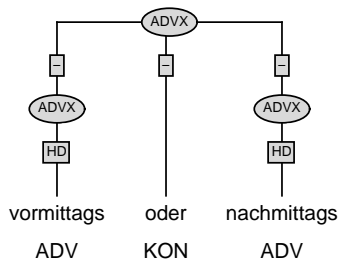
Prepositional Phrases



Adjectival Phrases

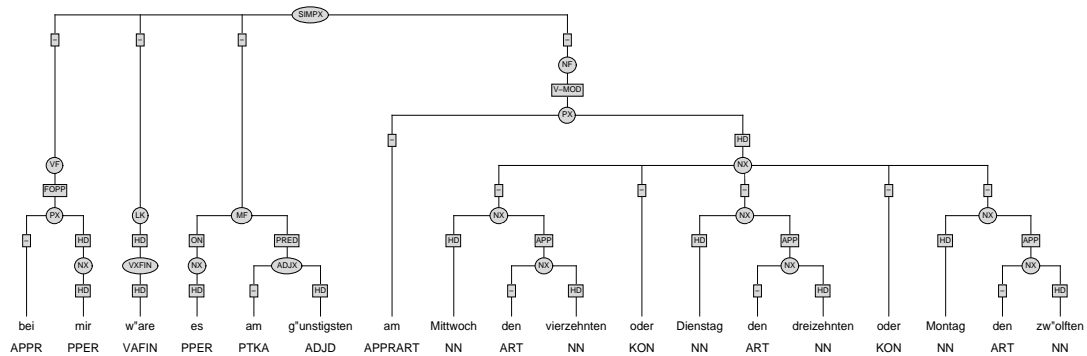


Adverbial Phrases

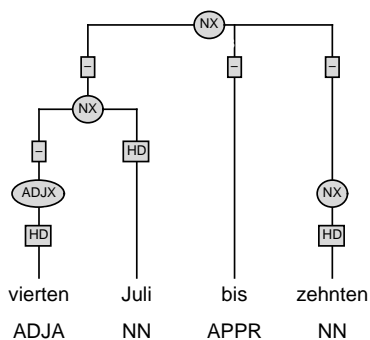
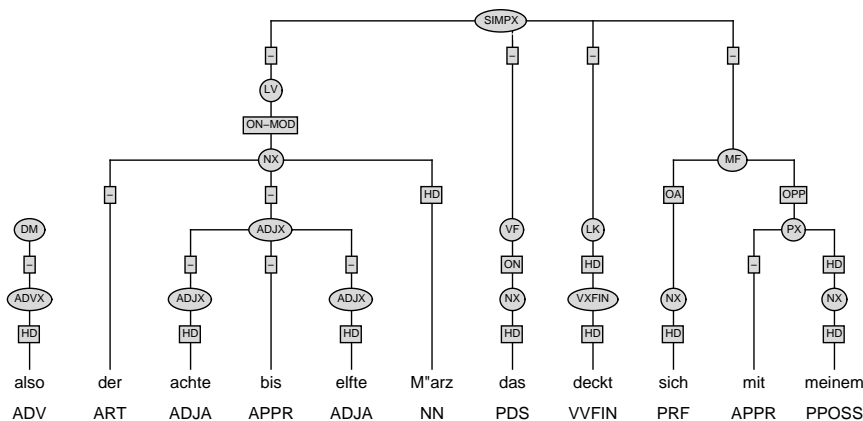


6.10.2 Specific Coordination Phenomena

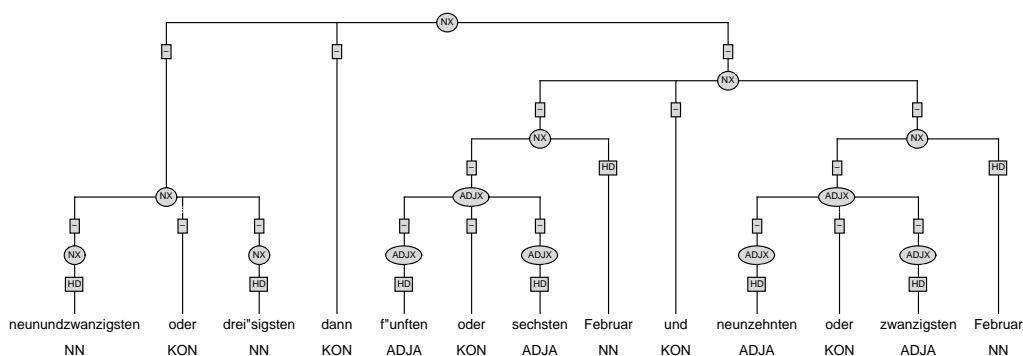
Coordinations with more than two conjuncts are treated as flat n -ary branching structures:



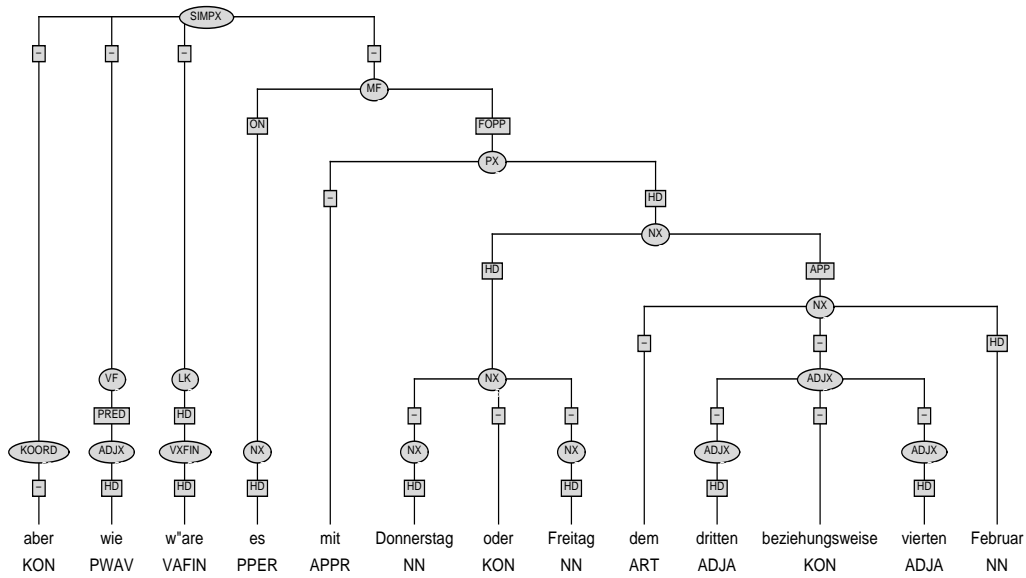
Coordinations with *bis/APPR* are very common in VERBMOBIL. They are treated as coordinations with *und/oder*:⁵



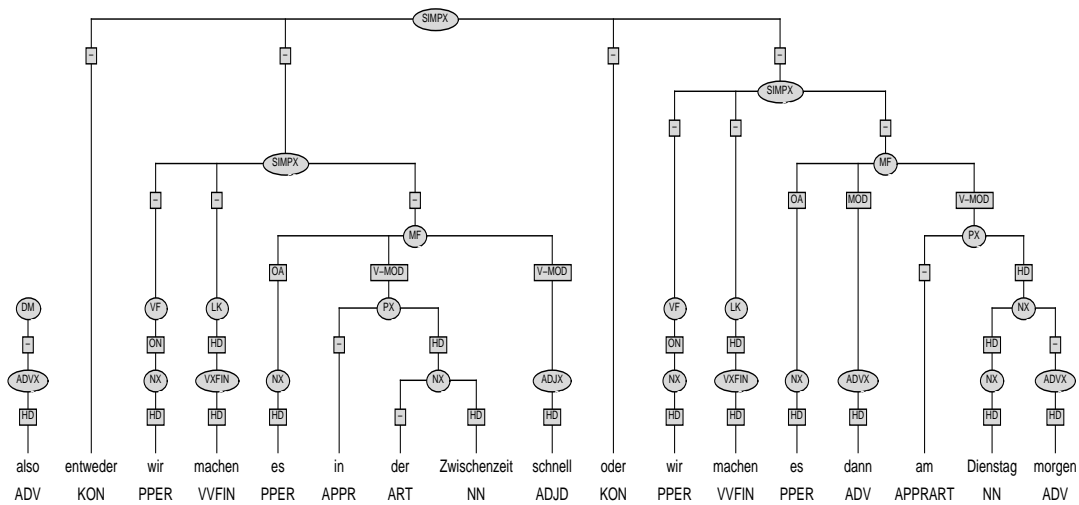
dann/KON and *beziehungsweise/KON* can be replaced by *und/oder* and therefore are treated as conjunctions:

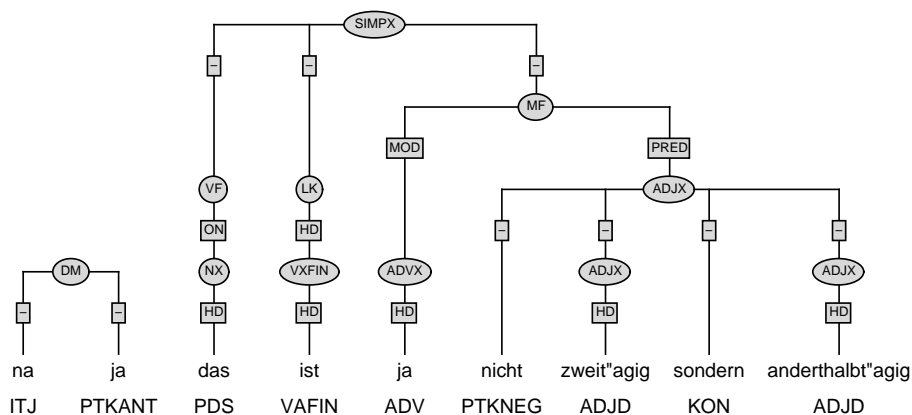


⁵But remember that *von ... bis ...* phrases are treated differently (cf. section 4.3.1)!

