Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler

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

Dezember 2003

Abstract

This stylebook describes the design principles and the annotation scheme for the German treebank TüBa-D/Z developed by the Division of Computational Linguistics (Lehrstuhl Prof. Hinrichs) at the Department of Linguistics (*Seminar für Sprachwissenschaft* – SfS) of the Eberhard-Karls-Universität Tübingen, Germany. The guidelines focus on the syntactic annotation of written language data taken from the German newspaper 'die tageszeitung' (taz). The unannotated taz newspaper material was taken from the Science CD (*Wissenschafts-CD*) of 'die tageszeitung' (taz) that can be licensed from contrapress media gmbh (email: abu@taz.de).

At present, the treebank comprises 15 260 sentences. The newspaper material is taken from the taz editions from May 3rd to May 7th 1999. The average sentence length is 17.5 words and the total number of tokens currently amounts to 266 441. The TüBa-D/Z treebank is still under development. Thus, the number of annotated sentences will increase over time. Periodic data updates and accompanying updates of this stylebook will be made available at:

`www.sfs.uni-tuebingen.de/en_tuebadz.shtml`

Please consult this website in order to ensure that you are using the most recent and most complete version of the treebank.

The annotation scheme for the TüBa-D/Z treebank is derived from the VERBMOBIL treebank for spoken German, developed earlier (1997–2000) by the Division of Computational Linguistics of the SfS (Hinrichs et al. 2000). The TüBa-D/Z annotation scheme has been extended along various dimensions to accommodate the characteristics of written texts. In order to ensure the re-usability of the data, a surface-oriented annotation scheme has been adopted that is inspired by the notion of topological fields and is enriched by a level of predicate-argument structure. The linguistic inventory used in the treebank annotation is based on a minimal set of assumptions that are uncontroversial among major syntactic theories. In this sense it is an attempt at theory-neutrality.

Acknowledgements

The annotation of the TüBa-D/Z treebank is carried out as part of the *Competence Center for Text- and Information Technology* (or officially: Kompetenzzentrum für Text- und Informationstechnologie – KIT"). KIT is a joint project of the Institute for Natural Language Processing (IMS) in Stuttgart, and the Seminar für Sprachwissenschaft (SfS) in Tübingen. The project is funded from 2000 through 2005 by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg and by the universities of Stuttgart and Tübingen. Additional support for the creation of the TüBa-D/Z treebank was provided by the special research center *Linguistic Data Structures* (Sonderforschungsbereich *Linguistische Datenstrukturen* – SFB 441) at the University of Tübingen funded by the German Research Council (Deutsche Forschungsgemeinschaft – DFG).

A project of this scale would not be possible without the generous support from many contributors:

Our special thanks go to 'die tageszeitung' (taz) who kindly granted permission to process the newspaper data and to release the treebank.

We would like to acknowledge Rosmary Stegmann for her many contributions to the treebank of spoken German in VERBMOBIL. Her research laid the foundations for the annotation scheme of that treebank, which has been summarized in the 'Stylebook for the German Treebank in VERBMOBIL' (Stegmann et al. 2000).

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 spoken as well as for written German.

Furthermore, we are indebted to our Tübingen colleague Tylman Ule for his assistance with part-of-speech tagging of the data and with data conversion.

We would also like to acknowledge the support of Martina Liepert and Jorn Veenstra, who initiated and developed the integration of named entities into the annotation scheme.

The quality of the treebank has been considerably improved by feature oriented consistency checks developed by Ventsislav Zhechev. Further consistency tests were contributed by Tylman Ule and Frank H. Müller in the course of their research work in the SFB 441. They deserve special mention for their support.

For their diligence and dedication to the arduous task of linguistic annotation and of post-editing we thank our research assistants Armin Buch, Silke Dutz, Steffen Frömel, Daniel Hüttl, Wolfgang Meier, Nicole Maruschka, Kai Metzger, Insa Starr.

The development of the TüBa-D/Z treebank was greatly facilitated by a number of former 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) 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.

Contents

Chapter 1

Introduction

The purpose of this report is to describe the design principles and annotation scheme for the TüBa-D/Z treebank of German. It is intended as a guide for the treebank annotators in Tübingen and for theoretical and computational linguists who want to use annotated treebank data for their own research. In addition, we hope that this report may be of some use for researchers who want to construct their own treebank for German or for some other language. We would like to emphasize that the annotation scheme is language-specific, and we advise against adopting this scheme without modification for some other language. However, we do believe that the type of design decisions that are reported here for German will arise for other languages as well. And it is in this sense that the current report could provide an useful point of reference.

The TüBa-D/Z treebank was developed by the Division of Computational Linguistics (Lehrstuhl Prof. Hinrichs) at the Department of Linguistics (*Seminar für Sprachwissenschaft* – SfS) of the Eberhard-Karls-Universität Tübingen, Germany. The guidelines focus on the syntactic annotation of written language data taken from the German newspaper 'die tageszeitung' (taz). The unannotated taz newspaper material was taken from the Science CD (*Wissenschafts-CD*) of 'die tageszeitung' (taz) that can be licensed from contrapress media gmbh (email: abu@taz.de).

At present, the treebank comprises 15 260 sentences. The newspaper material is taken from the taz editions from May 3rd to May 7th 1999. The average sentence length is 17.5 words. The total number of tokens currently amounts to 266 441. The TüBa-D/Z treebank is still under development. Thus, the number of annotated sentences will increase over time. Periodic data updates and accompanying updates of this stylebook will be made available at:
`www.sfs.uni-tuebingen.de/en\_tuebadz.shtml`

Please consult this website in order to ensure that you are using the most recent and most complete version of the treebank.

The annotation scheme for the TüBa-D/Z treebank is derived from the VERBMOBIL treebank for spoken German, developed earlier (1997–2000) by the Division of Computational Linguistics of the SfS (Hinrichs et al. 2000). The annotation scheme for the VERBMOBIL treebank has been summarized in the 'Stylebook for the German Treebank in VERBMOBIL' (Stegmann et al. 2000). The TüBa-D/Z annotation scheme has been extended along various dimensions to accommodate the characteristics of written texts. In order to ensure the reusability of the data, the linguistic inventory used in the treebank annotation is based on a minimal set of assumptions that are uncontroversial among major syntactic theories. In this sense it is an attempt at theory-neutrality.

The TüBa-D/Z treebank is released in three different data formats : the Negra export format, the Penn treebank format, and an XML format. More information about each data format is given in Appendix 8.1.3.

To the best of our knowledge, the VERBMOBIL treebank for spoken German is still the only treebank based on German speech data. For written texts, TüBa-D/Z is not the only treebank available for German. Two other treebanks are currently available, each with their own annotation scheme: the NEGRA treebank (`www.coli.uni-sb.de/sfb378/negra-corpus/`) and the TIGER treebank (`www.ims.uni-stuttgart.de/projekte/TIGER/`).

Chapter 2

Major Challenges and Design Decisions

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.

However, newspaper texts exhibit a number of phenomena that do no lend themselves easily to a purely sentence-based annotation. These phenomena include: headlines, titles, parentheses, discourse markers, and sentence conjunction by a colon. These cases are described in more detail in sections 3.4.3 to 3.4.5 of this stylebook.

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 two main considerations:

1. Linguistic adequacy and 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 common assumptions that syntacticians share across different frameworks concerning questions of constituenthood, phrase attachment, and grammatical functions. In this sense, the annotations should be theory-neutral and minimal. This desideratum is of utmost importance so as to ensure the reusability of the annotated data.

At the same time, the annotation scheme should reflect, as much as possible, those empirical generalizations that syntacticians, especially from a descriptive perspective, have identified as characteristic of the language in question.

2. Balancing the needs of potential users: Since the construction of a treebank is a labor-intensive and costly enterprise, ideally the TüBa-D/Z treebank should appeal to as many potential users with as a wide range of research interests as possible. Considering the renewed interest in the use of corpora for both theoretical and computational linguistics, choicepoints in the annotation scheme should be resolved in such a way that the needs of potential users are balanced as much as possible.

To support the use of the TüBa-D/Z treebank in computational linguistics, 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.

In order to satisfy the above desiderata, the annotation scheme is surface-oriented and context-free. The theoretical assumptions underlying the levels of annotation and the choice of labels themselves are based, as much as possible, on a rich tradition of theoretical and empirical research on German syntax.

For the treatment of word regularities of German, a language with relatively free word order, an inventory of topological fields is incorporated into the annotation scheme. Topological fields in the sense of Herling (1821), Erdmann (1886), Drach (1937), and Höhle (1986) are widely used in descriptive studies of German syntax. Such fields constitute an intermediate layer of analysis above the level of individual phrases and below the clause level. The concept of topological fields favors tree-based annotations, i.e. bracketings that do not rely on crossing or discontinuous dependencies. Instead, such non-linear dependencies are to be expressed at the level of predicate-argument structure which constitutes a second level of annotation with its own descriptive inventory of grammatical functions.

Since 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. One of the purpose of this stylebook is to add to these reference materials.

Currently, a total of 25 syntactic node labels for the encoding of constituent structures are being used. These include labels for topological fields as well as labels for phrases and their constituent parts.

In order to capture grammatical functions of individual phrases and syntactic dependencies between phrases, constituent structure trees are enriched by a set of edge labels between constituent structure nodes. The current inventory of edge labels comprises 36 distinct categories. In addition to these primary edge labels, 4 secondary edge labels are used. These labels indicate either phrase-internal government of elements in the verb complex , express phrase-internal modification of noun phrases, resolve long-distance dependencies among modifiers, or relate the phrasal complements of so-called *third-construction* control verbs.

For certain computational applications, robust identification of *named entities*, e.g. person names, names of companies and institutions, names of geographical locations, is of central importance. Therefore, such named entities are identified by a special node label, and their internal structure is sometimes identified by an additional secondary edge label that is used exclusively for named entities.

At the word level, part-of-speech labels are assigned according the Stuttgart-Tübingen tagset, which is widely accepted for part-of-speech tagging for German and which provides an inventory of 54 distinct part-of-speech labels.

Detailed information about the complete inventory of node and edge labels is given in section 3.4.2 of this stylebook.

The remainder of this stylebook is organized as follows: chapter 3 offers an overview of the theoretical foundations of the annotation scheme, focusing on the concept of topological fields (3.1) and its relation to constituent structure (3.2), on general annotation principles (3.3), as well as an overview of the annotation layers and of the inventory of the annotation labels for each layer (3.4). Chapter 4 concerns the annotation of the internal structure of phrases, broken down into major word classes and their phrasal projections. Chapter 5 addresses the principles for relating individual phrases to each other, particularly for modifier and complement attachment. Chapter 6 discusses the annotation of entire sentences, focusing on the relationship between sentence types and topological fields, coordination (including phrasal conjunction) and elliptical constructions. Chapter 7 is devoted to the annotation of miscellaneous syntactic constructions such as comparatives, verbal and adjectival participles, topicalization, newspaper headlines, discourse markers, and parentheses, which each pose special challenges for the annotation tasks. Chapter 8 describes the criteria used for distinguishing different grammatical functions. The stylebook concludes with a bibliography, a subject index, and an appendix which describes the three different data formats in which the TüBa-D/Z treebank is distributed.

Chapter 3

The Theoretical Basis of the Annotation Scheme

3.1 Topological Fields

The annotation scheme for the TüBa-D/Z 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 (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 Erdmann (1886). Herling (1821) developed an adequate topological theory for complex sentences in which clauses form a topological unit carrying a syntactic function and he mentioned the special position of the finite verb in verb-second und verb-final clauses. Erdmann (1886) established the basics of a theory of topological fields and pointed out that the first position in a clause is not necessarily the subject position. The so called Herling/Erdmann scheme already covers a set of word order regularities which apply for all three clause types of German. Later Drach (1937) introduced the notion of *field*. Finally, Höhle (1986) developed topological schemes for the three clause types.

3.1.1 The Concept of Topological Fields

In a German clause, 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 clauses into 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* (final field). The Vorfeld and the Mittelfeld are divided by the *linke Satzklammer* (left sentence bracket), which is the finite verb, the *rechte Satzklammer* (left sentence bracket) is the verb complex between the

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

Mittelfeld and the Nachfeld. Thus, the theory of topological fields states the fundamental regularities of German word order. It is an important basis for the topological analysis of any German sentence, since subclauses and embedded clauses are treated within the bounds of fields. Identical word order regularities within a specific field can be realized in all three clause types. But the fields themselves differ in their possible elements and grammatical rules. Therefore, the theory is a descriptive rather than explanative theory for a specific language.

Höhle (1986) denotes the three clause types as E-Sätze (verb-final clauses), F1-Sätze (verb-initial clauses), and F2-Sätze (verb-second clauses). The topological schemes of these types are listed in 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 (1986), we assume the existence of the following topological fields (cf. Table 3.2):

The following description of the topological fields does not claim completeness 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, the subject can also

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

Table 3.2: Topological fields

Field	Description
VF	Vorfeld (initial field)
LK	Linke (Satz-)Klammer (left sentence bracket)
MF	Mittelfeld (middle field)
VC	Verbkomplex (verb complex)
NF	Nachfeld (final field)
LV	Linksversetzungsfeld (field for resumptive constructions)
C	C-Feld (field for complementizers, left from MF)
KOORD	Koordinationsfeld (field for coordinating particles)
	left-most element, optionally in all clause types, (e.g. *und, oder*)
PARORD	Koordinationsfeld (field for non-coordinating particles)
	left-most element, optionally only in verb-second (e.g. *denn, weil*)

occur in the Mittelfeld, thus allowing almost every other constituent to occupy the Vorfeld.

LK: The *Linke Klammer* is the position of the finite verb in verb-second and verb-first clauses or a conjunction in verb-final clauses. It 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*. It consists of a sequence of any number of constituents including none. The linear order of the constituents depends on the specific word order principles for German and their interaction.

VC: The *Verbkomplex* is a sequence of verb forms. In verb-second and verb-first clauses it consists of one or more non-finite elements or - depending on the verb - of a separable prefix. In verb-final clauses it 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 usually the right-most element (exception: *Ersatzinfinitiv* constructions (*daß er sich ein neues Konzept wird überlegen müssen*) (cf. 4.7.3).

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 any number of constituents including none.

LV: The *Linksversetzungsfeld* is a field for resumptive constructions. A *Linksversetzung* is a pendent constituent. It can be regarded as a syntactic anticipation of a part of a sentence (cf. 6.1.4). There are many restrictions which apply for this position.

C: The *C-Feld* only occurs in verb-final clauses (exception: the conjunction *als* in subordinated sentences of comparison *als wäre es nie geschehen.*). It is obligatorily occupied in finite verb-final clauses. In non-finite verb-final clauses the C-position may be empty. This field can 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 Geld' geht es dabei? / ..., 'an der' Max Daniel Professor für Klavier ist.* (cf. 6.1.1).

KOORD: The KOORD-field is the field for coordinating particles. In contrast to the PARORD-field, it can optionally occur as the left-most element of all clause types (cf. 6.1.2).

PARORD: The PARORD-field is the field for non-coordinating particles which optionally occur as the left-most element of a verb-second clause (cf. 6.1.3).

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

The combination of these fields in order to constitute verb-first, verb-second, or verb-final clauses is described in Höhle (1986).

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 consitutent 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 and edge 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 contain only **one** constituent.
   More than one constituent is allowed within MF and NF.

2. **Types of constituents within the fields:**
   Phrasal constituents occur in VF, MF, NF and C (interrogative or relative phrases). Embedded clauses either belong to NF, VF, LV, or in some cases to MF. Usually, outside the spoken language context, verb-final clauses do not occur isolated. They need to be attached if possible.

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. Constituents which cannot be assigned a grammatical function within a syntactic construction 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* prescibes 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 treebank. The sequence of the fields in the three clause types never violates the topological schemes given by Höhle (1986). Within each sentence structure, in general at least two topological fields are occupied (exception: infinitive constructions, (cf. 4.7.4). Others may be left empty (elliptical constructions, cf. 6.5). Table 3.3 lists the four levels of annotation which we distinguish within the structure of an annotated syntactic tree:

Table 3.3: Levels of annotation

Level	Inventory
clause level	root node labels for different types of clauses
field level	node labels for topological fields (including labels for conjuncts of 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)

Node labels denote the syntactic category of a phrase or sentence, a topological field, or a grammatical property. 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** tags 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 3.3.1 (cf. (Schiller et al. 1995)). The tagging of the data was performed by the tnt tagger ((Brants 1998)) and manually corrected in the *Annotate* tool.

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. Table 3.4 lists all node labels which are used in the 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, including **secondary edge labels**. The latter ones are used to resolve ambiguities on a different level of description.

Two specific edge labels denote whether a constituent has the function of a head (HD), e.g. a phrase (NX, PX, ADJX, ADVX, VXFIN, VXINF), or a non-head (-), e.g. a determiner or a modifier attached to a phrase. On any annotation level, there is at most one head. Within phrases, these two labels indicate the internal dependency structure of the phrase. The head of a sentence structure (e.g. SIMPX) is always the finite verb. In coordinations, each conjunct depends on the head of the whole construction and is

[2]PAV 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.

denoted with a specific edge label (KONJ) in order to distinguish them from conjunctions and modifying elements within a coordination (cf. 6.4.1 and 6.4.3). Edge labels below all root node labels carry only non-head labels (cf. (Kübler and Telljohann 2002)).

In order to mark proper nouns, and named entities within the treebank, the node label (EN-ADD) and the secondary edge label (EN) are defined (see Table 3.6).

EN-ADD is inserted between two nodes to indicate that the node below represents a complex proper noun, (e.g. *Ute Wedemeier, The Jim Wane Swingtett*), a single proper noun tagged as NN with respect to the STTS (e.g. Sögestraße), or a named entity (e.g. *Auf die stürmische Art, Built to Spill*) (cf. 4.2.5). EN-ADD is either directly attached to a head noun of a phrase or to a field. If it has a postmodifier, its mother node is NX which represents the nominal status of EN-ADD. The internal syntactic structure of EN-ADD is governed by the general annotation rules.

The secondary edge label **EN** gives information about the relation of two parts of a proper noun within a complex phrase consisting, for instance, of an article and/or an attributive adjective which do not belong to the proper noun itself, e.g. *der [zweite Weltkrieg EN]* (cf. 4.2.5).

Table 3.3.1: The STTS tagset

POS =	description	examples
ADJA	attributive adjective	*[das] große [Haus]*
ADJD	adverbial or predicative adjective	*[er fährt] schnell, [er ist] schnell*
ADV	adverb	*schon, bald, doch*
APPR	preposition; left circumposition	*in [der Stadt], ohne [mich]*
APPRART	preposition + article	*im [Haus], zur [Sache]*
APPO	postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	right circumposition	*[von jetzt] an*
ART	definite or indefinite article	*der, die, das, ein, eine*
CARD	cardinal number	*zwei [Männer], [im Jahre] 1994*
FM	foreign language material	*[Er hat das mit "] A big fish [" übersetzt]*
ITJ	interjection	*mhm, ach, tja*
KOUI	subordinating conjunction with *zu* + infinitive	*um [zu leben], anstatt [zu fragen]*
KOUS	subordinating conjunction with clause	*weil, daß, damit, wenn, ob*
KON	coordinative conjunction	*und, oder, aber*
KOKOM	particle of comparison, no clause	*als, wie*
NN	noun	*Tisch, Herr, [das] Reisen*
NE	proper noun	*Hans, Hamburg, HSV*
PDS	substituting demonstrative pronoun	*dieser, jener*
PDAT	attributive demonstrative pronoun	*jener [Mensch]*
PIS	substituting indefinite pronoun	*keiner, viele, man, niemand*
PIAT	attributive indefinite pronoun without determiner	*kein [Mensch], irgendein [Glas]*
PIDAT	attributive indefinite pronoun with determiner	*[ein] wenig [Wasser], [die] beiden [Brüder]*
PPER	irreflexive personal pronoun	*ich, er, ihm, mich, dir*
PPOSS	substituting possessive pronoun	*meins, deiner*
PPOSAT	attributive possessive pronoun	*mein [Buch], deine [Mutter]*
	relative pronoun	
PRELS	substituting	*[der Hund,] der*
PRELAT	attributive	*[der Mann ,] dessen [Hund]*
PRF	reflexive personal pronoun	*sich, einander, dich, mir*
PWS	substituting interrogative pronoun	*wer, was*
PWAT	attributive interrogative pronoun	*welche [Farbe], wessen [Hut]*

POS =	description	examples
PWAV	adverbial interrogative or relative pronoun	*warum, wo, wann, worüber, wobei*
PROP	pronominal adverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	*zu* + infinitive	*zu [gehen]*
PTKNEG	negation particle	*nicht*
PTKVZ	separated verb particle	*[er kommt] an, [er fährt] rad*
PTKANT	answer particle	*ja, nein, danke, bitte*
PTKA	particle with adjective or adverb	*am [schönsten], zu [schnell]*
TRUNC	truncated word - first part	*An– [und Abreise]*
VVFIN	finite main verb	*[du] gehst, [wir] kommen [an]*
VVIMP	imperative, main verb	*komm [!]*
VVINF	infinitive, main	*gehen, ankommen*
VVIZU	infinitive + *zu*, main	*anzukommen, loszulassen*
VVPP	past participle, main	*gegangen, angekommen*
VAFIN	finite verb, aux	*[du] bist, [wir] werden*
VAIMP	imperative, aux	*sei [ruhig !]*
VAINF	infinitive, aux	*werden, sein*
VAPP	past participle, aux	*gewesen*
VMFIN	finite verb, modal	*dürfen*
VMINF	infinitive, modal	*wollen*
VMPP	past participle, modal	*[er hat] gekonnt*
XY	non-word containing special characters	*D2XW3, letters*
$,	comma	,
$.	sentence-final punctuation	. ? ! ; :
$(	other sentence internal punctuation	- [ ] ( )

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
FX	foreign language phrase
DP	determiner phrase (e.g. *gar keine*)
	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 coordinating particles
PARORD	field for non-coordinating particles
FKOORD	coordination consisting of conjuncts of fields
MFE	middle field between VCE and VC
VCE	verb complex with the split finite verb of *Ersatzinfinitiv* constructions
FKONJ	conjunct consisting of more than one field
	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 Heads and Conjuncts	
HD	head
-	non-head
KONJ	conjunct
Complement Edge Labels	
ON	nominative object
OD	dative object
OA	accusative object
OG	genitive object
OS	sentential object
OPP	prepositional object
OADVP	adverbial object
OADJP	adjectival object
PRED	predicate
OV	verbal object
FOPP	optional prepositional object
VPT	separable verb prefix
APP	apposition
Modifier Edge Labels	
MOD	ambiguous modifier
ON-MOD, OA-MOD, OD-MOD, OG-MOD, OPP-MOD, OS-MOD, PRED-MOD, FOPP-MOD, OADJP-MOD, V-MOD, MOD-MOD	modifiers modifying complements or modifiers e.g. V-MOD = modifier of the verb
Edge Labels in Split-up Coordinations	
ONK, ODK, OAK, FOPPK, OADVPK, PREDK, MODK, V-MODK	second conjunct (K) in split-up coordinations e.g. ONK = second conjunct of a nominative object (subject)
Secondary Edge Labels	
	dependency relation between:
REFVC	two verbal objects in VC
REFMOD	two ambiguous modifiers
REFINT	a phrase internal part and its modifier
REFCONTR	control verb and its complement across clause boundaries

Table 3.6: Labels for proper nouns and named entities

Labels	Description
Phrase Node Labels	
EN-ADD	proper noun or named entity (additional label)
Secondary Edge Label	
EN	phrase internal relation between two parts of a proper noun

3.4.3 What Is a Syntactic Unit?

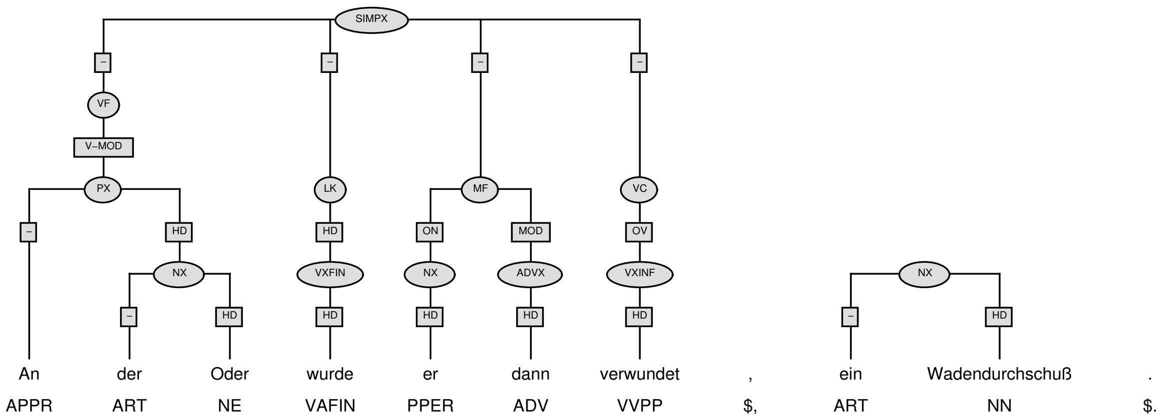
The newspaper articles of the taz have been defined as the primary segmentation domain of the data. They are preprocessed into syntactic units delimited by punctuation marks (. ? ! ; - ... /) for which specific rules demand or forbid segmentation. Each syntactic unit is assigned a specific code which identifies its origin in the newspaper data, eg. T990507.123 (T (taz) 99 (year) 05 (month) 07 (day) 123 (article)).

A syntactic unit usually consists of one complete sentence structure with a root node (SIMPX, R-SIMPX, P-SIMPX). But it may also consist of one or more sentences and/or phrases, e.g. headlines, titles, sentences with parentheses, sentences with discourse markers, or sentence conjunction by a colon.

An annotated tree is a complete syntactically and semantically well-formed construction according to the *longest match principle*. The model of topological fields does not prescribe that all fields have to be occupied. The fact that fields can be left empty, also helps us to cope with elliptical constructions (cf. 6.5).

Punctuation is not annotated, i.e., all punctuation marks are not attached to the tree structure. Exceptions are punctuation marks which carry a semantic meaning within a sentence, e.g. - (*bis, und*) in expressions like *15.30 - 17.30 Uhr*. They are tagged according to the part of speech that they represent in the text (cf. 4.4.1).

Constituents are not attached to a tree if they are not assigned a grammatical function within the specific syntactic construction. The following tree diagram shows two annotated trees in one syntactic unit:[3]



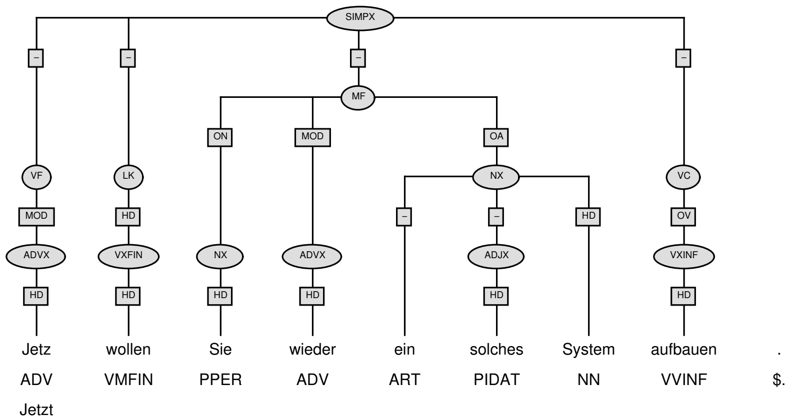
[3]These tree diagrams and all following tree diagrams in this report were generated with the aid of the Negra *Annotate* tool.

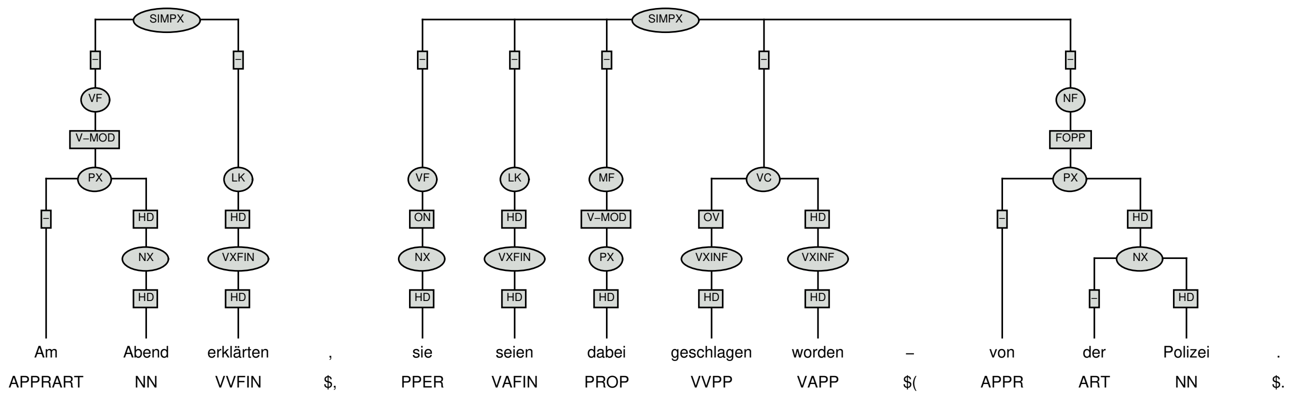
The leaves of the trees consist of pairs of non-terminal symbols and part-of-speech tags. Non-terminal symbols are represented by spherical nodes, edge labels by rectangular nodes. The tree diagram consists of two trees, a SIMPX and an isolated phrase. In accordance with the four annotation levels shown in Table 3.3, the sentence is annotated top-down by the root node (SIMPX), the field nodes (VF, LK, MF, and VC), the phrase nodes (PX, VXFIN, NX, ADVX, and VXINF), and finally the tagged lexical entries. The edge labels between the field level and the phrase level indicate that the syntactic structure contains one unambiguous modifier (V-MOD), a subject (ON), one ambiguous modifier (MOD), a verbal object (OV), and the finite verb, which itself is the head (HD) of the entire syntactic construction. The noun phrase (*ein Wadendurchschuß*) 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.

3.4.4 Printing and Spelling Errors

In contrast to spoken language data like in the VERBMOBIL (cf. (Stegmann et al. 2000)) which exhibit fragmentary utterances, false starts, repetitions, interruptions, and hesitation noises as its characteristic properties, data taken from newspaper corpora does not include unintentionally formed syntactic constructions.

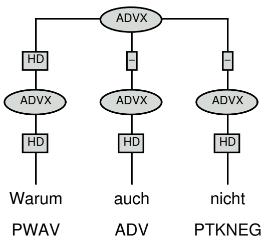
Deviations from syntactic wellformedness are either intended by the author or are caused by printing errors. While incorrect writing of words is neglected in the syntactic analysis (the respective lexical entry is marked with the correct writing of the word in a comment line below), lexical elements which do not belong to the syntactic construction (intentional or unintentional) are structured as much as possible, but are not attached to the surrounding constituents:

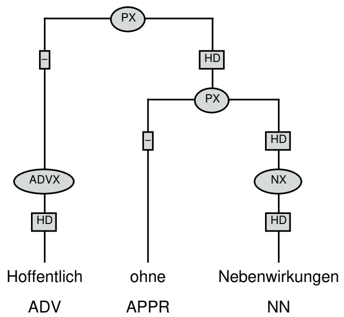




3.4.5 Isolated Phrases

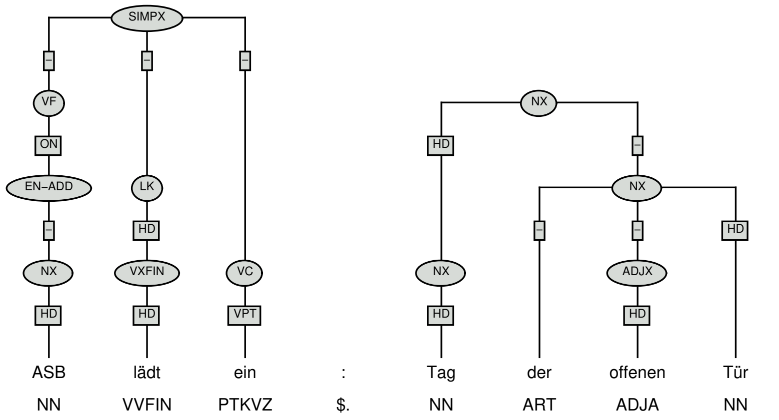
There are textual fragments in newspaper data which cannot be analysed as a SIMPX or as a constituent of a SIMPX because they are lacking a verbal constituent or they are not assigned a specific grammatical function within a well-formed sentence. These fragments are annotated as isolated phrases. The isolated elements 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, so that a conflict with the topological field analysis is avoided. Their root node carries a phrasal category of their lexical head (NX, PX, ADVX, etc.):

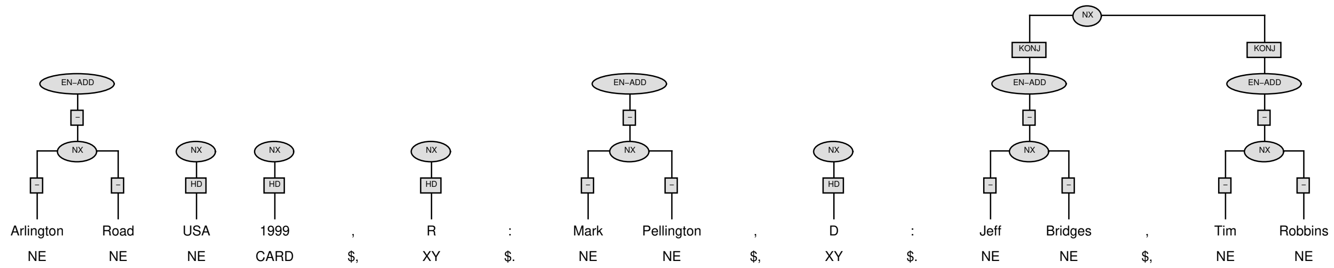


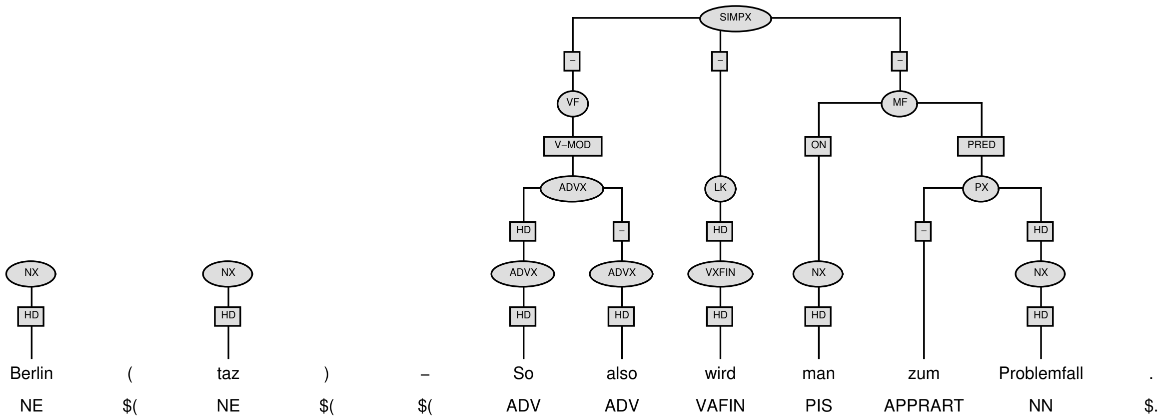


In accordance with the *longest match principle*, as many parts of the fragment as possible are projected to the phrase level and are included into a tree structure. It has to be decided which part of the whole construction is the head and which parts depend on this head.