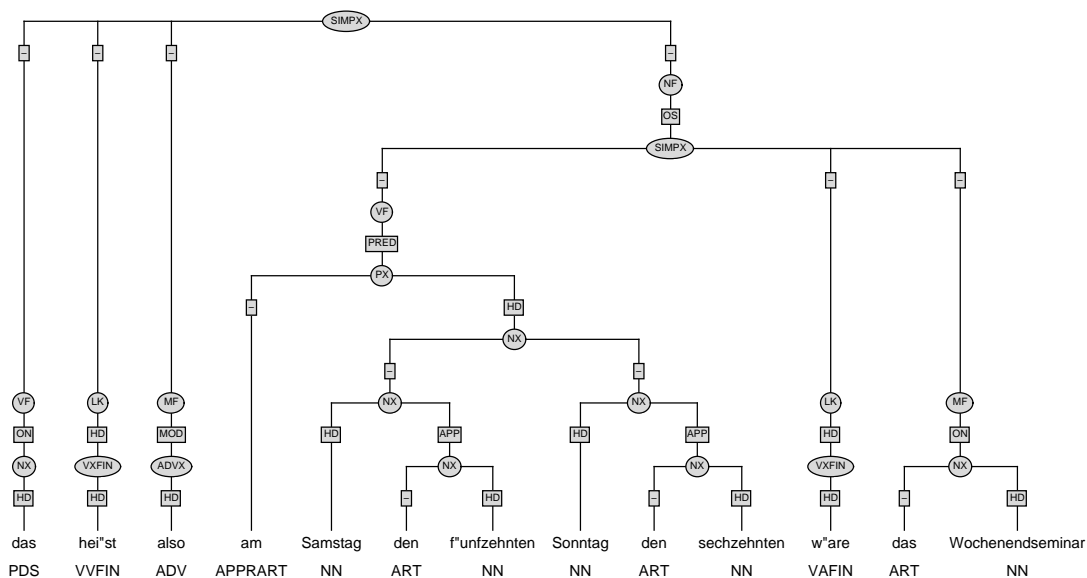


entweder - *oder* forms a unit that is semantically closely related to coordinations and thus can be treated in the same way. This seems to be the case for *nicht* - *sondern* as well. Note that *entweder*, *oder*, and *sondern* are tagged as KON, whereas *nicht* is always tagged as PTKNEG:

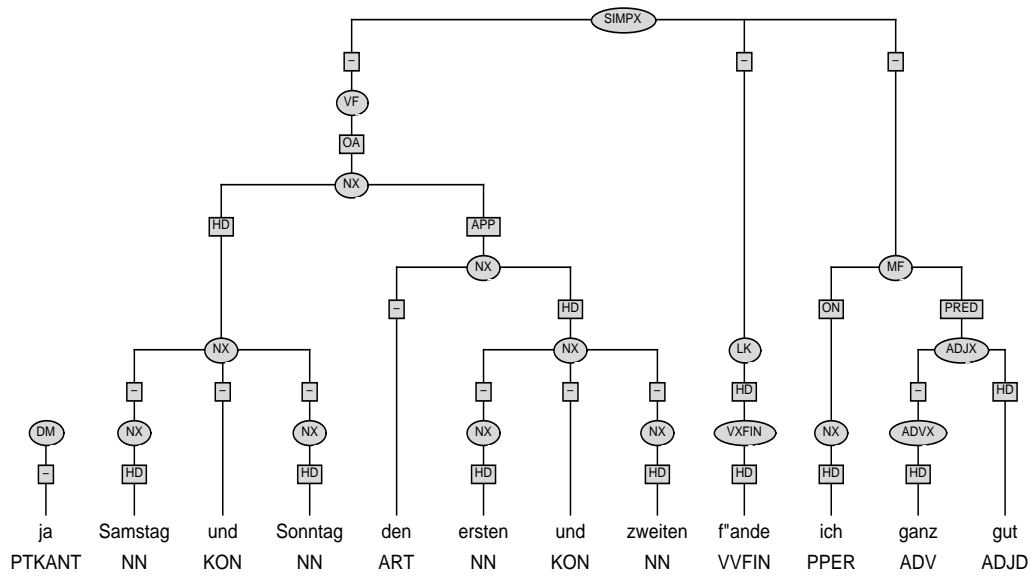
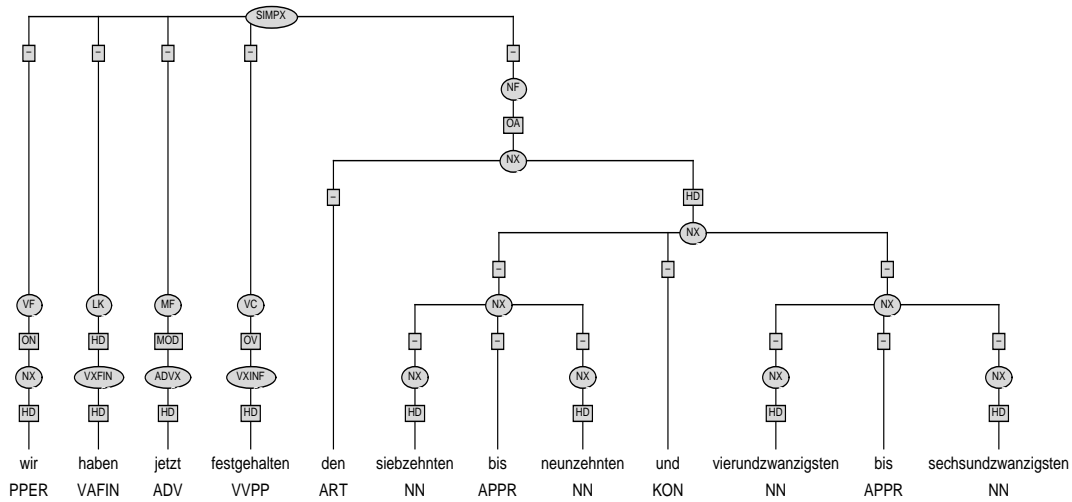




Enumerations are considered coordinations without conjunctions:



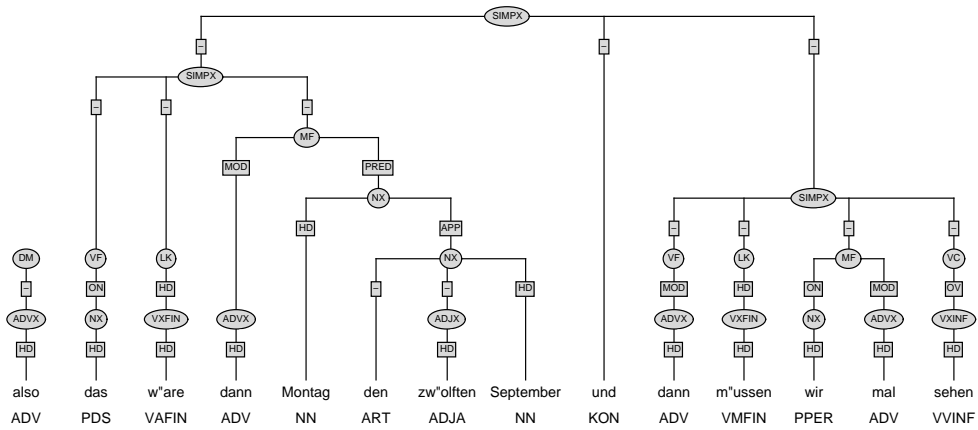
Two or more nominal conjuncts may occur together with a common determiner. This determiner is not attached to the first conjunct only, but to the node above the coordination. Thus, its scope extends over the entire coordination:



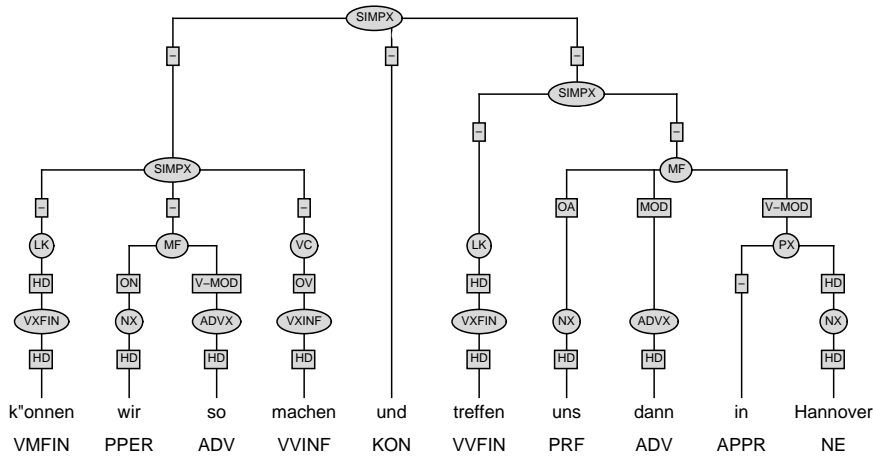
6.10.3 Coordination of Sentences

In accordance with the *longest match principle*, also complete sentence constructions can be coordinated. In the following example, two main clauses, both containing a subject, are the conjuncts of a coordination of sentences:

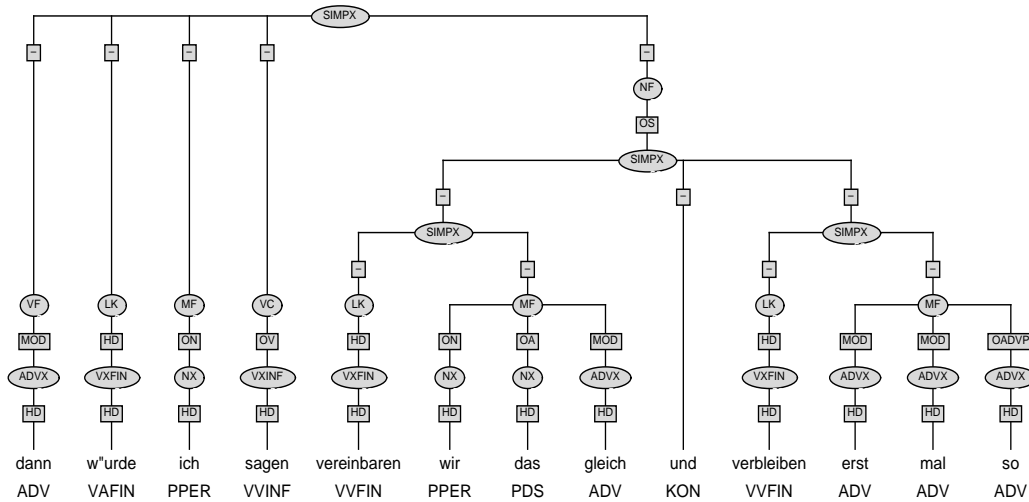
Stylebook for the German Treebank



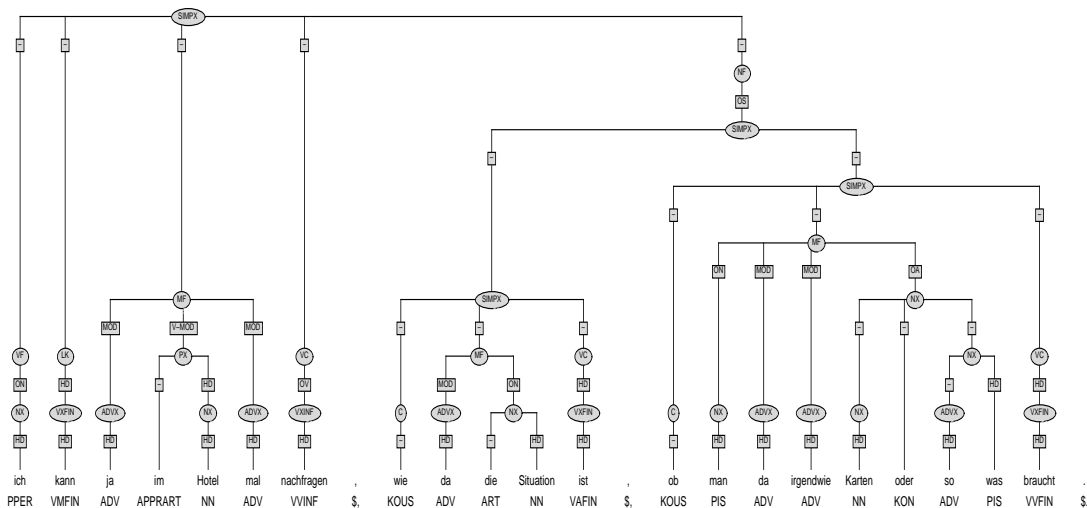
But a coordination may also consist of two main clauses with the subject of the whole construction only occurring in the left conjunct of the coordination.



Sentences within a field (either in the VF or in the NF) may also be coordinated by a conjunction:



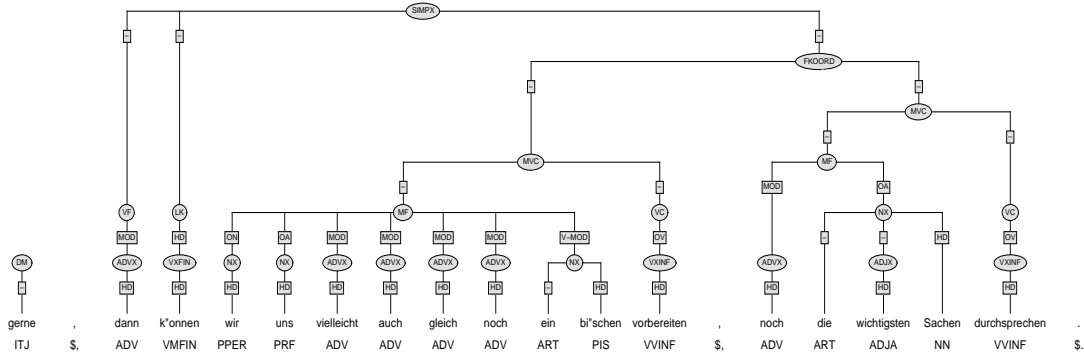
Subordinate clauses (either in the VF or in the NF) with or even without a conjunction can also be treated as conjuncts.



6.10.4 Coordination of Topological Fields

The conjuncts of a coordination of topological fields are either single fields or - like in the following example and most other cases - a combination of fields. Possible combinations are, for instance, MVC (MF + VC), LKM (LK + MF), LKMVC (LK + MF + VC). The node labels for these conjuncts are derived by the concatenation of the field labels of the conjunct. In the following example equal field combinations (MVC and MVC) are attached to the general coordination field

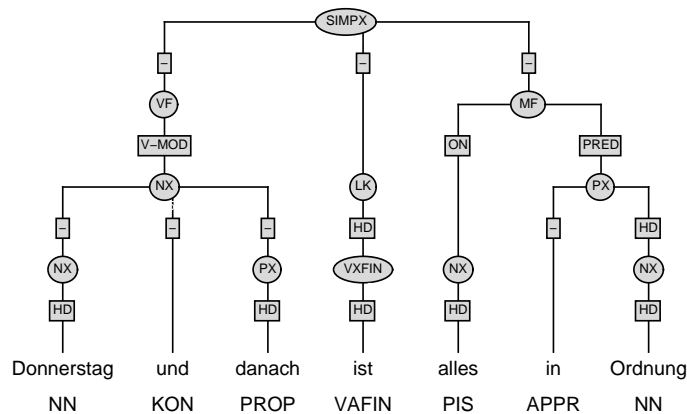
FKOORD on the next higher node.



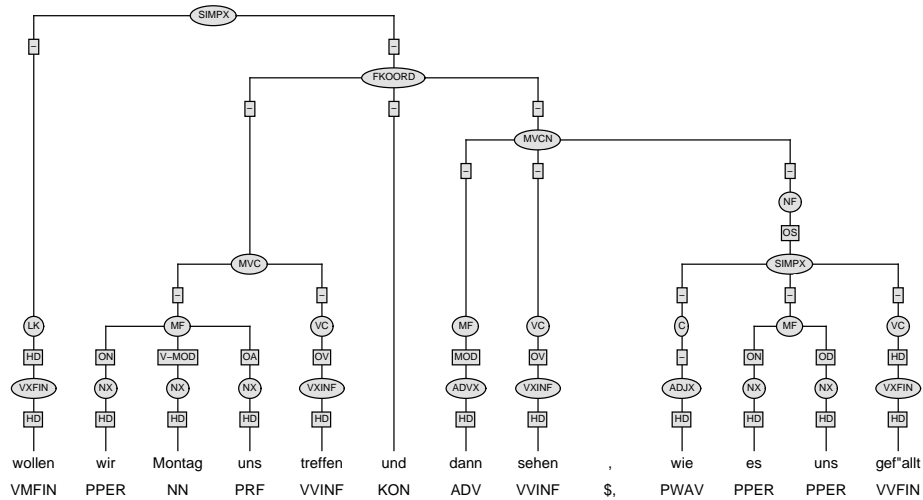
6.10.5 Coordinations with Unequal Conjuncts

The Problem of Unequal Conjuncts

Within some coordinations, the conjuncts are not of the same syntactic category. The problem is to choose the right label for the mother node. In this case, the default strategy has been adopted to choose the syntactic category of the **left-most** conjunct as the category of the entire coordination:



There are even more difficult cases, which can only be solved by the coordination of topological fields. In the following example, the conjuncts are unequal inasmuch the subject refers to both conjuncts but only occurs in the left conjunct. Moreover, the right conjunct consists of a NF in addition:

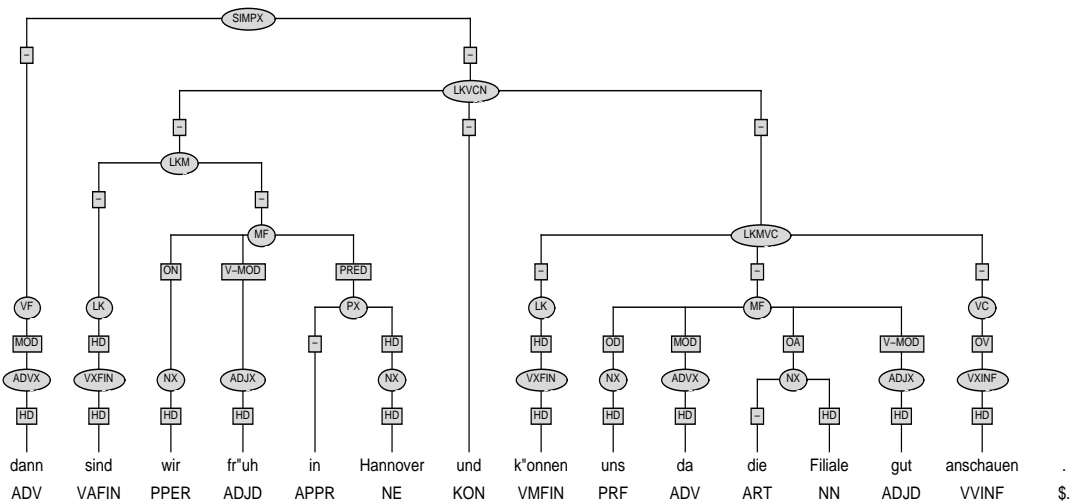
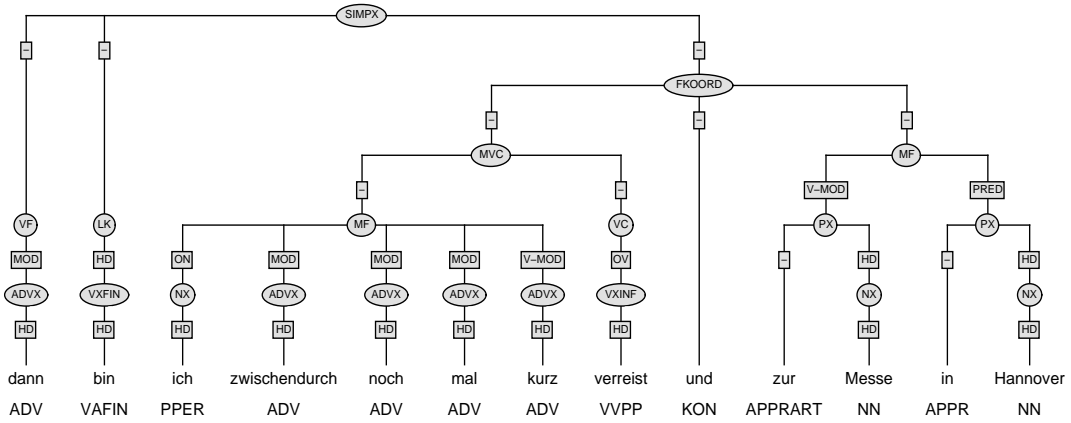
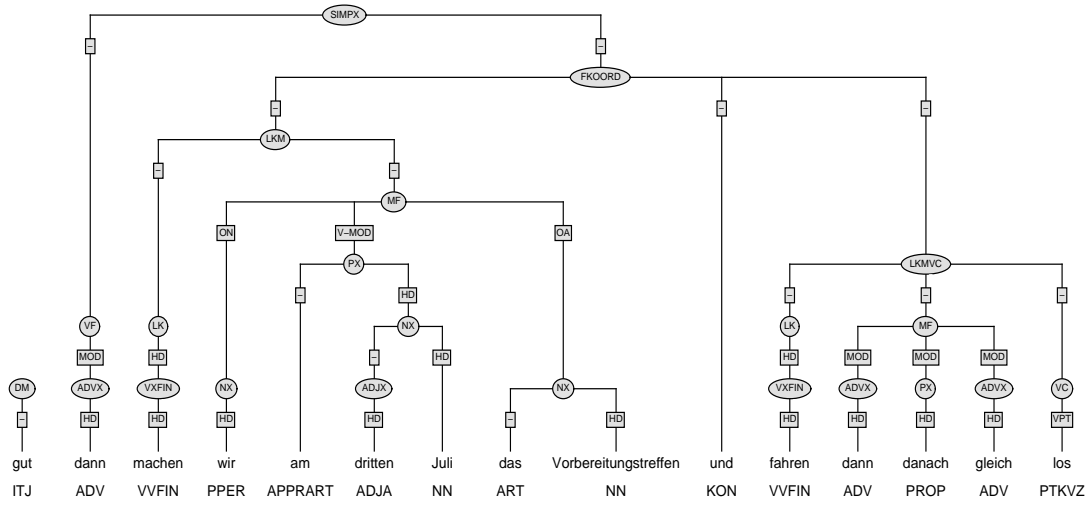


Therefore, this phenomenon is analysed as a coordination of fields rather than a coordination of constituents. In a coordination of fields, the following annotation steps are involved:

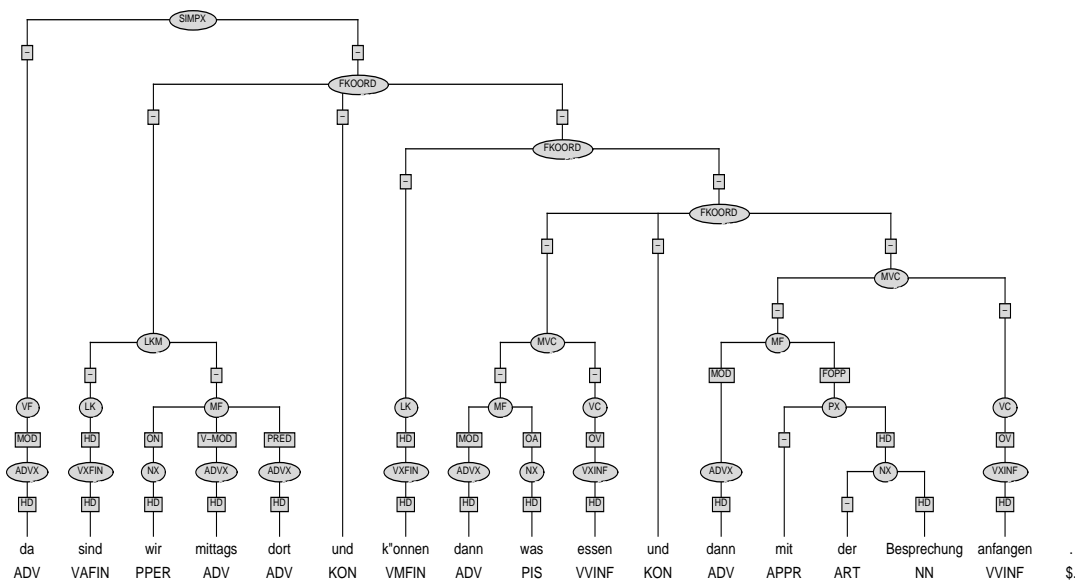
1. The constituents are attached to the fields they occur in (MF, VC, NF etc.).
2. Each conjunct (concatenation of fields) is reduced to a common denominator which subsumes the fields it consists of (e.g. MVC (M+VC), MVCN (M+VC+N)).
3. The conjuncts (in the example above: MVC and MVCN) are attached to the general coordination field FKOORD. The FKOORD-label subsumes all possible combinations auf field conjuncts.

Further examples:

Stylebook for the German Treebank

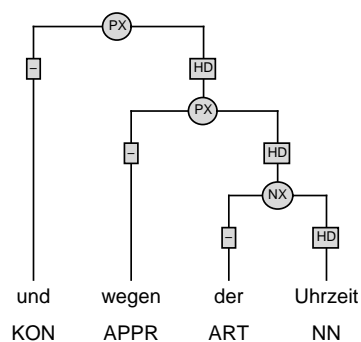


The coordination of fields may even be more complex. FKOORD can also be the daughter node of another FKOORD:

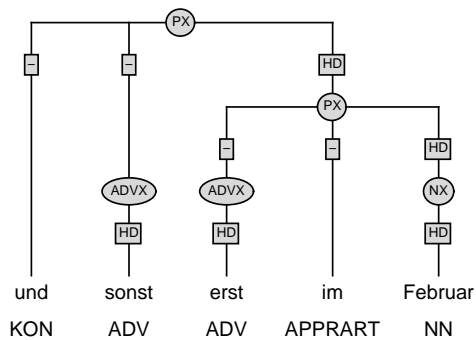
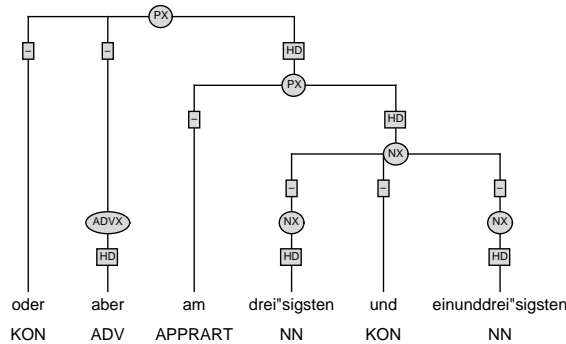
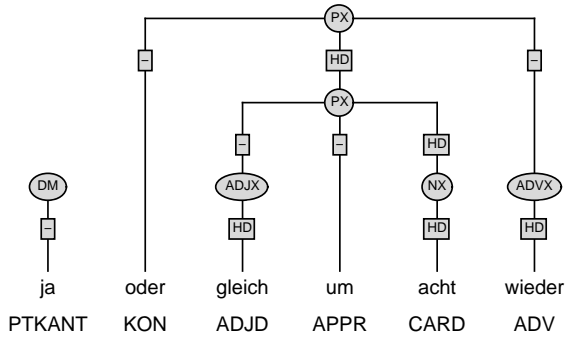


6.10.6 Conjunctions Occurring with Isolated Phrases

In VERBMOBIL, conjuncts often occur isolated, i.e. not as a coordination with at least two conjuncts. In this case, the conjunction is attached to the conjunct on a higher node like in complete coordinations. However, for isolated conjuncts, the conjunct is annotated as the head of the construction (whereas complete coordinations have no heads!). In the following example, *und* is attached high to the conjunct *wegen der Uhrzeit*:



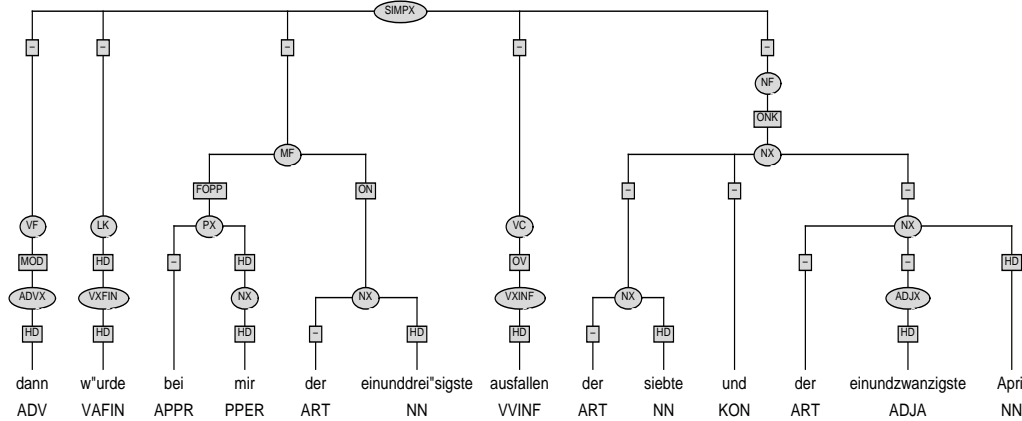
If there are modifiers that do not belong to the conjunct itself because they are ambiguous or might modify something else rather than the conjunct, they are attached on the same (high) level as the conjunction:



6.10.7 Split-up Coordinations

Closely related to isolated conjuncts are *split-up* coordinations. Generally, the left conjunct of a split-up coordination is located in the MF, in some cases in the VF, and the right conjunct occurs in the NF. To express the relation between them,

the left conjunct carries the label of its grammatical function (ON, OA, OD, etc.) whereas the right conjunct carries a label that denotes that it is the conjunct of this grammatical function (e.g. ONK, OAK, ODK, etc.):



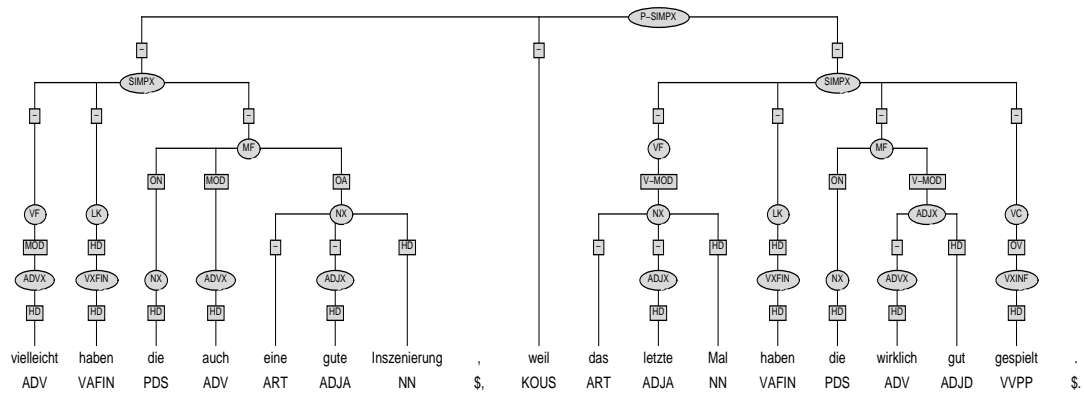
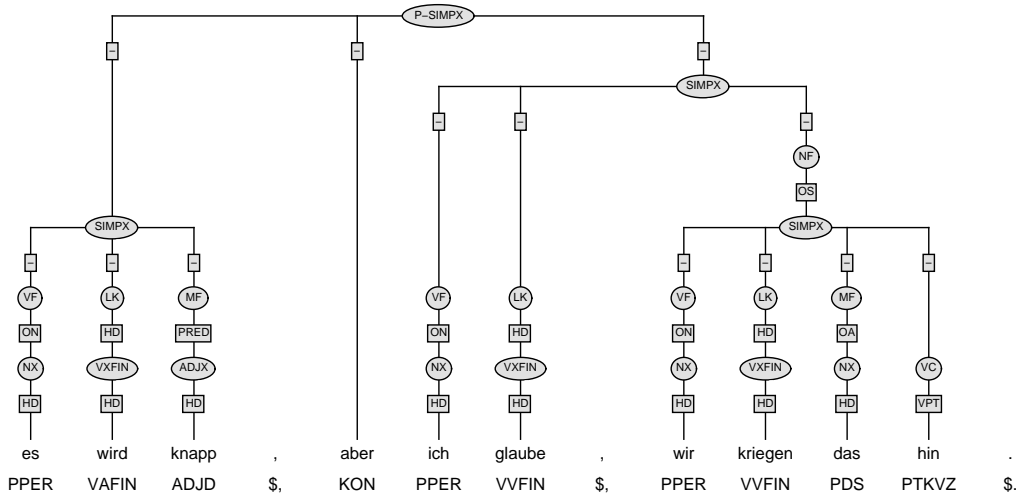
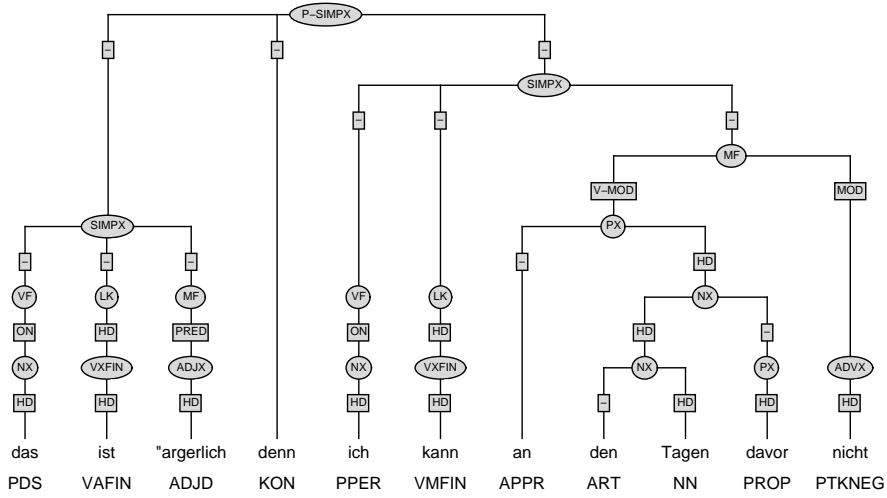
Sometimes it is difficult to distinguish between split-up coordinations and coordinations of topological fields. We use the following rules to distinguish the two cases:

- A **split-up coordination** is given if there are two phrases of the same type (e.g. two NPs) that are in a coordination relation to each other, even if they are not adjacent.
- A **coordination of topological fields** is given if the conjuncts contain different combinations of fields.

6.11 Paratactic Constructions

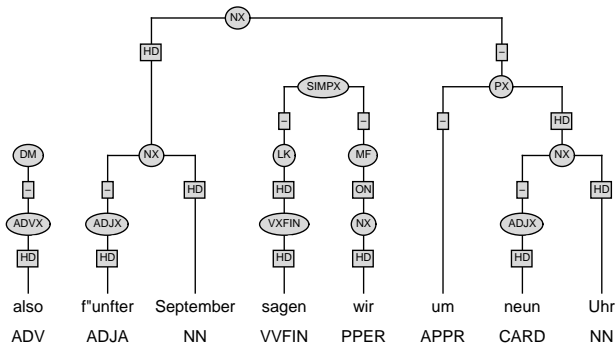
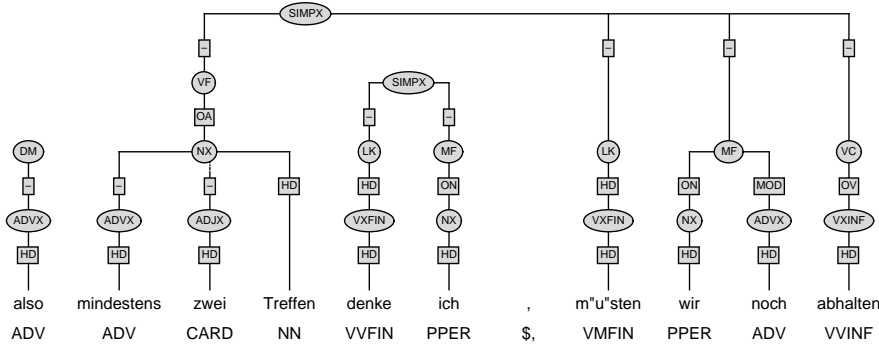
In paratactic constructions, verb-second clauses are treated as equal conjuncts. They may, for instance, be conjoined by *denn*, *aber*, *nur*, *weil*, etc. These are conjunctions that can occur in the PARORD-field in the beginning of a sentence.

Stylebook for the German Treebank



6.12 Parentheses

Parentheses occur as interjective utterances within sentences. They are not attached to the surrounding constituents. There is no dependency relation between the parenthesis and the rest of the construction. Often parentheses occur as SIMPX-clauses:



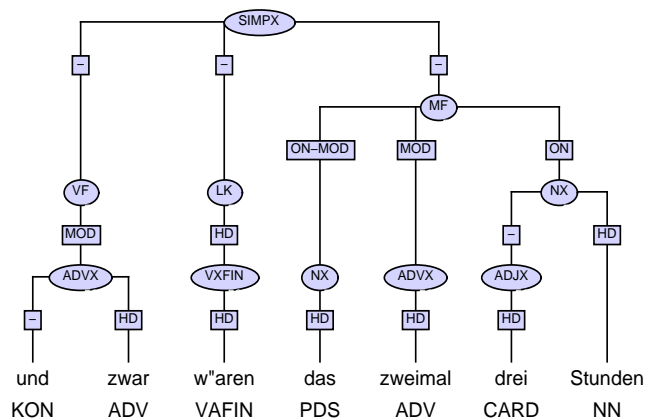
Chapter 7

The Annotation of Specific Syntactic Phenomena

7.1 *und zwar*-Constructions

Und zwar may occur in two different ways:

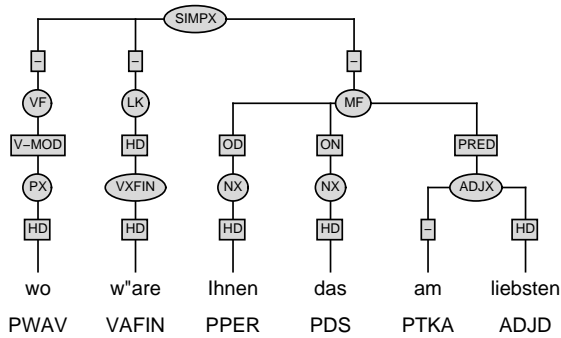
1. as a discourse marker (cf. *und zwar* in section 7.5)
2. as a modifier within the VF of a sentence (as described below)



In the second case *und* is directly attached to *zwar* in order to express that *und zwar* is a unit that typically occurs within the VF of verb-second clauses.