Phrases within a syntactic unit are not attached on a higher level if they do not show dependency relation. This is often the case with syntactic elements which are separated by a colon or a dash (cf. 5.3.2):





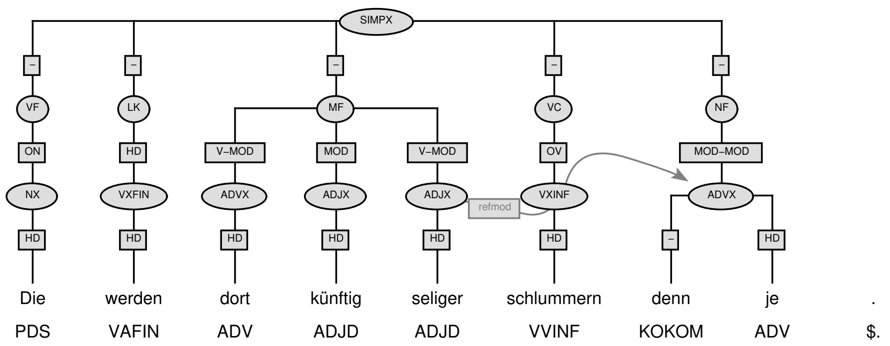


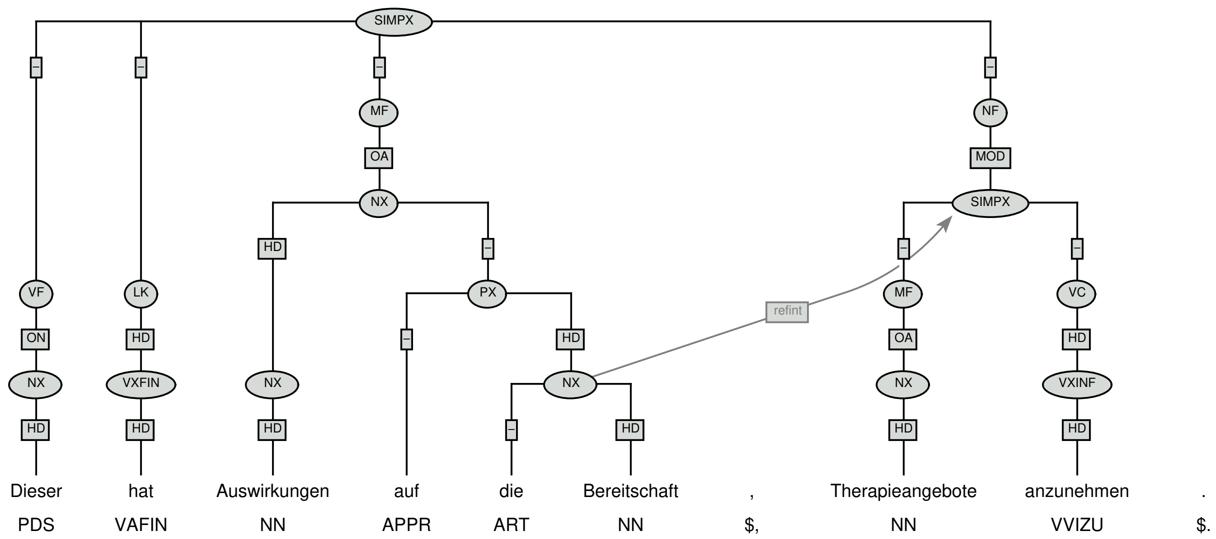
3.4.6 Long-Distance Dependencies

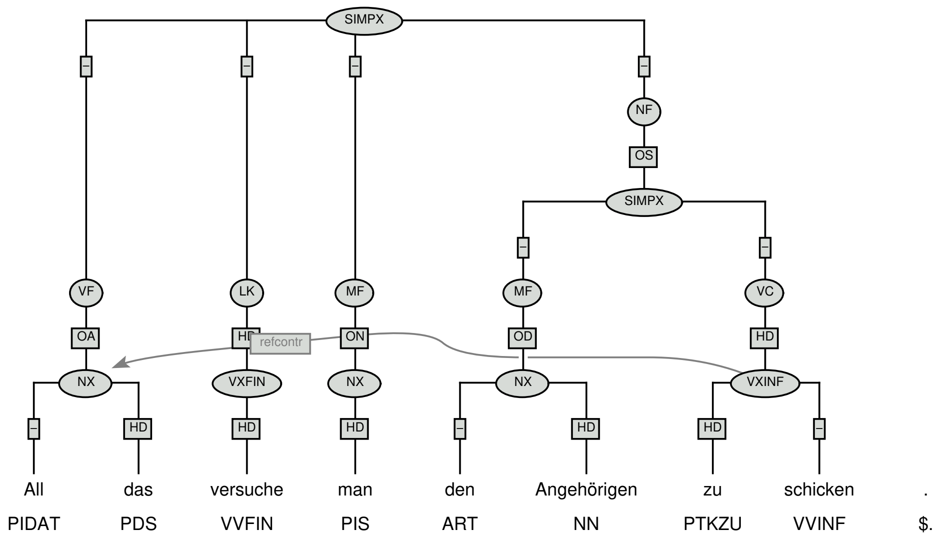
Our annotation scheme facilitates a surface-oriented representation of long-distance dependencies without crossing branches and traces. If a modifying constituent is not adjacent to the modified constituent, their dependency relation, which can even go beyond the border of topological fields, is encoded by special naming conventions for edge labels. We use edge labels such as OA-MOD (referring to OA) or PRED-MOD (referring to PRED) etc. expressing the non-ambiguity of the modifier.

Beyond this, we make use of secondary edge labels for ambiguity resolution. These labels just serve as additional information to the grammatical functions encoded in the edge labels. These secondary edge labels indicate underspecified long distance dependencies in the following cases:

1. If the above mentioned edge labels need further disambiguation, e.g. if there are two OAs or V-MODs below one SIMPX node (REFMOD).
2. If the dependency relation exists between two nodes of which at least one is phrase internal and therefore carries only head or non-head information (REFINT).
3. If there is a dependency relation outside of SIMPX in control verb constructions (REFCONTR).







3.4.7 Empty Categories

In general, an empty category analysis, e.g. for phrases without heads, is being avoided in the TüBa-D/Z treebank. But specifiers, prepositions,[4] complementizers discourse markers, KOORD and PARORD constituents, conjunctions, and unambiguous modifiers (that are attached to phrases immediately rather than to topological fields) are not labeled with grammatical functions. Furthermore, the edges below the SIMPX node are empty. They are not labeled in order to speed up annotation where the information is unnecessary or self-evident.

Furthermore, empty edge labels are used in elliptical phrases, e.g. noun phrases only consisting of an article and and an attributive adjective (cf. 6.5).

[4]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, the annotated head of a prepositional phrase is the NX (or complement) rather than the preposition itself. Therefore, prepositions carry no edge label.

Chapter 4

The Annotation of the Internal Structure of Phrases

4.1 Premodification and Postmodification in Phrases

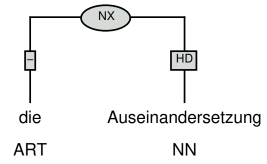
The annotation of phrases is also carried out following the *flat clustering principle* in order to keep the number of hierarchy levels in a syntactic structure as small as possible. As will be shown in the following sections, phrases may include adjectival or nominal premodifiers and/or postmodifiers of any syntactic category. Both kinds of modifiers are in principle projected to their phrase levels. Since the modification scope of premodifiers is unambiguous, they are directly attached to the head of the phrase which they modify. By contrast, postmodifiers are always attached on a higher level to preserve ambiguity. This decision, referred to in 3.3 as the *high attachment principle*, was taken to avoid the problematic distinction whether a postmodifier is a free adjunct or a complement of the modified phrase. The attachment strategy for premodifiers and postmodifiers is applied for all categories of phrases.

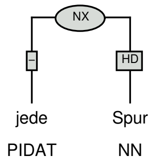
4.2 Noun Phrases

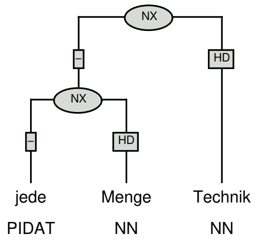
A simple noun phrase (NX) consists of a head noun (noun, proper noun, or a pronoun), (optionally) a determiner and (optionally) an adjectival or a nominal premodifier of any complexity preceding the head noun. A complex noun phrase is a simple noun phrase with a postmodifier of any syntactic category and complexity.

4.2.1 Prenominal Modification

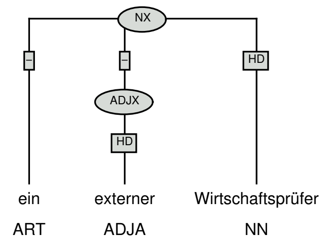
In a simple noun phrase, both the determiner and the head noun are directly attached on the same level to NX so that the label of the head noun carries the edge label HD and the edge label of the determiner is empty.

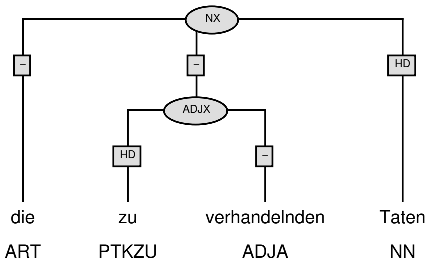


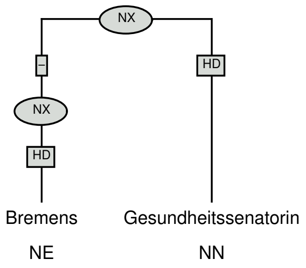




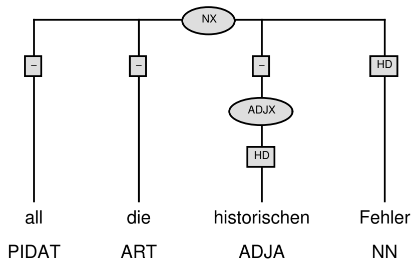
Since prenominal modifiers are directly attached to the head noun on the same level, their edge labels are empty (whereas the edge labels of modifiers that are attached to topological fields are non-empty (cf. 8.1.2)). Prenominal modifiers are either attributive adjectives or preceding genitive phrases:



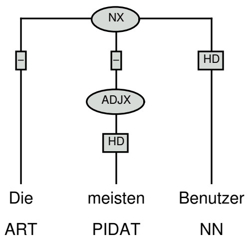


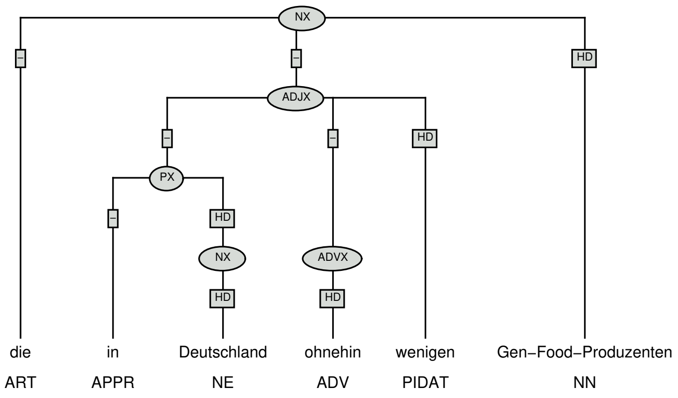


If there is a PIDAT preceding the article it is directly attached to the noun phrase.

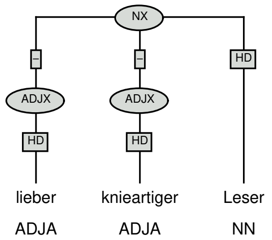


If a PIDAT is following the article in adjective position it is projected to its phrase level (ADJX) with possible premodifiers and then directly attached like an attributive adjective to the noun phrase.

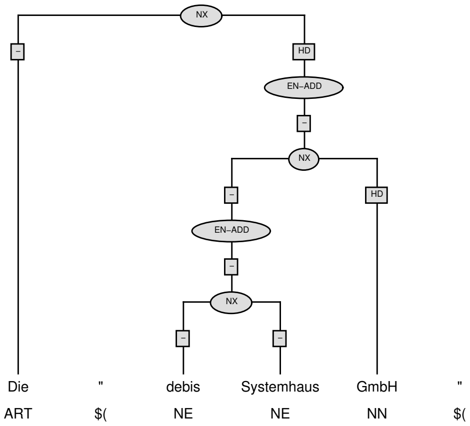


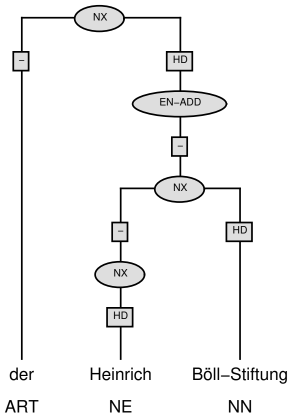


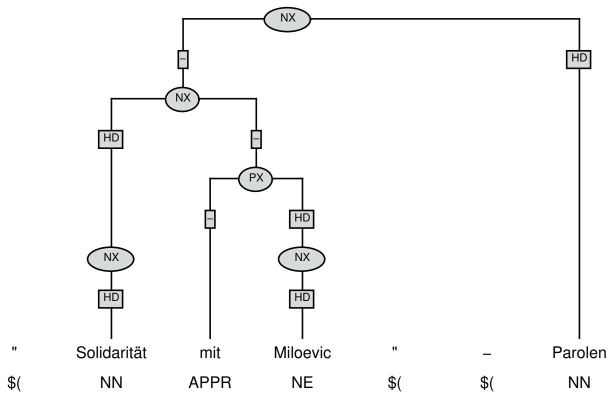
If there is more than one prenominal modifier, the one left of the noun is modifying the following noun, the one left of the modifier is modifying both, the modifier and the noun, and so on. All of these modifiers are attached to the head noun on the same level which yields a rather flat noun phrase structure. This stategy is justified by the fact that these modifiers have a scope of modification beyond the adjectival phrase, e.g. as in coordinated noun phrases like *insgesamt 12.000 Studienplätze und 15.000 Lehrstellen*, the adverb *insgesamt* modifies *12.000 Studienplätze* as well as *15.000 Lehrstellen*.

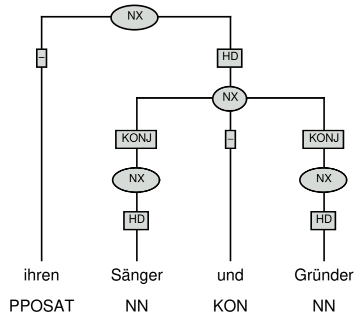


In case of complex head nouns, e.g. complex (proper) nouns consisting of two nominal parts or coordinated head nouns (cf. 6.4.5), first the complex noun respectively the coordination (cf. 6.4) is annotated with its own internal dependency structure then the determiner and possible premodifying adjectival phrases are attached on a higher level.



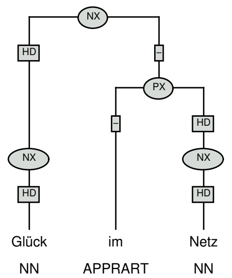


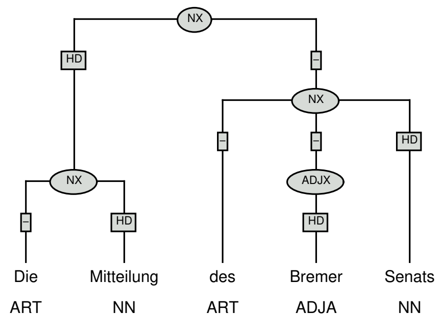




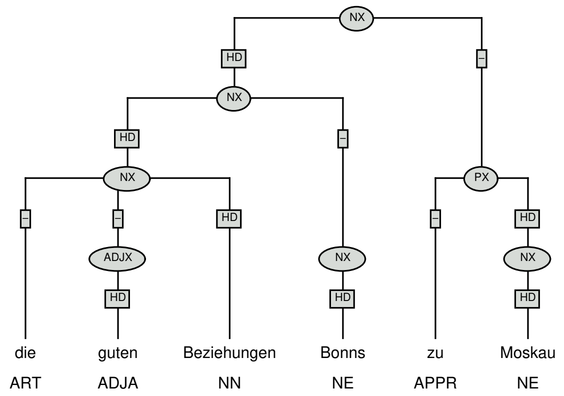
4.2.2 Postnominal Modification

Whereas prenominal modifiers are always directly attached to the head noun on the same level, postnominal modifiers are attached to the head noun on a higher level. Postnominal modifiers are also always first projected to the phrase level before they are attached to the head noun on a higher level. Phrase internal postmodifiers can be of any phrasal category. The following tree structures show a prepositional phrase (PX) and a genitive phrase (NX) as postmodifiers:

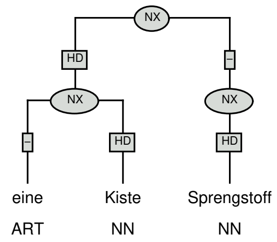




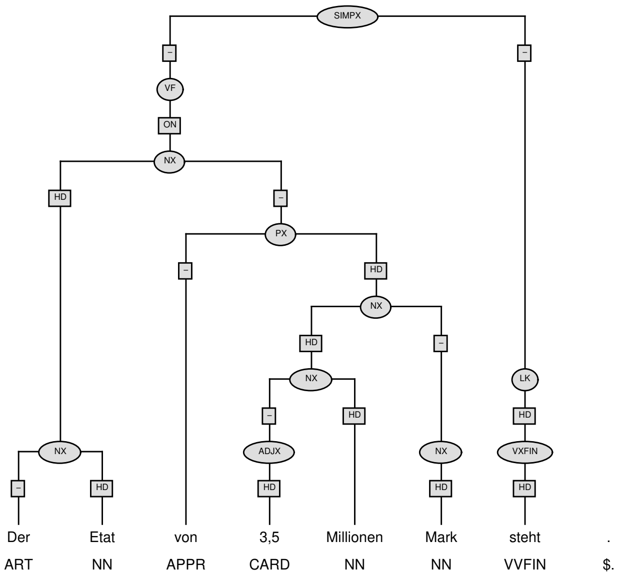
If there is more than one postmodifier of a noun, these modifiers usually show a hierarchical structure, for example, the first modifier modifies the head noun, the second modifier modifies the complete preceding noun phrase structure, and so on.