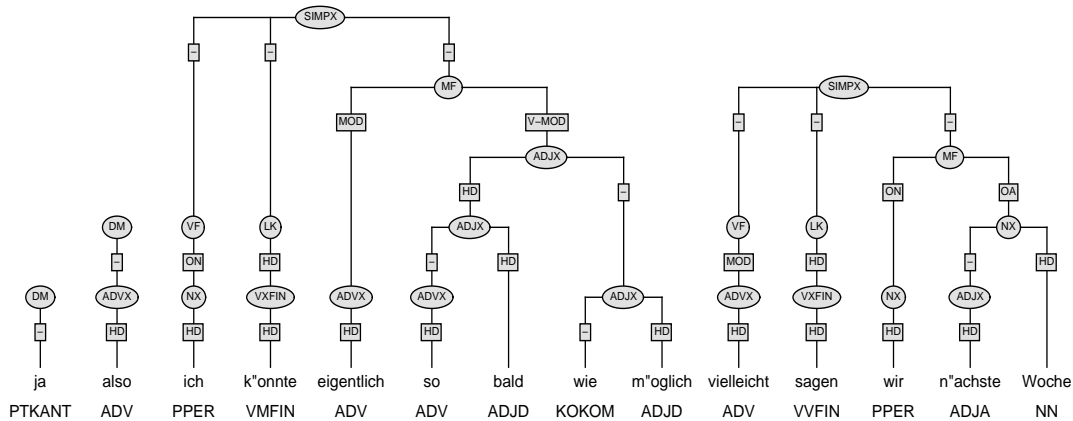
7.2 Superlative and Comparative Forms

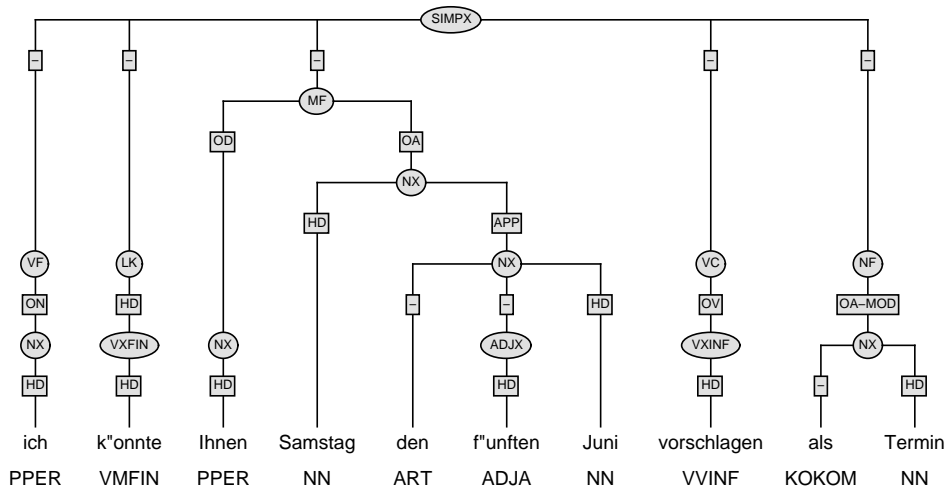
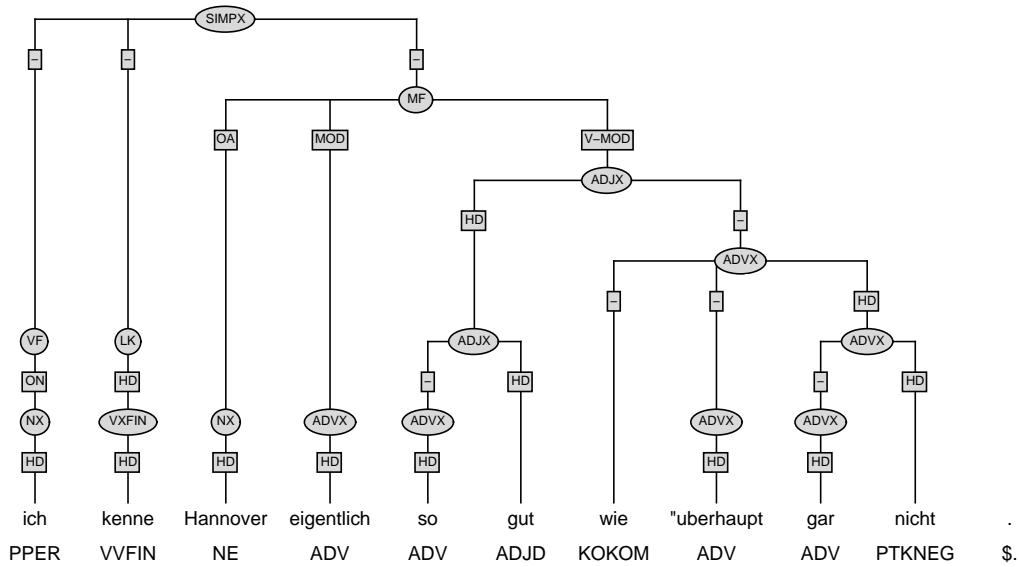
Note that the part-of-speech tag of *am*, which occurs as a particle with an adjective or an adverb in superlative constructions, is PTKA, not APPRART:



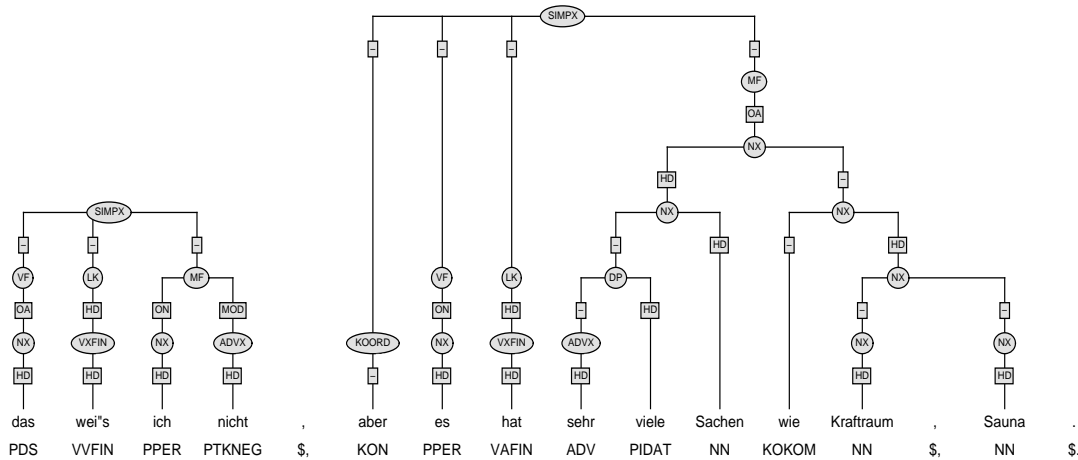
7.2.1 The Comparative Particles *wie* and *als*

In German only *als* and *wie* are comparative particles. They do not introduce a sentence and they are tagged as KOKOM. These particles can occur with all types of syntactic phrases (NX, ADVX, PX, etc.). They function as prenominal modifiers, directly attached to the phrase. This phrase postmodifies the phrase it is related to. If they are not adjacent, the long-distance dependency is denoted by the respective edge labels:





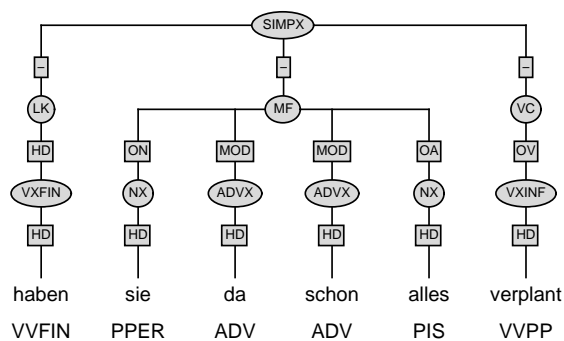
The comparative particle can also modify a coordination of phrases. In this case, first the two conjuncts are coordinated. Then the particle is attached on the high node.

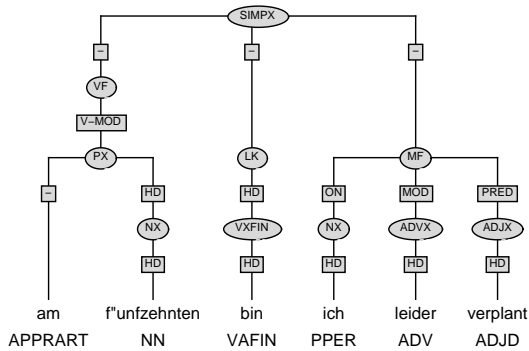


7.3 Verbal and Adjectival Use of Participles

In German all verbal participles which are passive verb forms can be used as adjectives either as an attribute adjective (*der verplante Tag*) or as a predicate/adverbial adjective (*der Tag ist verplant/er fährt schnell*). Morphologically there is no difference between converted adjectives and passive verbal participles. In constructions with adjectival passives the auxiliary *sein* is used in contrast to the *passive werden*, which is the auxiliary in verbal passives. Adjectival passive constructions are called *Zustandspassiv*. Therefore adjectival passives are tagged as ADJD and are projected to an adjectival phrase. Concerning the differences between adjectival and verbal passives in English cf. Bresnan (1995).