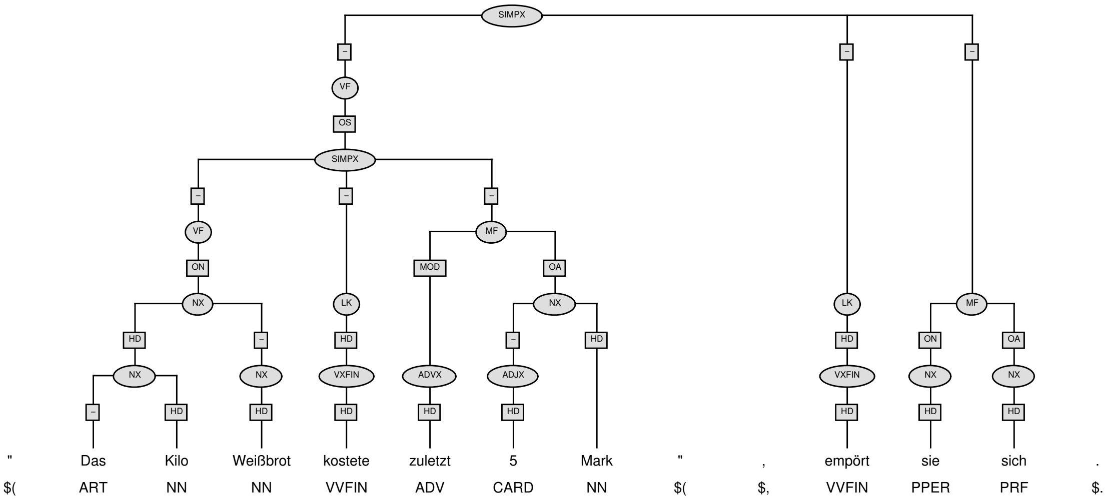


Attributes of degree and quantity nouns are also defined as postnominal modifiers:

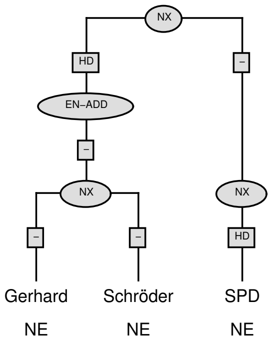


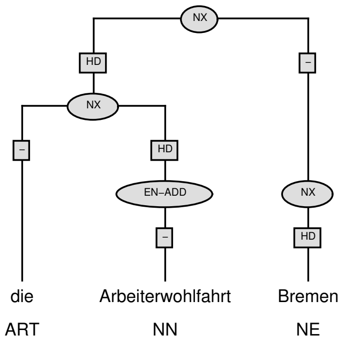
Cardinal numbers either appear as quantity nouns or premodifying adjectival attributes, e.g. the cardinal number *1,000,000* can also be expressed by the quantity noun *eine Million*. Therefore, we have to distinguish the following two ways of annotation:

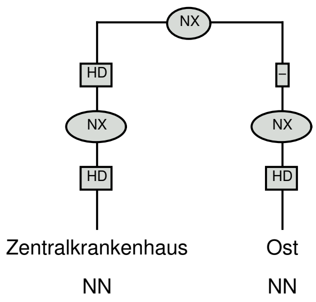


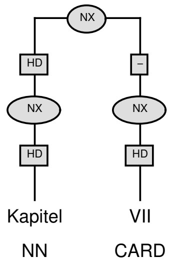


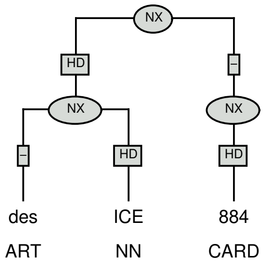
For nominal postmodifiers apart from genitive phrases the same attachment rule is applied. This kind of postmodifiers which may also appear in brackets, e.g. *Gerhard Schröder (SPD)*, is semantically closely related to the preceding head noun phrase. *die Arbeiterwohlfahrt Bremen*, for instance, means *die Arbeiterwohlfahrt* which is located in *Bremen*, but does not mean *die Arbeiterwohlfahrt* which is called *Bremen*. Therefore, these postmodifiers have to be distinguished from appositions (cf. 4.2.3) and complex proper nouns (cf. 4.2.5).







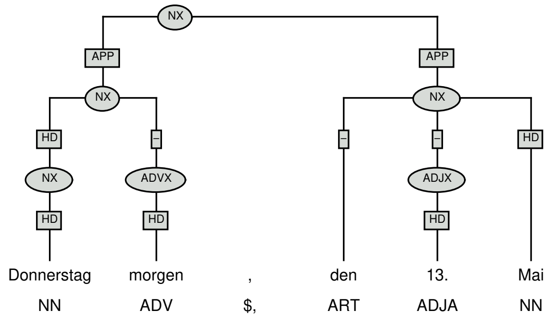


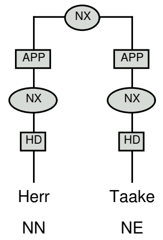


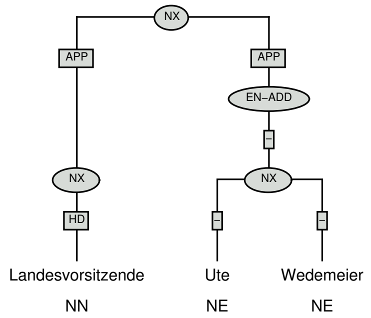
4.2.3 Appositional Constructions

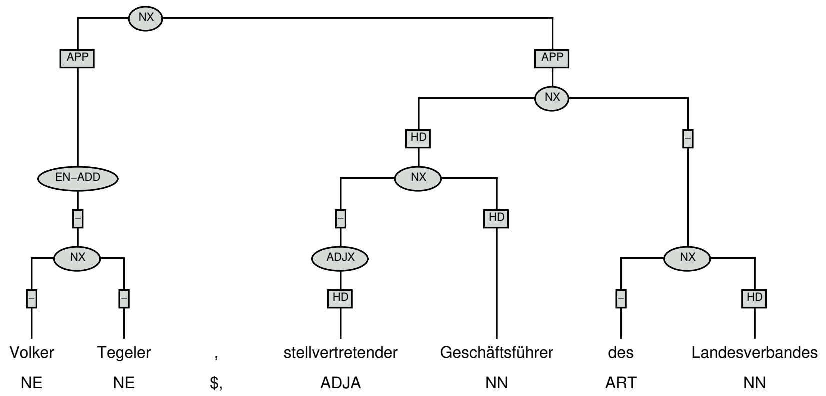
An apposition is a specific kind of an attribute to a noun, which agrees in case with this noun and does not change its overall meaning. There is no consensus among grammarians of what is exactly meant by the notion *apposition* (cf. (Eisenberg 1999-2001)). Eisenberg (1999-2001), for instance, claims that, e.g. *Ute Wedemeier die Landesvorsitzende* and *die Landesvorsitzende Ute Wedemeier* are both appositions but it is not clear which part is the apposition and which part is the head noun. The Duden Grammar (1995) distinguishes between loosely constructed appositions (*lockere Apposition*) (e.g. *Ute Wedemeier, die Landesvorsitzende,*), which follow the head noun separated by a comma, and tightly constructed appositions (*enge Apposition*) (e.g. *(die) Landesvorsitzende Ute Wedemeier*), which precede the head noun (cf. (Drosdowski 1995)). According to Helbig/Buscha (1998) there is case agreement between loosely constructed apposition and head noun which are separated by a punctuation mark. By contrast, Engel (1996) thinks that only loosely constructed appositions can be regarded as appositions. He treats tightly constructed appositions as *nomen varians* or *nomen invarians*.

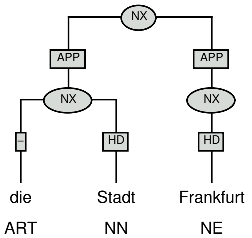
Because of these different definitions of the notion of apposition, we do not decide on what is the head noun and what is the apposition. We assume referential identity between the two parts. Loosely constructed appositions as well as tightly constructed appositions are treated as appositional constructions, i.e., the head noun and its apposition form a complex structure which does not give any information about head assignment. Therefore, both parts are first projected to their phrase level and then coordinated on a higher level, each of them labeled as apposition (APP), i.e. as a part of an appositional structure. What is important is the referential identity in meaning. Thus, *Nummer 1* is an appositional construction, whereas *Seite 1* is a noun phrase with the postmodifier *1*. Forms of address for persons and titles, e.g. *Herr, Frau, Doktor (Dr.), Professor (Prof.), Bundeskanzler*, are also treated as appositional constructions. Here are some examples:

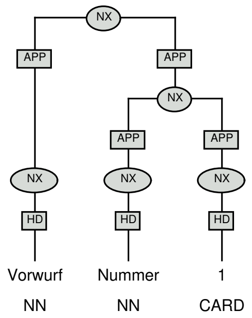


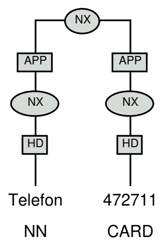




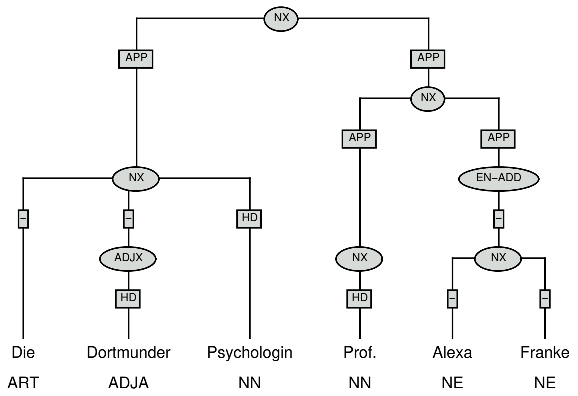




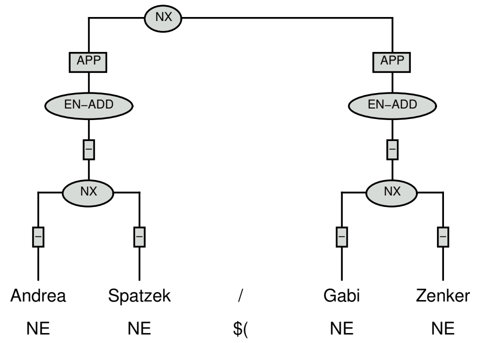




In case of a form of address combined with one or more titles preceding a name, we annotate an embedded appositional construction:

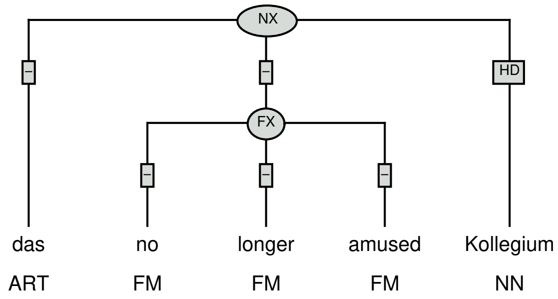


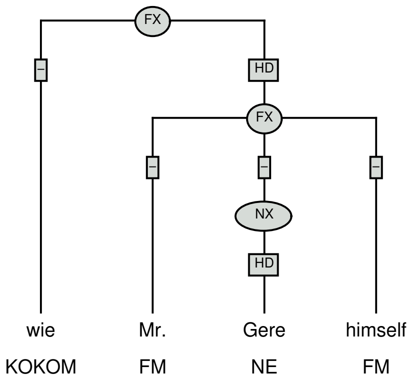
The same way, we treat proper nouns in brackets which are identical to the preceding proper noun, for example, an actor's name and role:



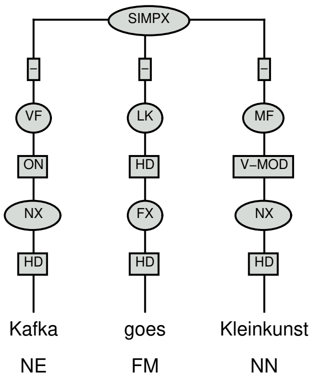
4.2.4 Foreign Language Material

Words or parts of a text written in a foreign language except foreign language proper nouns are tagged as foreign language material (FM), e.g. *hello* (FM), *no* (FM) *longer* (FM) *amused* (FM). All parts of foreign language proper nouns (4.2.5) are tagged as NE (e.g. *Mary*, *New* (NE) *York (NE)*, *University* (NE) *of* (NE) *Illinois* (NE)). Single foreign words are projected to a syntactic level assigned the node label FX, which is an universal label for any syntactic category (phrasal and sentential) in the respective foreign language. More complex parts of a text tagged as FM are attached on the same level without any internal syntactic structure and head assignment. Their mother node is also assigned the label FM, e.g. *no longer amused*. For foreign language constructions containing a foreign language proper noun, the annotation strategy is the following: in a first step, all NEs are projected to the phrase level (NX), in a second step, these phrase node labels together with all FMs are projected to the next higher level with the node label FX. Again, there is no head assignment directly below the FX node, e.g. *Mister Gere himself*.

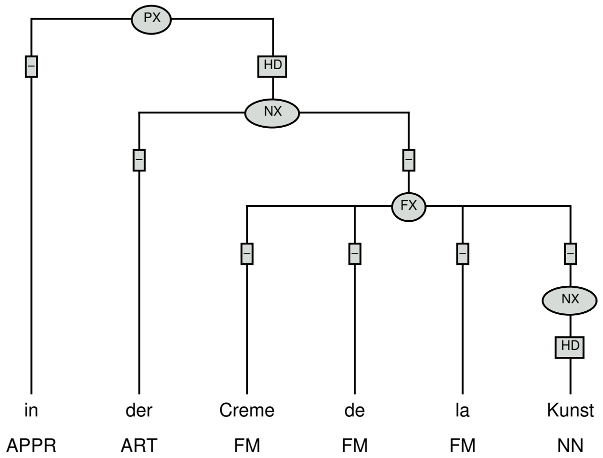


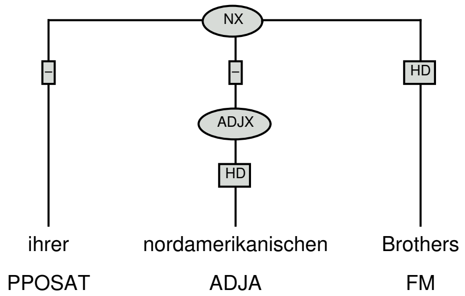


Often, foreign language material is a part of a German syntactic construction and plays the role of a grammatical function. Therefore, the FX node is attached as a constituent to the tree structure. If it is directly attached to a field or a sentence bracket, the edge label above the FX node denotes its grammatical function within the clause, e.g. *Kafka* **goes** *Kleinkunst* (head of the clause).

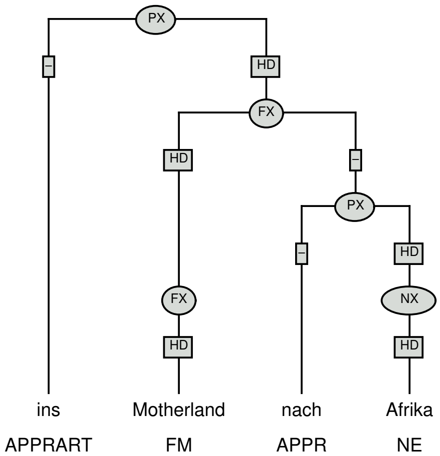


If a single FM is head of a phrase which can be identified as a German phrase, e.g. by an article and/or an adjective (noun phrase), it is projected to the specific phrasal category, e.g. NX instead of FX in a construction like *in der Creme de la Kunst, die nordamerikanischen Brothers.*





If FX is modified by a German postmodifier the mother node of the complex phrase is also FX, which again may be preceded by a German preposition resulting in a PX, e.g. *ins Motherland nach Afrika.*



4.2.5 Proper Nouns and Named Entities

Proper nouns denote individual living beings, objects, etc. which exist only once as entities with their own specific properties. The distinction between proper nouns and common nouns is not always clear-cut. On the one hand, proper nouns can also become common nouns, e.g. *Opel* as the company is a proper noun PoS-tagged as *Opel* (NE), on the other hand, *Opel* as the car is a common noun PoS-tagged as *Opel* (NN).

In addition to the categories of proper nouns listed in the STTS guidelines (first and last names of persons, names of companies, and geographical names), we also define names of streets and places, individual names of institutions (e.g. *Max-Planck-Institut, Deutsches Museum*, events (e.g. *Zweiter Weltkrieg*), and titles of books, movies, etc. as specific categories of proper nouns.

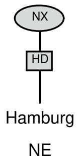
Since the PoS-tagging of proper nouns follows the categories of proper nouns in the STTS, some nouns which belong to our class of proper nouns (e.g. composed forms of NE + NN like *Sögestraße*) are tagged as NN. Complex proper nouns forming a syntagma as well as titles, names of historical events, institutions, and so on, are PoS-tagged according to their distribution (e.g. *der* (ART) *Potsdamer* (ADJA) *Platz* (NN), *Auf* (APPR) *die* (ART) *stürmische* (ADJA) *Art* (NN)). This kind of proper nouns, we define as named entities.

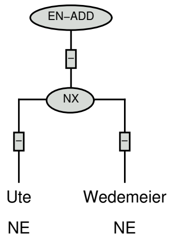
In order to distinguish common nouns from proper nouns and named entities, the latter ones are assigned an additional node label EN-ADD above their mother node. EN-ADD subsumes all single proper nouns which are not tagged as NE, complex proper nouns tagged as NE as well as named entities tagged according to their distribution.

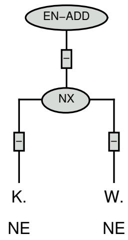
The secondary edge label EN is used for complex proper nouns if EN-ADD cannot be used because of the annotation rules for the internal structure of nouns phrases (e.g. *den gleichgeschalteten Hamburger Reichssender*) (cf. 4.2.5). Since German and foreign language named entities differ concerning their PoS-tagging, their internal syntactic structure also differs in consequence. Our annotation strategy of proper nouns and named entities will be demonstrated in the following.

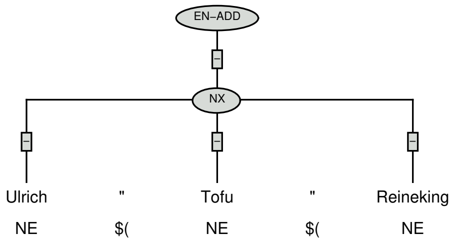
German Proper Nouns

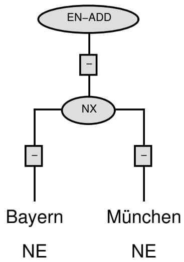
German proper nouns denoting individual entities in the above mentioned sense consist of one or more lexical elements tagged as NE. In case of a single NE, this NE is projected to its phrase level like common nouns carrying the node label NX. Proper nouns consisting of two or more NEs are attached on the same level. None of them carries a head label to indicate that there is no obvious dependency relation between them, e.g. first name and last name of a person, initials, (included) nick names, and names of institutions consisting of NEs. As mentioned above, these complex proper nouns are assigned the additional node label EN-ADD.



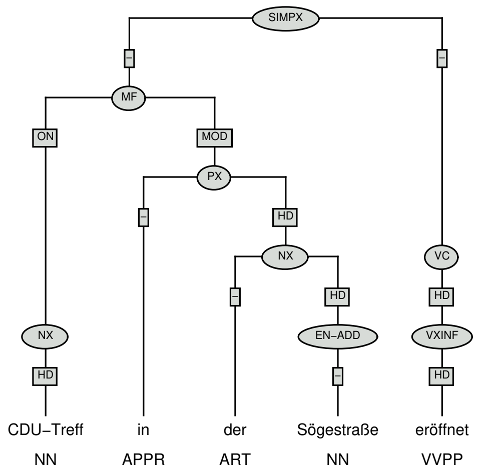




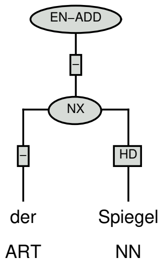




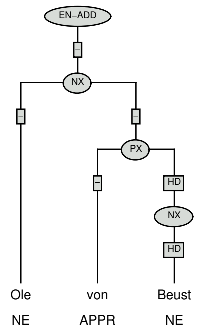
Proper nouns which are not tagged as NE, e.g. composed forms of NE+NN (*Sögestraße*) or complex phrases (*Der Spiegel*) also get an additional EN-ADD node. If a preceding article does not belong to the proper noun itself, EN-ADD is directly projected from the lexical level of the noun:



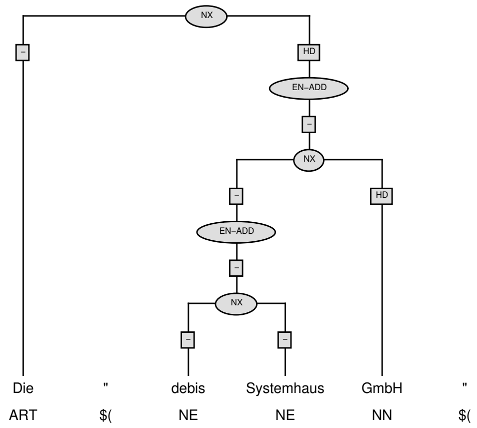
If the proper noun, is a complex syntactic structure, e.g. a phrase or a sentence, the lexical elements of this syntactic structure are tagged according to their distribution. First the whole phrase/sentence is annotated before EN-ADD is added:



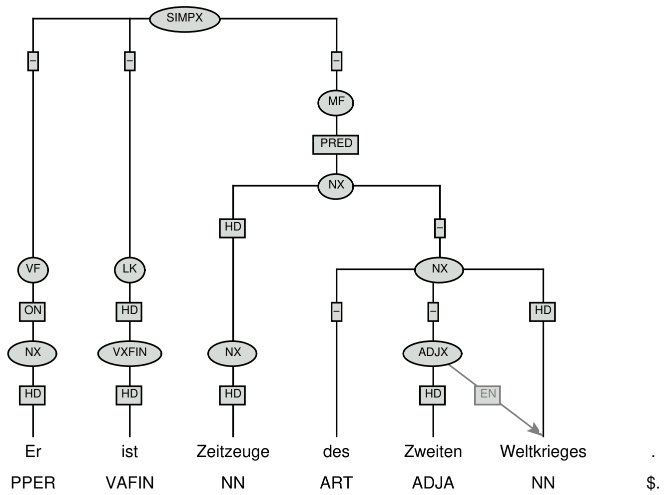
If proper nouns include other parts of speech than NEs, these are tagged according to their distribution. Therefore, proper nouns with a preposition include a prepositional phrase.

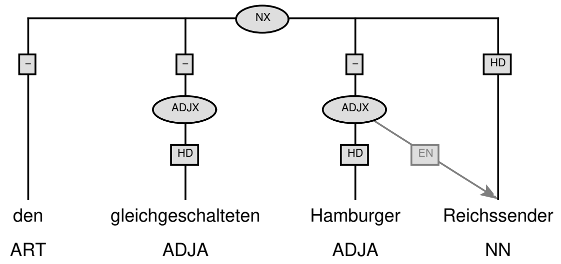


If a proper noun occurs within a more complex proper noun, EN-ADD is annotated on both levels of proper nouns:

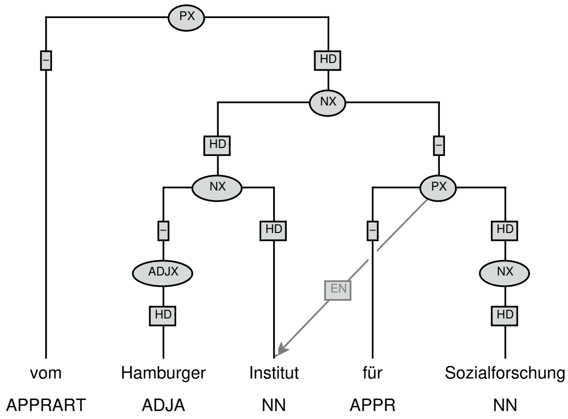


If the original form of a proper noun (e.g. *Zweiter Weltkrieg, Hamburger Reichssender*) is inflected and/or premodified by an article and/or attributive adjective, the included proper noun is indicated by the secondary edge label EN. EN always points from the dependend part of the proper noun to its head noun:

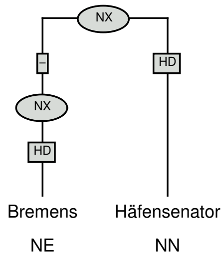




The secondary edge label is also used to mark that a postmodifier is part of a proper noun:



Proper nouns may occur as a prenominal genitive with the function of an attributive adjective:

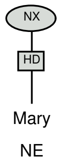


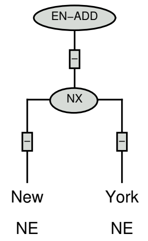
Proper nouns may be modified postnominally by a another proper noun, e.g. *Arbeiterwohlfahrt Bremen* (cp. 4.2.2), which is no proper noun as a whole because *Bremen* only specifies a specific location of *die Arbeiterwohlfahrt*, whereas *München* in *Bayern München* is part of the name of the soccer club. *Bremen* in a construction like *das Motorschiff Bremen* belongs to an appositional construction (cf. 4.2.3).

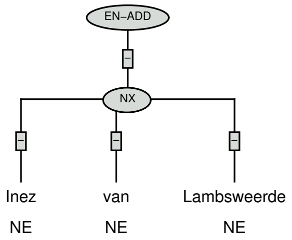
Foreign Language Proper Nouns

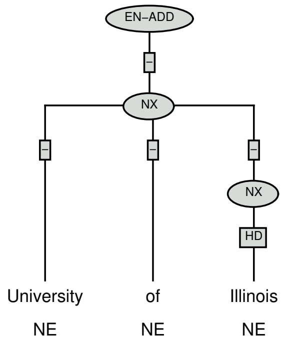
Since all elements of foreign language proper nouns are tagged as NE, they are annotated in the same way as German proper nouns, i.e., they are either a single proper noun or a complex proper noun with an EN-ADD node.

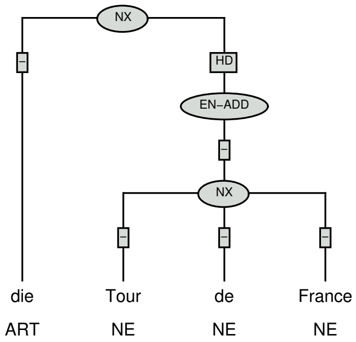
Single foreign language proper nouns are projected to their phrase level, complex ones are attached on the same level without head assignment. Their mother node is NX with the additional node EN-ADD. Note the difference between *Ole von Beust* (cf. 4.2.5) and *Inez van Lambsweerde* which results from the PoS-tagging rules. If the proper noun consists of more than one lexical element and if it has a German article, the article is attached on a higher level.


NX
HD
Mary
NE


EN–ADD
NX
New
York
NE
NE

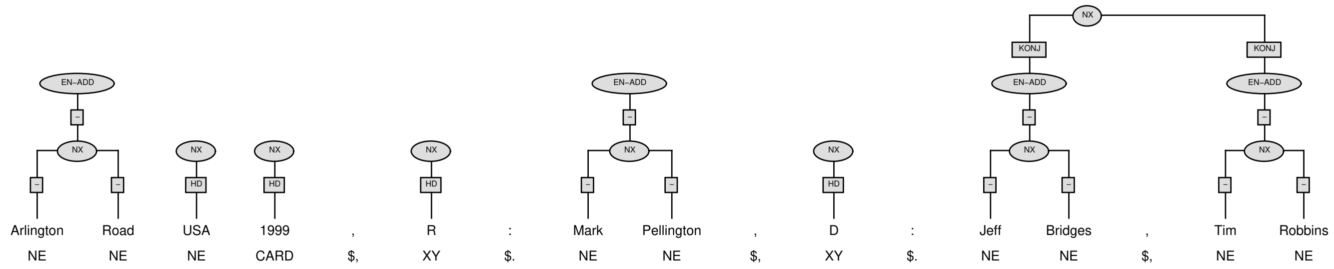

EN–ADD
NX
Inez
van
Lambsweerde
NE
NE
NE

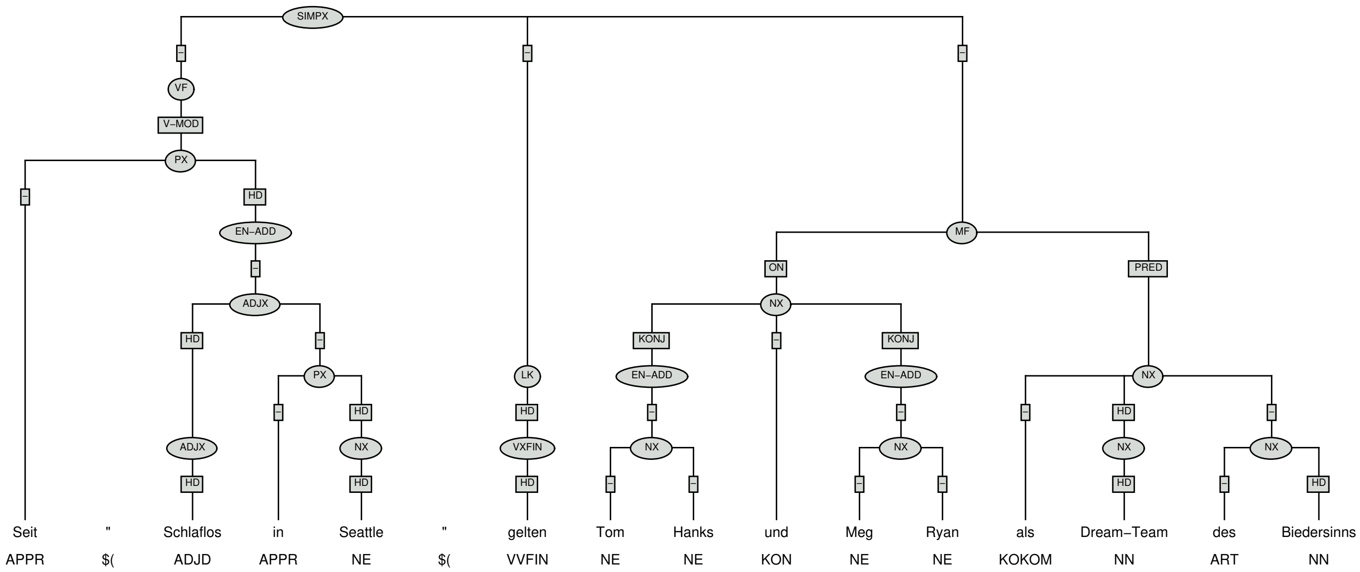

EN–ADD
NX
NX
HD
University
of
Illinois
NE
NE
NE


NX
HD
EN–ADD
NX
die
Tour
de
France
ART
NE
NE
NE

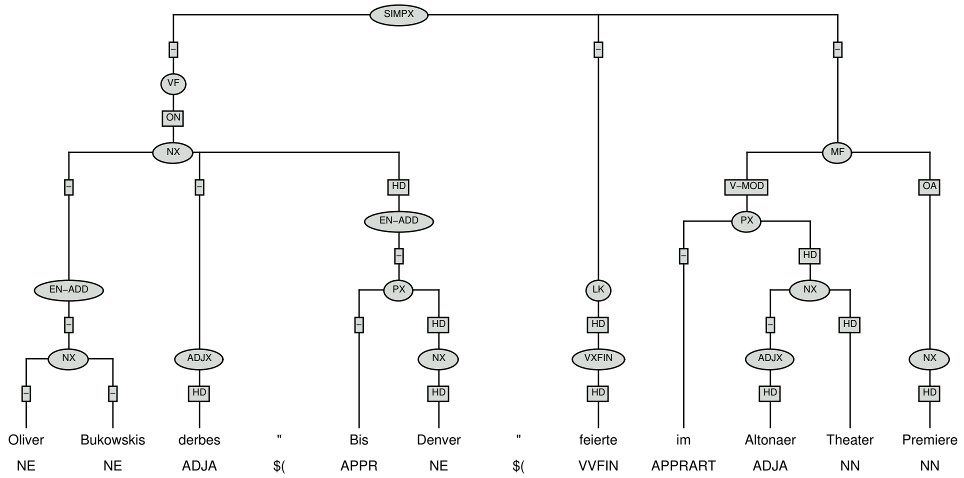
German Named Entities

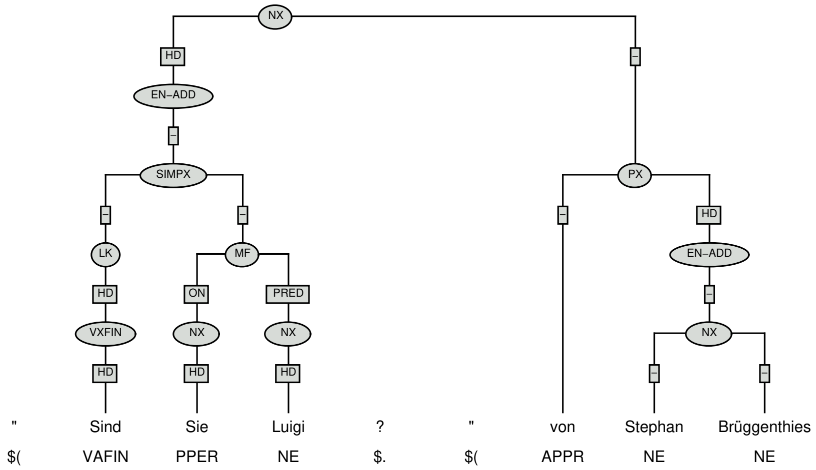
As mentioned above, German named entities are always tagged according to their distribution and annotated with their internal syntactic structure as noun phrases, prepositional phrases, clauses, etc. In order to indicate their status as named entities, the additional node EN-ADD is inserted between their mother node and the next higher annotation level. If two EN-ADD nodes are coordinated, their mother node is NX which represents the nominal status of EN-ADD. Here are some examples:

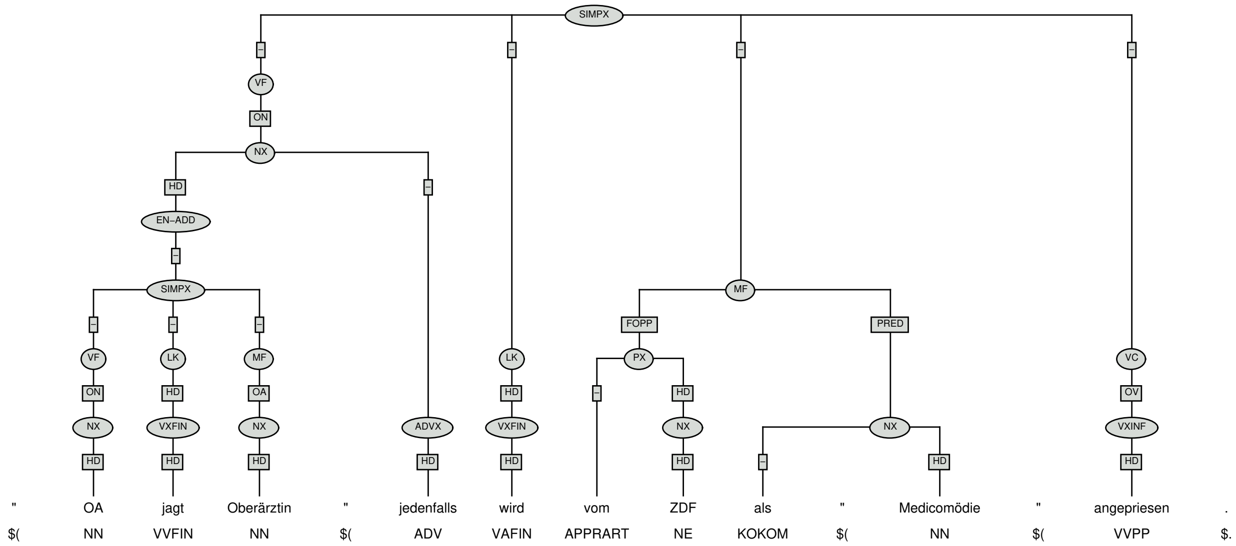




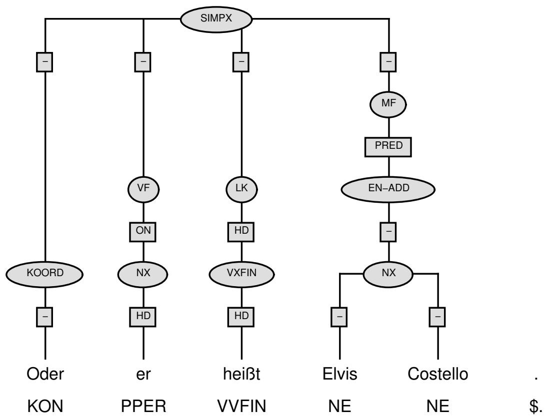
If EN-ADD has a premodifier, which can be an EN-ADD itself or a postmodifier, its mother node is also always NX:



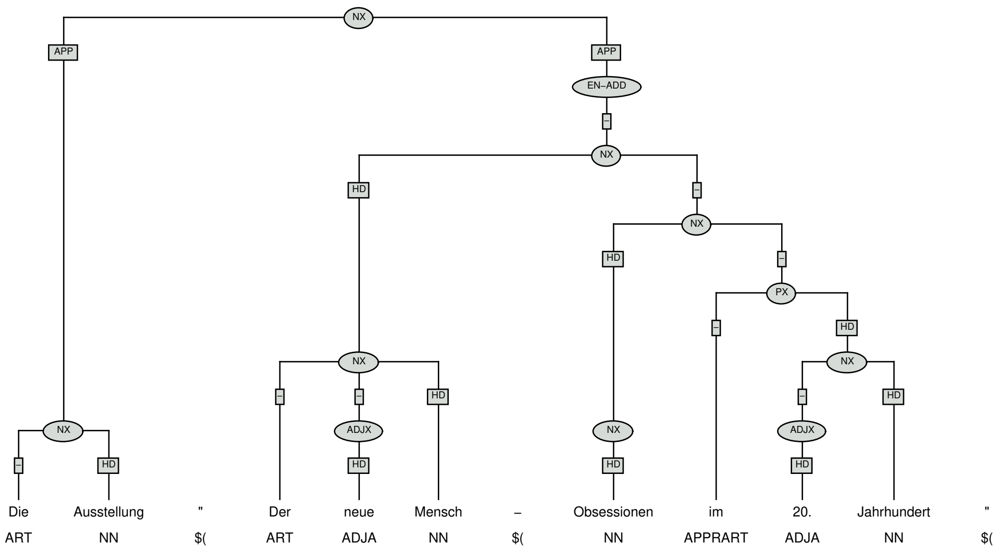




EN-ADDs which are directly attached to a field are assigned the grammatical function they have in the German sentence:

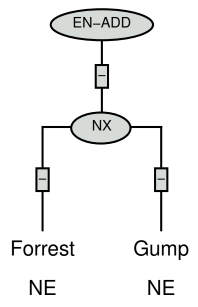


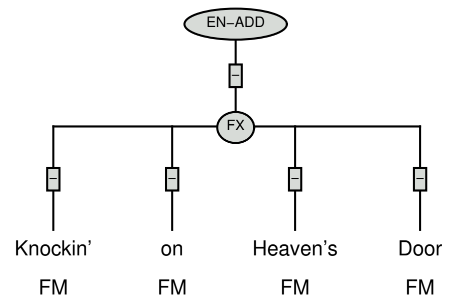
If named entities, e.g. a title, consist of two separate phrases and/or clauses, (e.g. title and subtitle), the first part is annotated as the head of the second part to express their dependency relation:

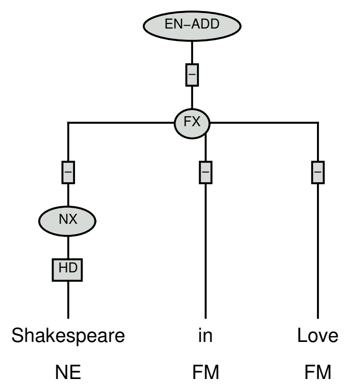


Foreign Language Named Entities

The syntactic annotation of foreign language named entities differs from the annotation of German named entities in the following aspects. As mentioned above, foreign language proper nouns are tagged as NE, while all other lexical entries of a foreign language are tagged as foreign language material (FM). A foreign language named entity which consists only of a proper noun, e.g. the title of a movie (*Forrest Gumpp*) is assigned an EN-ADD label. If a foreign language named entity consists of only FM tagged elements,these elements are directly attached on the same level without internal syntactic structure. Its mother node is marked as FX, e.g. *Knockin' on Heaven's Door*. If a foreign language named entity consists of NE as well as FM tagged elements, e.g. *Shakespeare (NE) in (FM) Love (FM)*, first the annotation stategy described in 4.2.4 is applied. Then, in a second step, the same strategy as for German named entities is applied: the insertion of the EN-ADD node.

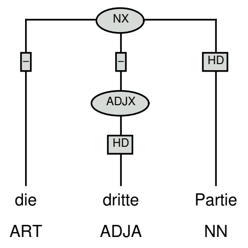


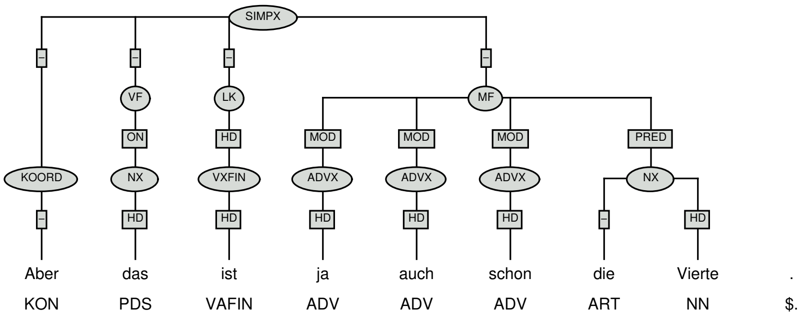




4.2.6 Ordinal Numbers

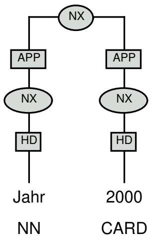
According to their distribution, ordinal numbers occur either as a premodifying attributive adjective (e.g. *die dritte (ADJA) Partie*) or as a head noun (e.g. *er ist der sechste (NN)*). In the first case, the premodifier is projected to an adjectival phrase, in the latter case it is projected to a noun phrase.

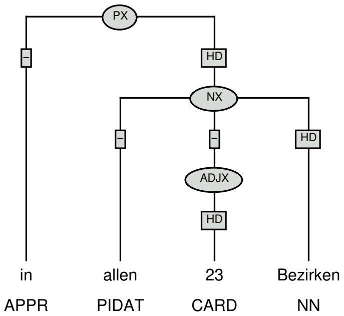


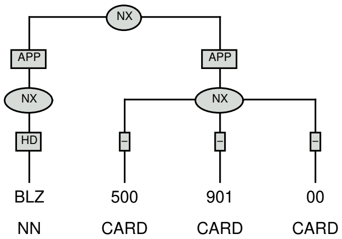


4.2.7 Cardinal Numbers

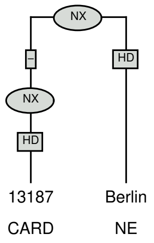
According to their syntactic function (nominal or adjectival), cardinal numbers, (CARD) are either projected to NX or ADJX. If their numerals are written separately or in groups, e.g. numbers of bank accounts, they are attached on the same level like proper names without internal head assignment.



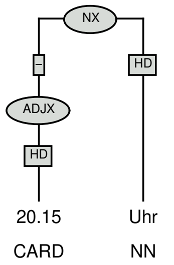


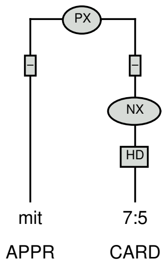


A premodifying cardinal number is nominal if it does not express a quantity like in the example above, but a characteristic of the following noun, e.g. the number of a zip code:



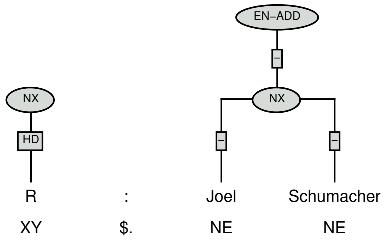
Complex time expressions or results of competitions are also treated as cardinal numbers:

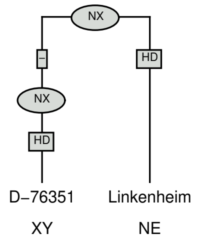


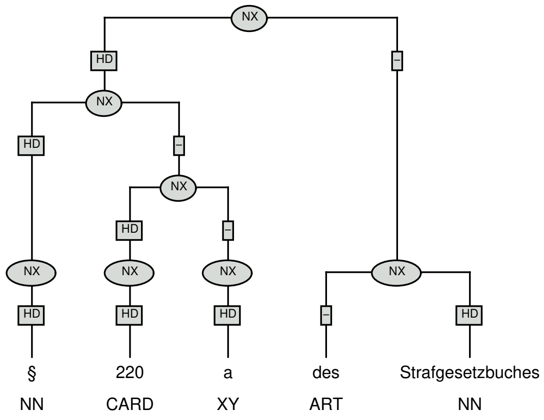


4.2.8 Letters and Non-Words

Letters and non-words are tagged as XY. They are projected to their phrase level and assigned the syntactic category to which they belong in the construction. Signs which represent a lexical element, e.g. the sign for paragraph, are tagged with the respective part-of-speech tag:







4.2.9 Expletive *es*

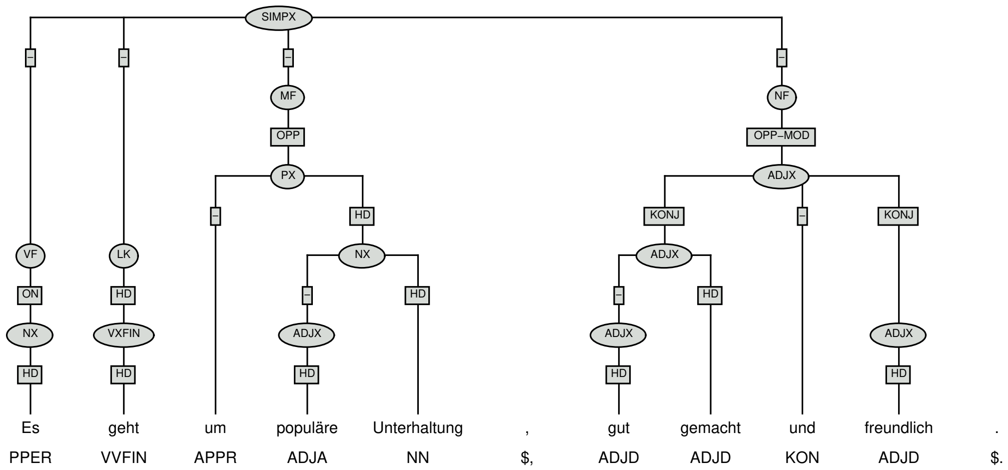
The initial field can be occupied by an obligatory element with purely morphosyntactic function: the expletive *es*. As the expletive *es* has no semantic content, it cannot be regarded as a proper argument of the verb, although its syntactic role is that of a purely formal grammatical subject. Two different 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. *es regnet*) or other verbs that can denote agentless events *es geht um populäre Unterhaltung*. It may also occur in MF (*morgen geht es um populäre Unterhaltung*). The important point is 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.

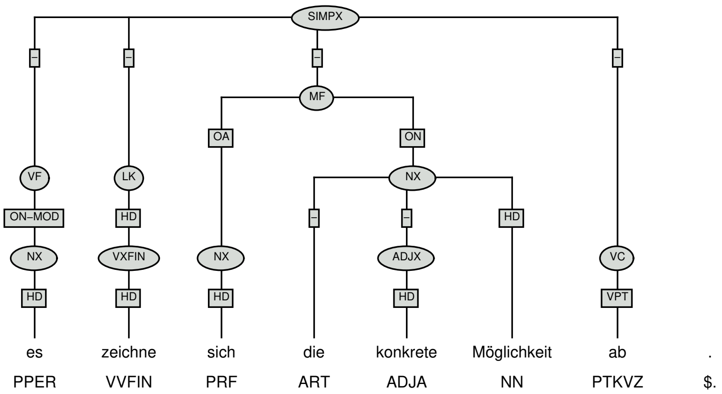
2. *es* as a subject or object expletive:
   The subject expletive can only occur in the first position (**Die konkrete Möglichkeit zeichnet es sich ab.*) while the subject of the sentence is a constituent in MF. There is no agreement relation between *es* and the finite verb but between the subject and the finite verb (e.g. *Es zeichnet sich die konkrete Möglichkeit ab.*).
   The object expletive occurs in MF while the object is realized as a sentential object (e.g. *Ich finde es gut, daß das Aktionsprogramm nun in der breiten Öffentlichkeit diskutiert wird.*)

In order to indicate the differences between *es* as a formal subject, a subject expletive, and an object expletive in the tree structure, different edge labels are used.

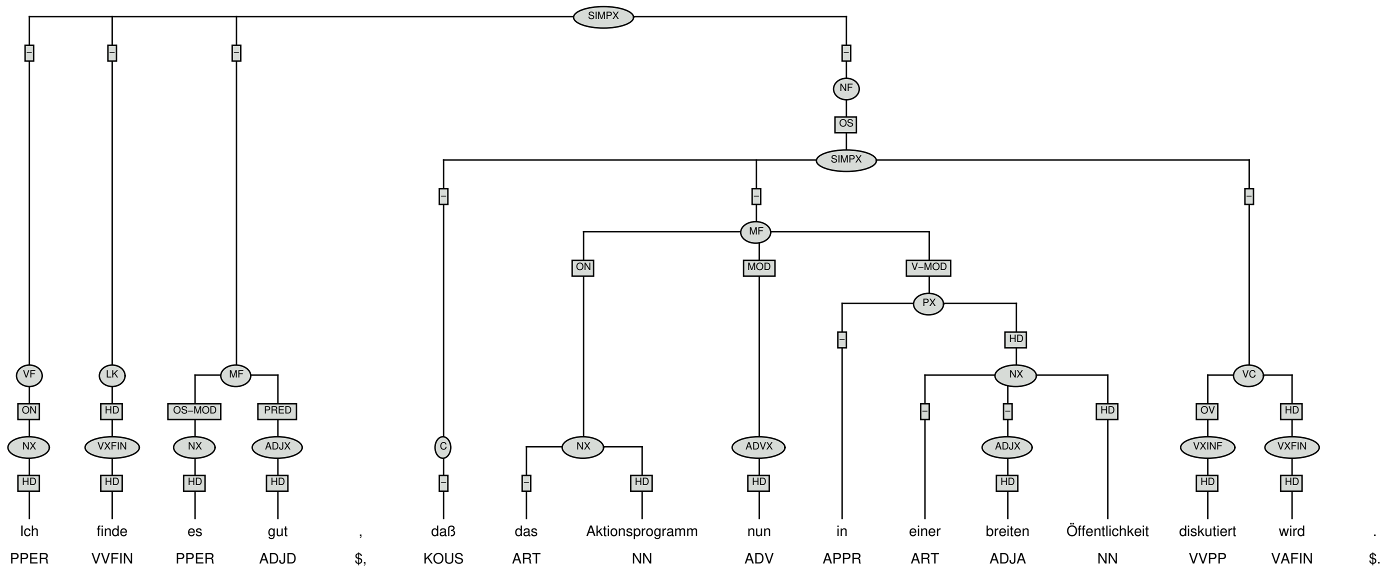
- The verb in the following example is lacking a subject. Therefore, the subject position is filled with the formal subject *es*, marked with ON as edge label.