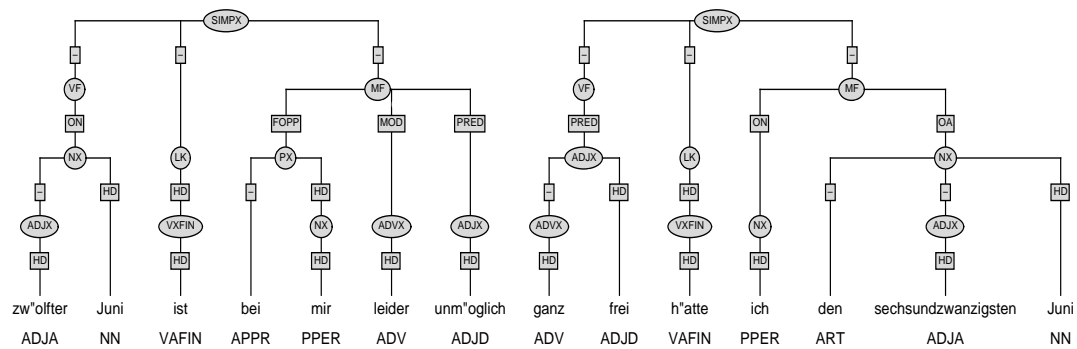
The first example shows a non-passivized verbal participle, which is tagged as VVPP, the second a *Zustandspassiv*, in which the particle is tagged as ADJD.



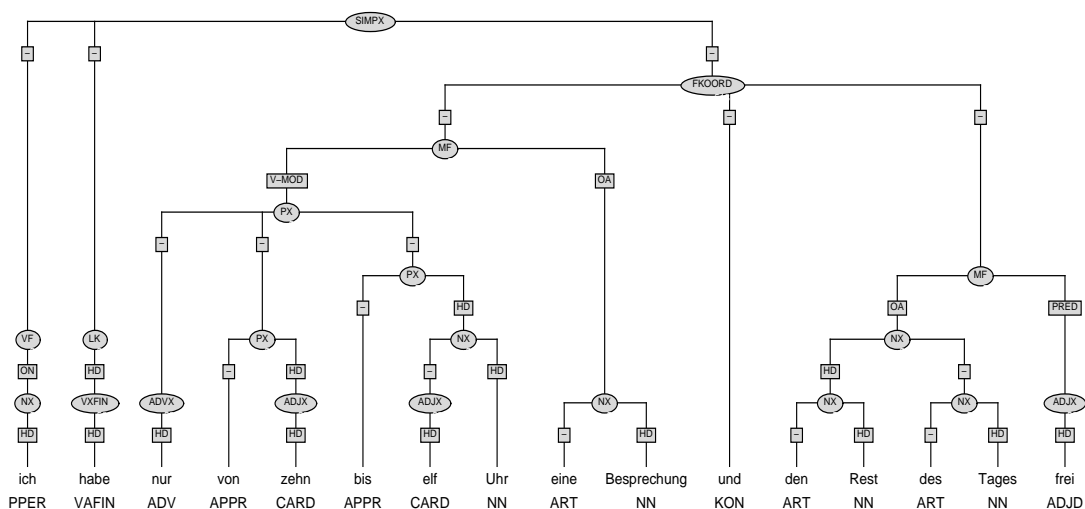


7.4 Adjectival Use of Verbal Particles

Separable verbs consist of a verb and a particle which can be a prepositional, an adjectival, or an adverbial element. In special syntactic constructions like topicalizations, the particle has phrasal status. In this case, it is used, for instance, as an adjective. Verbal particles with adjectival use are not tagged as PTKVZ but as ADJD like *frei* in the following example.



In the following coordination, the separable particle is also treated as a complement of the verb:



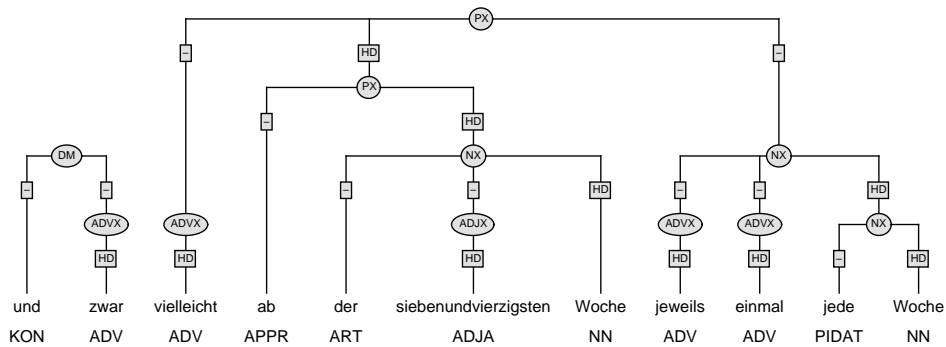
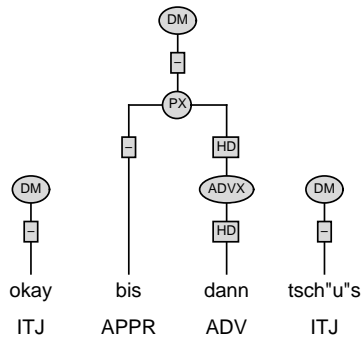
7.5 Discourse Markers

Discourse markers are sentence external expressions or phrases that contribute information about the course of conversation, e.g. whether a suggested date is being accepted or rejected. Expressions of greeting, apologizing, thanking, sentence introducing expressions, short emotional utterances, and interjections are also considered discourse markers.

Discourse markers are never attached to sentences, but they may occur as isolated (mainly interruptive) expressions within a sentence. The daughter of a discourse marker may be a phrase or a single expression. The edge label of a discourse marker is empty (i.e. it does not have a HD).

Typical discourse markers are:

ja, nein, guten Tag, hallo, auf Wiederhören, ah so, okay, bis dann, ne, genau, einverstanden, also, tut mir leid, sehr gut, alles klar, danke, und zwar ...



Single words that are directly projected to a DM are tagged as ITJ, e.g. adjectives like *einverstanden*, *okay*. Note that date suggestions such as *Montag*, *vielleicht später in der Woche wieder*, etc. are **not** defined as discourse markers because they carry semantic information that arises from their internal constituent structure. They cannot be considered to have only special discourse marking function.

Chapter 8

Problematic Issues

8.1 Problems with Grammatical Functions

8.1.1 Distinguishing FOPP, OPP, and V-MOD

One of the major problems is to distinguish, whether a given PP is an obligatory (OPP) or an optional (FOPP) complement of a specific verb in a specific reading, or whether it is a free adjunct (V-MOD) of that verb.

The *Verblast* document, which is electronically available with the treebank, lists all verbs occurring in the German treebank with their specific subcategorization frames. The list is intended as a reference for these problematic cases.

In the following, we will briefly describe what criteria have been used in order to decide about the subcategorization with respect to PP complements/modifiers:

1. A PP is called **OPP** within a sentence if the sentence were ungrammatical without the OPP (or if there was at least a very noticeable change of meaning). For instance, *Er legt den Termin [MOD auf Donnerstag]./ Etwas kommt [OPP in Frage]./ Sie rechnet [OPP mit Montag].*
2. A PP is called **FOPP** if it can be left out of this specific sentence without causing ungrammaticality (or a very noticeable change of meaning) **and if its preposition is selected by this specific verb**. For instance, *Sie wird fertig mit dem Report./ Er freut sich auf den Termin./ Sie einigen sich auf Montag*. Here, the prepositions are selected by these verbs and the PPs cannot be added to any arbitrary verb (which is possible for free adjuncts).
3. A PP is called **V-MOD** if its preposition is not selected by this specific verb, i.e. it can be exchanged by any other modifying PP, and similarly, this PP can occur with arbitrary verbs. Typical V-MODs are temporal or local

adjuncts specifying time and location of the action/event/state expressed by the verb.

8.1.2 Distinguishing MOD, MOD-MOD, and V-MOD

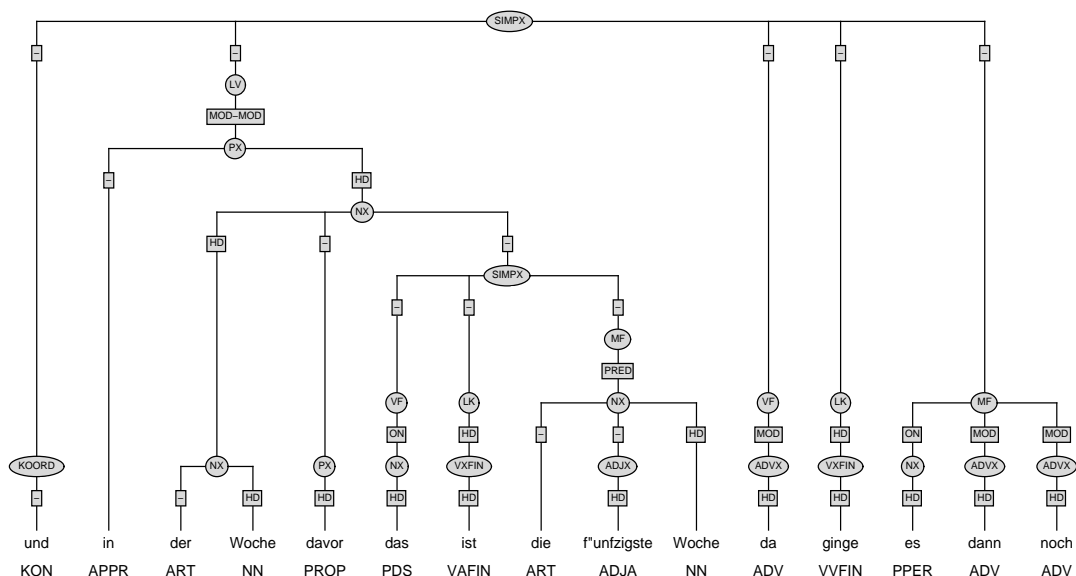
A typical VERBMOBIL case of ambiguity is a date consisting of a day and a period of time that further specifies that day. In general this means: a larger (or more vague) period/point of time specified by a more specific period/point of time. Here, the larger period/point is X-MOD (with X=V for instance), whereas the more specific period/point is MOD-MOD:

1. [V-MOD *am Freitag*] *habe ich* [MOD-MOD *den ganzen Tag*] *Zeit*.
 [V-MOD *am Freitag*] *habe ich* [MOD-MOD *um drei Uhr*] *Zeit*.
 [MOD-MOD *Wann*] *haben sie* [V-MOD *montags*] *Zeit?*
2. [MOD *da*] *habe ich* [V-MOD *morgens*] *keine Zeit*.
 [MOD *danach*] *habe ich* [MOD-MOD *in der zweiten Woche*] *Zeit*.
 [V-MOD *wann*] *haben Sie* [MOD *da/dort*] *Zeit?*
da is MOD because it might be time **or** location. *danach* is MOD, because it also refers to some sentence external expression that has been mentioned before.