- In the example below, the expletive *es* occupies the subject position but the subject of the sentence is in MF. Therefore, the expletive *es* is labeled as a constituent (ON-MOD) referring to the subject (ON).



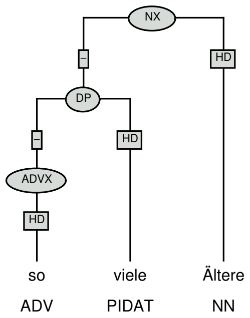
- If there is an expletive *es* in object position in MF referring to a sentential object (OS) in NF, *es* is labeled as (OS-MOD).

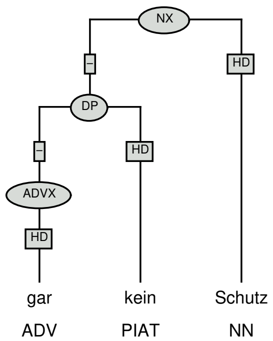


4.3 Determiner Phrases

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

In the case of *so viele Ältere*, the premodifying adverb *so* is attached to the indefinite pronoun *viele*. Together, they form a determiner phrase (DP), which is attached to the head noun *Ältere* on the same level:

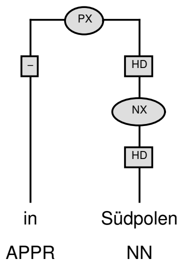




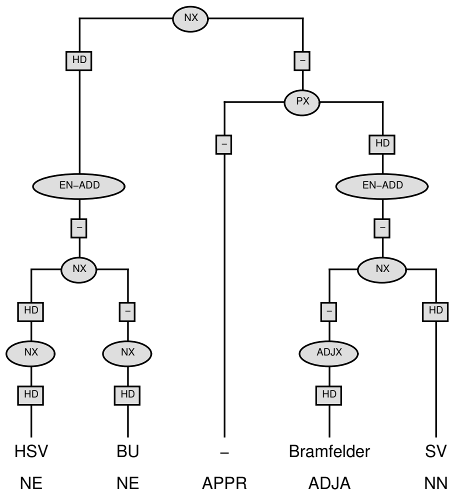
4.4 Prepositional Phrases

4.4.1 Prepositions

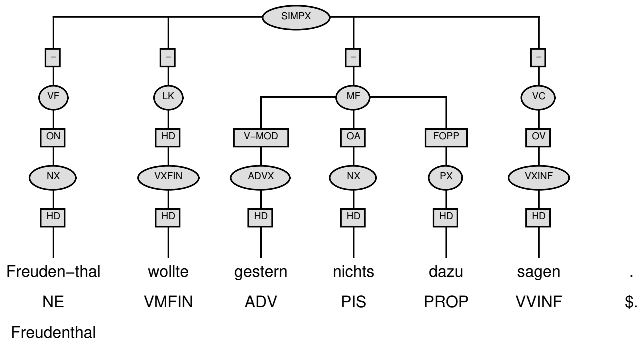
In order to facilitate the identification of dependencies between verbs and their nominal complements and adjuncts and in keeping with basic assumptions in Dependency Grammar, not the preposition itself but the complement in prepositional phrases is annotated as the head of the phrase:



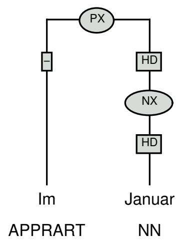
If the preposition is realized as a non-alphabetic sign, e.g. - (*bis, gegen*), this sign is tagged as APPR and annotated like a preposition:



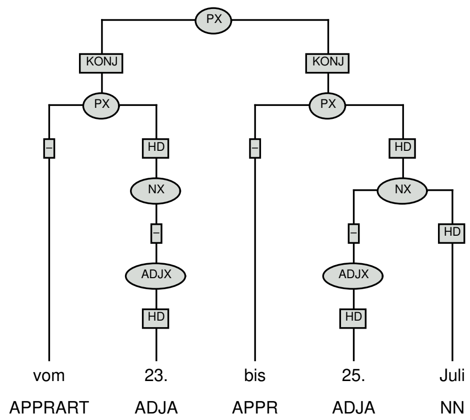
Since pronominal adverbs (PROP) are pronominal forms of a prepostional phrase, they are directly projected to PX without a preposition:

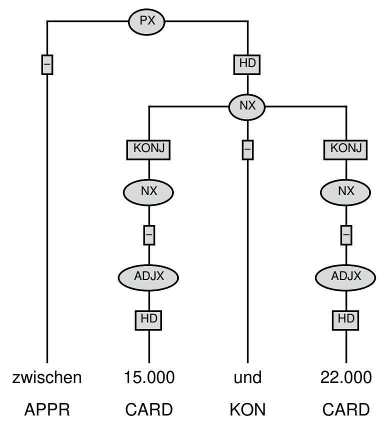


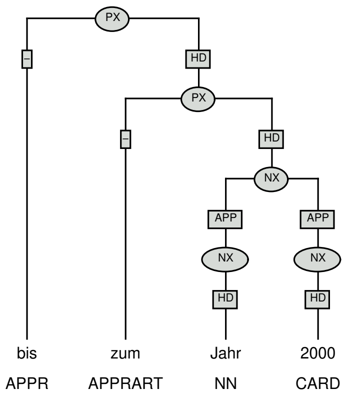
In German, there are so-called *Verschmelzungsformen*, i.e. merged forms of a preposition and a determiner, e.g. *in dem Januar* merges to *im Januar*. The merged form is assigned the part-of-speech tag APPRART and is annotated like a preposition:



Prepositional phrases expressing intervals, e.g. with *von/bis, von/bis zu* or *zwischen*, are annotated like coordinated phrases (cf. 6.4.1), i.e. without head assignment on the level of coordination, since the two phrases are considered to be conjuncts. If two prepositions follow each other (e.g. *bis zum*), the result is an embedded structure of a prepositional phrase taking another preposition.



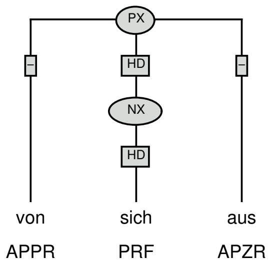




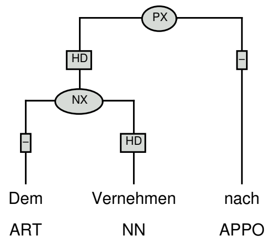
By contrast, intervals like *dritter bis fünfter November* are annotated as a coordinated attributive adjective within a simple noun phrase (cf. 6.4.1).

4.4.2 Circumpositions and Postpositions

Circumpositions are treated as ternary branching prepositional phrases. The circumposition on the left hand side is tagged as APPR and the circumposition on the right hand side as APZR:



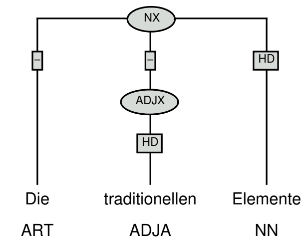
Postpositions are tagged as APPO. The complement of the postposition occurs on the left side and constitutes the head of the prepositional phrase:

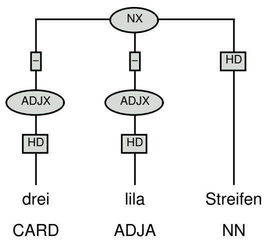


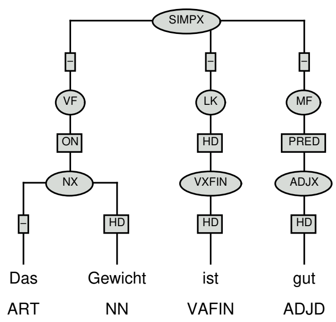
4.5 Adjectival Phrases

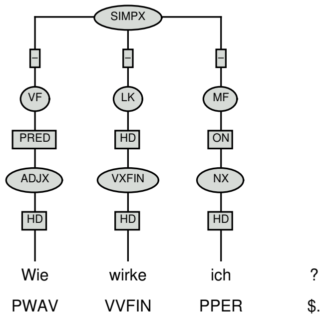
We distinguish between attributive and adverbial adjectives respectively predicative adjectives. Attibutive adjectives are tagged as ADJA (*die* **traditionellen** *Elemente*) or CARD (**20.15** *Uhr*), whereas adverbial or predicative adjectives are tagged as ADJD (*das Gewicht ist* **gut**) or PWAV (**wie** *wirke ich*).

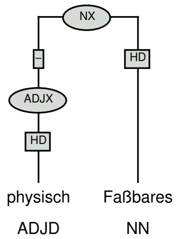
In general, German adjectives are inflected when they are an attribute of a noun. They are not inflected either when they function as a predicative adjective or a premodifier of an adjective or an adverb or when they belong to a small class of noninflected adjectives, e.g. some ancient form or some adjectives denoting a colour (*mit einer rosa Karte*). All adjectives have to be projected to their phrase level before they are attached to another phrase or to a field.

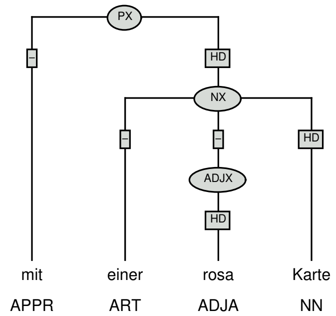




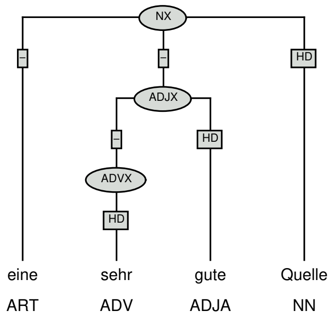


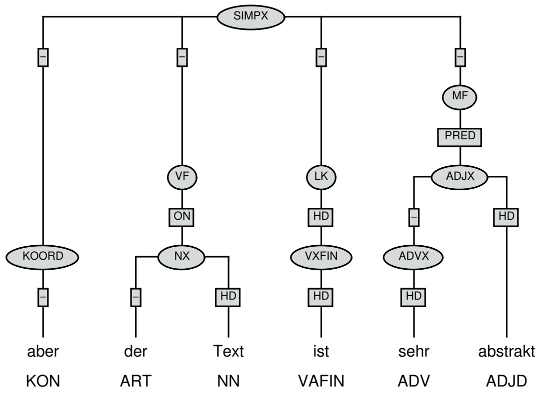


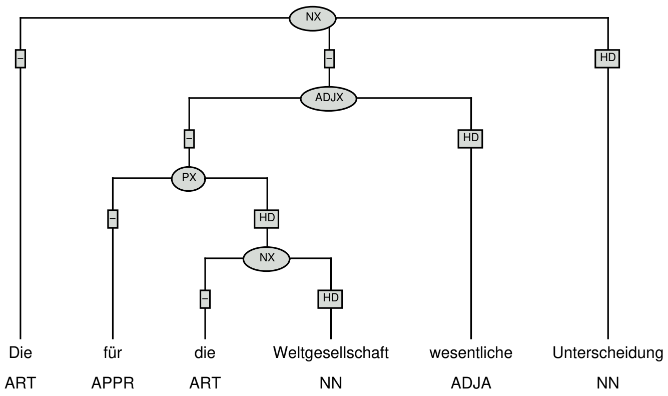




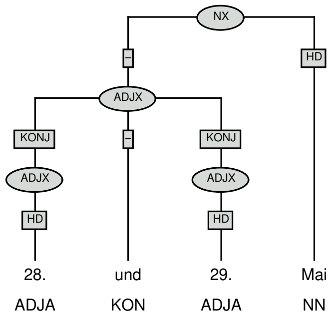
Whenever an adjective is modified by another modifier, the same annotation strategy as for noun phrases is applied, i.e., the modifier is directly attached to the adjectival phrase. The adjectival phrase as a whole is the premodifier of the noun phrase. For instance:

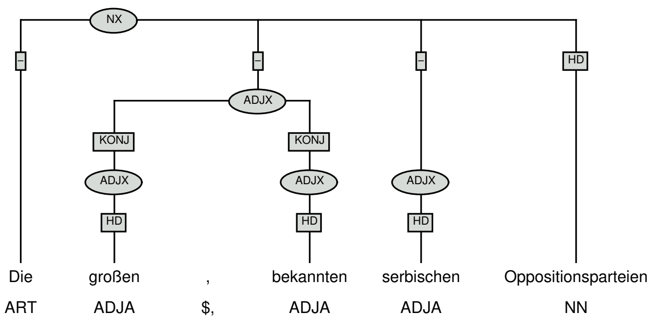


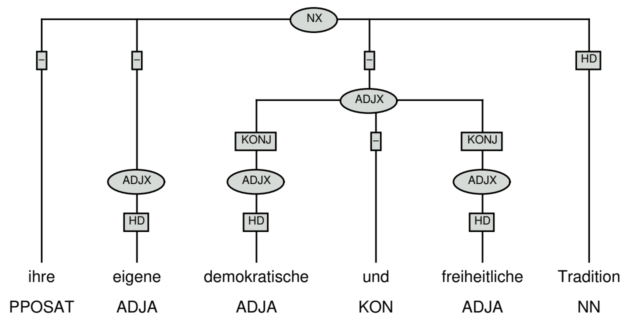




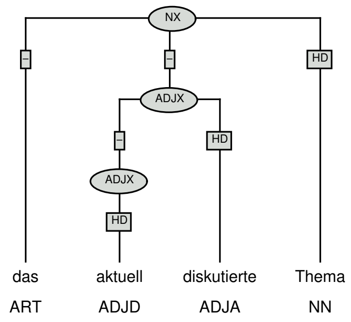
Premodifying adjectives may occur in a linear order and/or as a coordination (cf. 6.4.1) of attributive adjectives:

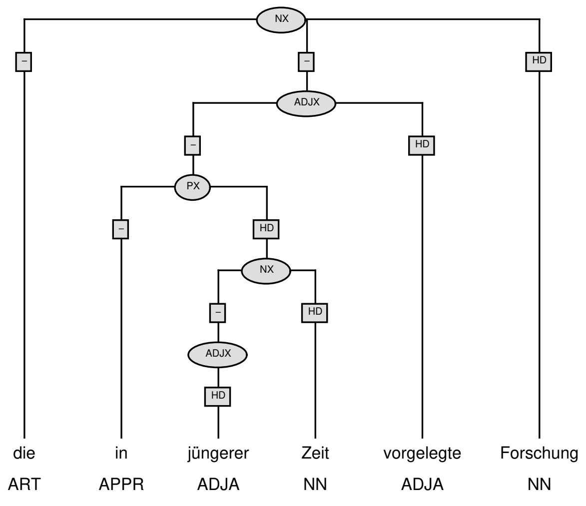




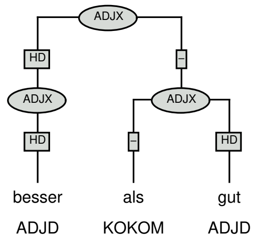


If the premodifying adjective is deverbal the adjectival phrase can be of any complexity. In this case, the adjectival phrase has its own internal dependency structure. All elements which depend on the adjective are annotated as its premodifiers. Deverbal adjectives are either attributive or adverbial respectively predicative and occur as the present participle or past participle form of a verb.



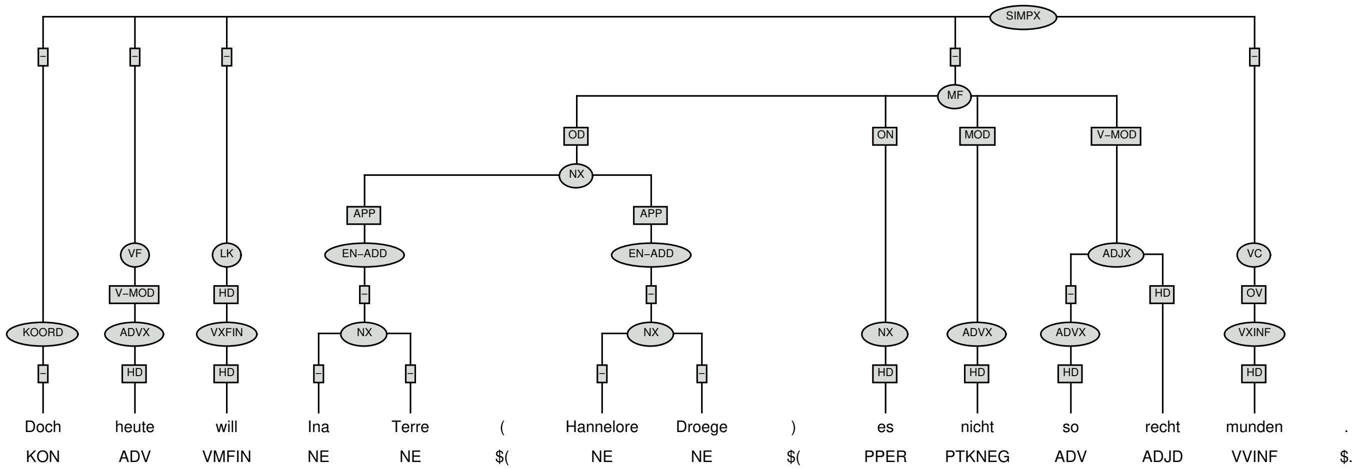


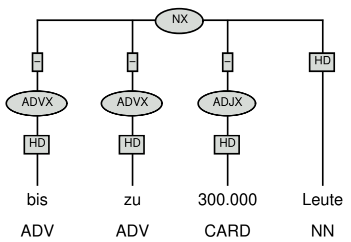
In the following example, postmodification of an adjectival phrases is shown:

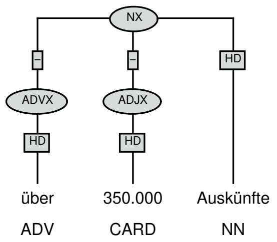


4.6 Adverbial Phrases

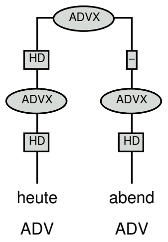
Besides adverbials also negation particles (PTKNEG) project to an adverbial phrase. They either occur as premodifiers[1] or postmodifiers or they are directly attached to a field.