The following two cases should be carefully distinguished:

1. [V-MOD *Dienstag*] *kann ich* [MOD-MOD *ab 13 Uhr*].
 [MOD *danach*] *kann ich* [MOD-MOD *ab 13 Uhr*].
Dienstag, *danach*, etc. can only be interpreted as **temporal** expressions, thus the following *time* expression *ab 13 Uhr* can only refer to them as MOD-MOD.
2. [MOD *da*] *kann ich* [V-MOD *ab 13 Uhr*].
 [MOD *dann*] *kann ich* [V-MOD *ab 13 Uhr*].
da, *dann*, etc. can be either **temporal**, **causal**, **consequential**, or **local** expressions. So one cannot make sure whether the following *time* expression *ab 13 Uhr* really refers to them. All that is obvious is that the *time* expression is a V-MOD in any case.

For resumptive constructions (LV), there is also a clear criterion concerning the modification relations. Within a verb-second clause, a modifier occurring in the VF is MOD/X-MOD, whereas an additionally occurring modifier in LV is MOD-MOD, not vice versa, because the modifier in VF occurs within the “core” of the sentence, whereas the modifier in LV has to be licensed by some other constituent in the core sentence:



8.1.3 Problems with the Distinction of ON, PRED, and ON-MOD

It is not always trivial to distinguish which constituent is ON, PRED, or ON-MOD for predicative verbs. For this reason, a few criteria and examples are listed here that can be of help. Here are some properties of PRED, ON, ON-MOD:

1. Typically, PRED occurs in the MF, whereas ON occurs in the VF of verb-second clauses. So this should be considered for annotation, if no other criterion (as described below) applies.
2. Of course, subject-verb agreement always has to be taken into account. For instance, if the verb is in plural form, the subject has to be plural as well. If the constituent in the VF is not the subject (e.g. because of agreement mismatches), then it is considered ON-MOD rather than PRED. However, there are cases in which typical PREDs occur in the VF: [*PRED Gut*] wäre [*ON Montag*].
3. If there is a suitable NP that could serve as a subject, then this NP is annotated as the subject rather than any other constituent with a different syntactic category (PP, ADVP etc.).

For verb-second clauses, it is important to follow these two steps in exactly this order to stick to the distributional criterion that has been chosen for the

PRED/ON/ON-MOD distinction:

1. Have a look at the constituent in the VF. If it is an NP which might serve as subject **and** if it agrees with the verb, the annotate it as the subject (ON).
2. If it does **not** agree with the verb, annotate it as ON-MOD, but **only if** it is not a typical PRED (ADJP, ADVP, PP, etc.).¹

Examples:

1. *[ON vier Termine] müssten [PRED es] sein.
und zwar wären [PRED das] zweimal [ON drei Stunden].
diesmal sind [PRED es] [ON drei].*
Subject-verb agreement suggests that *vier Termine/drei Stunden/drei* is the subject, because it is in plural form.
2. *[PRED gut] wäre [ON Montag].
[PRED wie] ist [ON Ihr Urlaub]
[PRED hier] ist [ON Frau Müller].
[PRED da] wäre dann ja [ON Zeit].
dann ist [PRED dort] [ON der zweite Termin].*
ADJPs and ADVPs typically have PRED function when occurring together with predicative verbs and NP subjects.
3. *[PRED am ersten] ist [ON Allerheiligen].*
Allerheiligen is considered the subject, because it is an NP, whereas *am ersten* is a temporal expression in form of a PP. *Am ersten* is not ON-MOD, since it is a PP and a typical PRED.
4. *[ON das] wäre [PRED ein guter Termin].
[ON ein guter Termin] wäre [PRED das].
[ON das] wäre [PRED es].
[ON der sechzehnte] ist [PRED ein Freitag].*
The NP in VF position agrees with the verb and therefore has subject priority. As a consequence, the constituent in the MF is PRED.
5. *[ON-MOD das] wären also [ON drei Termine].*
The subject (agreement!) occurs in MF position, thus the constituent in VF position is ON-MOD. It is an NP and does not predicate anything, thus it is no typical PRED.²

¹This scheme explains why *es/das* in the VF is often ON-MOD, whereas *es/das* in the MF is mainly PRED.

²However, NPs can be PREDS in other cases.

6. Especially note the distinction between (a) and (b):

- (a) [*ON Mai/Anfang Dezember*] *ist* [*PRED ziemlich voll*].
 [*ON Nachmittag*] *wäre mir* [*PRED lieber*].
- (b) [*V-MOD heute*] *ist* [*PRED problematisch*].
 [*V-MOD am 15.*] *ist* [*PRED gut*].

In (a), temporal expressions that are NPs might serve as subjects. In (b), the temporal expressions are not NPs and therefore have to be considered verbal modifiers rather than subjects.

Test: Something is a subject, if the insertion of *es* violates the well-formedness of the sentence (**Mai ist es voll.*) and something is a V-MOD if *es* can be inserted without problems (*heute ist es problematisch.*).

8.1.4 Problems with the Distinction of ON-MOD, ON, and V-MOD within LV Constructions

In the following sentence, it is not immediately obvious whether *Sonntag und Montag* should be annotated ON-MOD or V-MOD:

Sonntag und Montag ist das auch in Ordnung.

In cases of uncertainty, the following **LV test** can be applied:

The idea is to put a potential ON-MOD in LV position and the subject (ON) in VF position. This way, it is easier to see whether the subject refers back to the constituent in LV position or not. Examples:

1. Original sentence: *Sonntag und Montag ist das auch in Ordnung.*
 LV test: [*LV Sonntag und Montag*] [*VF das*] *ist auch in Ordnung.*
 In this constituent order, it is obvious, that *das* (ON) refers back to *Sonntag und Montag*. Therefore, *Sonntag und Montag* is labelled as ON-MOD.
2. Original sentence: *Mittwoch der 25. wie sieht es da aus?*
 LV test: * [*LV Mittwoch der 25.*] [*VF es*] *sieht da schlecht aus.*
es does not refer back to *Mittwoch der 25.*, thus *Mittwoch der 25.* is labelled as MOD-MOD rather than ON-MOD.
3. Original sentence: *Donnerstag ist alles in Ordnung.*
 LV test: * [*LV Donnerstag*] [*VF alles*] *ist in Ordnung.*
alles does not refer to *Donnerstag*. *Donnerstag* is labelled as V-MOD rather than ON-MOD.

More examples:

1. *[ON-MOD Freitag den 28.] das wäre schön.*
2. *[V-MOD Donnerstag] wäre es schön.*

Of course, the same principle applies for constituents other than ON, for instance OPP:

[OPP-MOD Samstag der 27. August] [OPP damit] wäre es gut.

References

- Behaghel, O. 1932. *Deutsche Syntax (Eine geschichtliche Darstellung), Band 4*. Heidelberg: Carl Winter.
- Brants, T., and W. Skut. 1998. Automation of treebank annotation. In *Proceedings of the Conference on New Methods in Language Processing (NeMLaP-3/CoNLL98), January 14-17, 1998, pages 49-57, Sydney, Australia*, 49–57.
- Bresnan, J. 1995. Lexicality and Argument Structure. In *Invited Paper given at the Paris Syntax and Semantics Conference*. October 12-14, 1995. URL: <http://www-csli.stanford.edu/~bresnan/download.html>.
- Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy*.
- Drach, E. 1937. *Grundgedanken der Deutschen Satzlehre*. Frankfurt/M.
- Erdmann, O. 1886. *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt*. Stuttgart. Erste Abteilung.
- Grewendorf, G. 1991. *Aspekte der deutschen Syntax, Band 33 of Studien zur deutschen Grammatik*. Tübingen: Gunter Narr Verlag.
- Herling, S. H. A. 1821. Über die Topik der deutschen Sprache. In *Abhandlungen des frankfurterischen Gelehrtenvereins für deutsche Sprache*, 296–362, 394. Frankfurt/M. Drittes Stück.
- Hinrichs, E., W. D. Meurers, F. Richter, M. Sailer, and H. Winhart. 1997. *Sprachtheoretische Grundlagen für die Computerlinguistik. Ein HPSG-Fragment des Deutschen. Teil 1: Theorie*. SFB 340, Universität Tübingen.
- Höhle, T. N. 1985. Der Begriff ‘Mittelfeld’. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne (Ed.), *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen*, 329–340.
- Kathol, A. 1995. *Linearization-Based German Syntax*. PhD thesis, Ohio State University.
- Kawata, Y., and J. Bartels. 2000. Stylebook for the Japanese Treebank in VERBMOBIL. Technical report, Verbmobil-Report 240.
- Kordoni, V. 2000. Stylebook for the English Treebank in VERBMOBIL. Technical report, Verbmobil-Report 241.

- Plaehn, O. 1998. Annotate - Bedienungsanleitung, Universität des Saarlandes, FR 8.7 Computerlinguistik, Projekt C3 Nebenläufige Grammatische Verarbeitung, Sonderforschungsbereich 378, Ressourcenadaptive Kognitive Prozesse, 13. April 1998.
- Schiller, A., S. Teufel, and C. Thielen. 1995. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universitäten Stuttgart und Tübingen. URL: <http://www.sfs.nphil.uni-tuebingen.de/ELWIS/stts/stts.html>.
- Skut, W., B. Krenn, T. Brants, and H. Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP), Washington, D.C.*