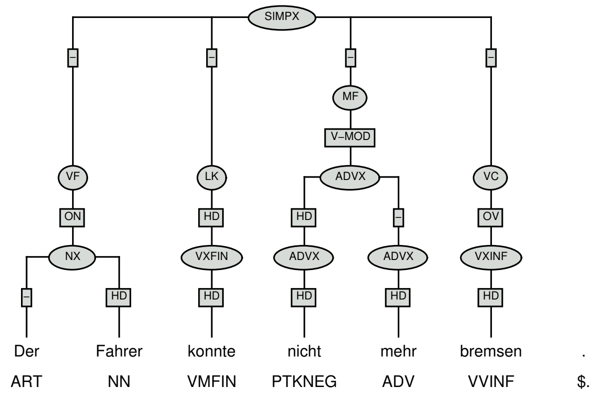






[1] *bis zu, über* are considered to be ADV rather than APPR because of their semantic meaning.





4.7 Verb Phrases

Whereas finite verb phrases are labeled VXFIN, non-finite verb phrases are labeled VX-INF.

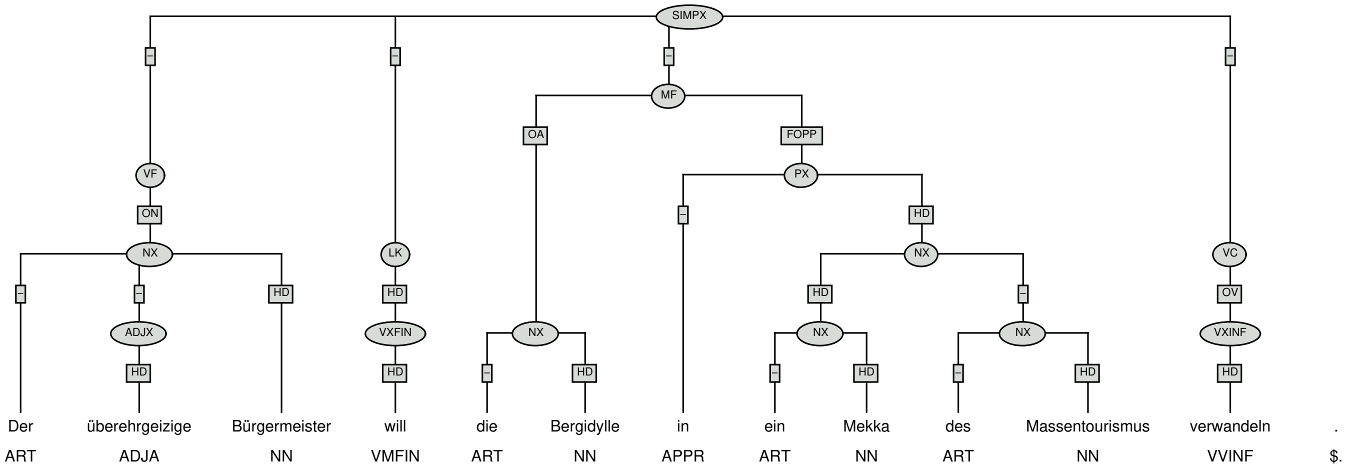
Since infinitives and past participles share certain properties (e.g. exchangeability in *Man hat nur noch das eigene Herz schlagen hören/gehört.*), they are assumed to carry the same phrase label (VXINF). The finite verb in LK as well as the infinite verbs in VC are always projected to their phrase level. All verb phrases of the verb complex are attached on the same level to form the verb complex. In order to follow the *flat clustering principle*, no internal hierarchy of the verb complex is annotated.

4.7.1 Head of a Sentence and Verb Complex

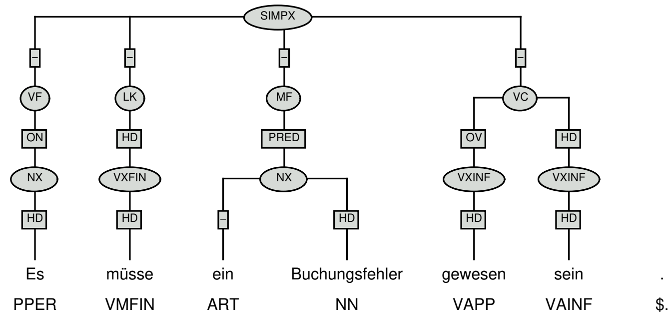
The finite verb which can either appear in LK (verb-first clauses and verb-second clauses) or in VC (verb-final clauses), is always the head of the entire sentence. Non-finite verbal elements belong to VC. If the finite verb is located in LK and if there is more than one non-finite element in VC, the 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. The head of VC selects the verbal object OV. This verbal object may select a another verbal object OV, and so on. To denote the dependency relations between verbal objects within the verb complex, we attach a secondary edge label REFVC between their phrase nodes.

4.7.2 Verb Complexes in Verb-second and Verb-final Clauses

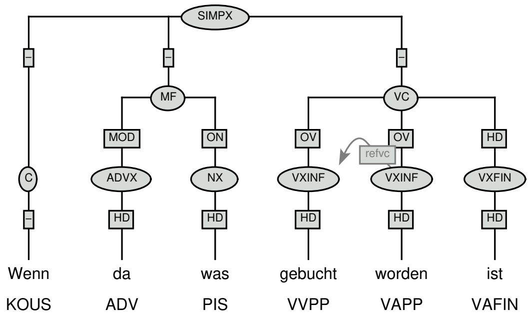
The following example shows a verb-second clause with the head of the sentence in LK and a verb complex consisting of a single non-finite element.

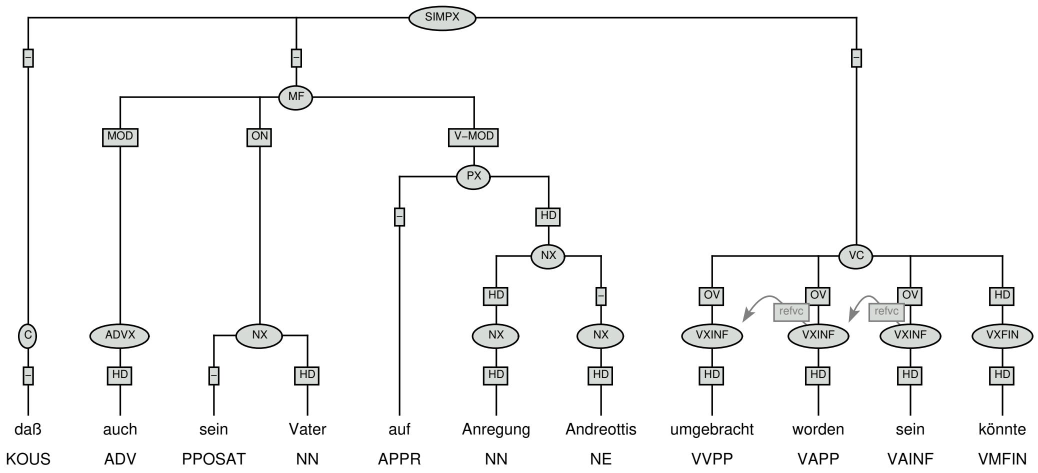


If the verb complex consists of more than one immediate daughter, the one that is selected by the finite verb is the head of VC.

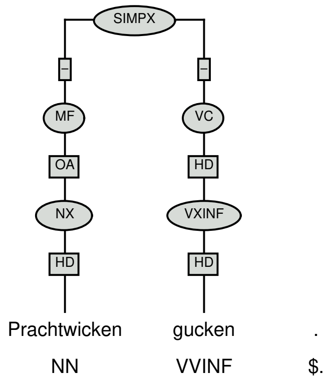


The following trees show verb complexes with two or more verbal objects. The secondary edge label REFVC is pointing from the selecting OV to the depending OV.



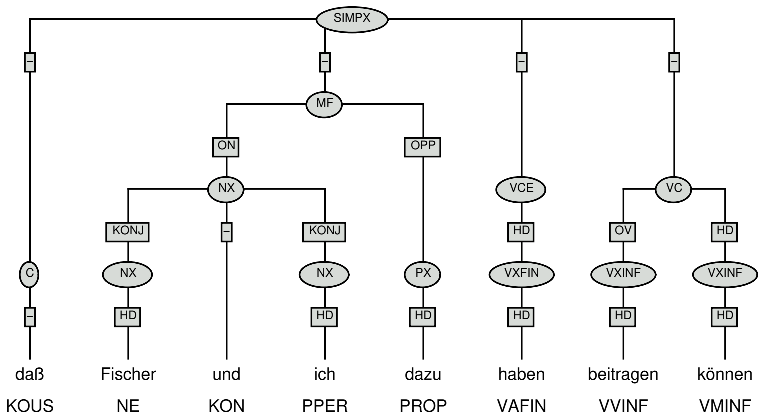


If there is no finite verb at all, the rightmost element of the verb complex (if there is more than one element) is annotated as the head of the sentence. This often occurs in headlines (cf. 5.2 and 7.4).

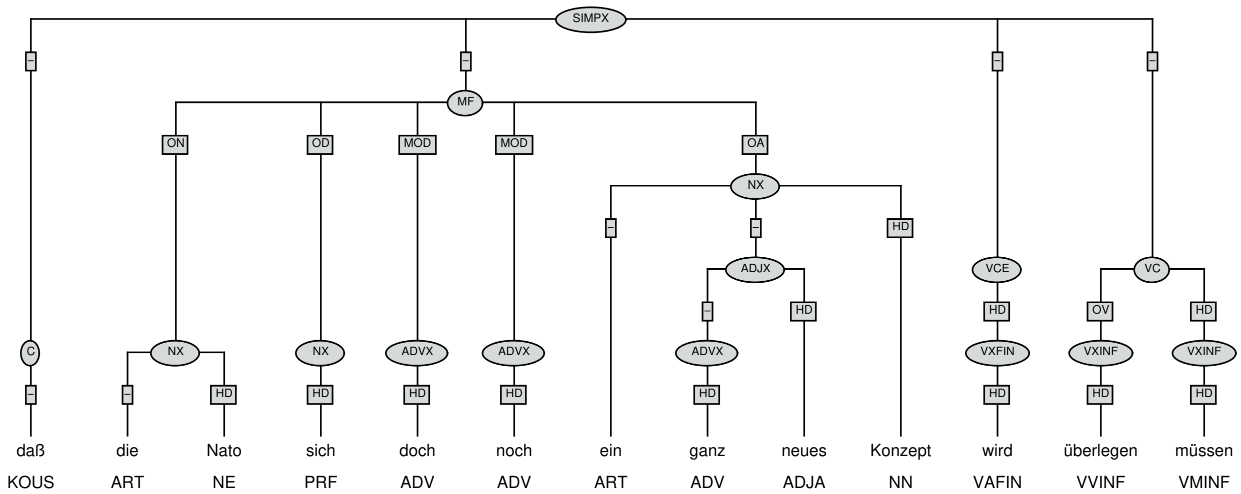


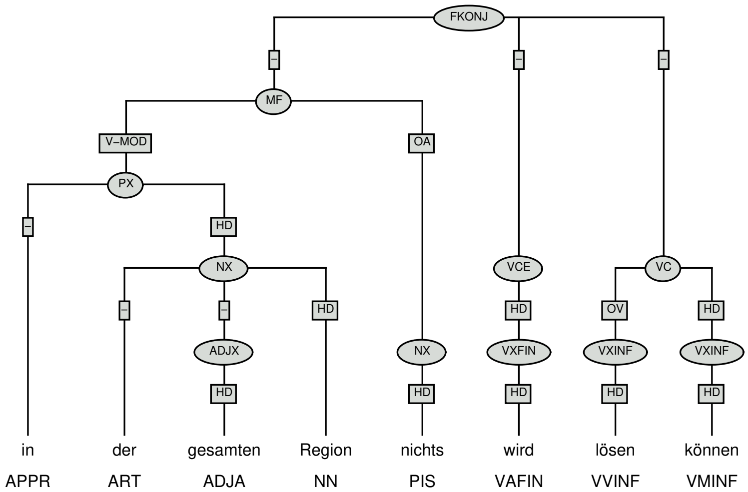
4.7.3 Ersatzinfinitiv Constructions

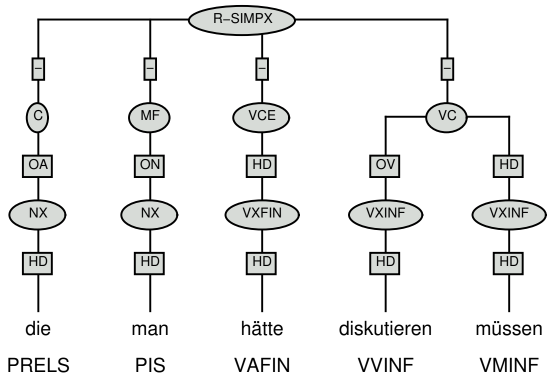
In order to indicate *Ersatzinfinitiv* constructions, two specific field node labels are introduced. VCE is the node label for the part of the verb complex consisting of the finite verb which subcategorizes for the *Ersatzinfinitiv*. MFE is the node label for the possible second part of MF between VCE and and the second part of the verb complex VC (e.g. *[C dass] [MF sie] [VCE hätten] [MFE alles] [VC diskutieren müssen]*).



In the example below, the finite verb precedes the non-finite verbs although *müssen* is no *Ersatzinfinitiv*. Since its position corresponds to the position of the finite verb in real *Ersatzinfinitiv* constructions and here also a second middle field is possible, we follow the same annotation strategy.

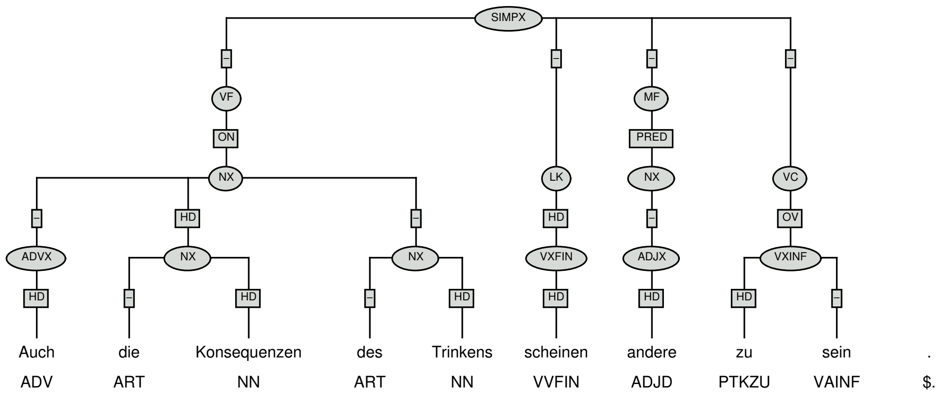


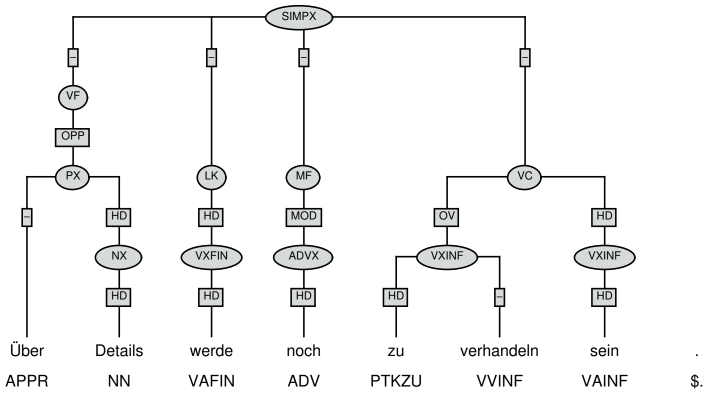




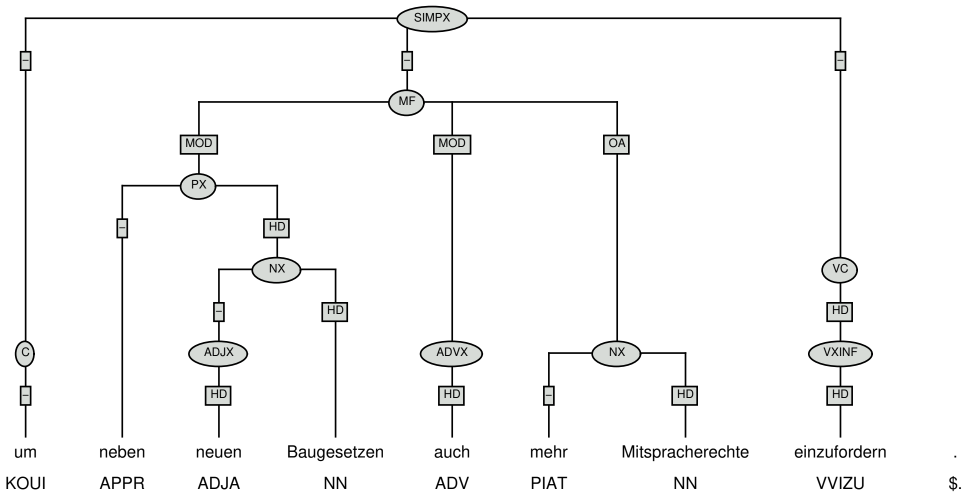
4.7.4 Infinitives with *zu*

For infinitives with *zu*, *zu* is considered the head, the infinitive is considered the complement since *zu* determines the infiniteness of the verb on the right hand side. Like other infinitives, they occur in the verb complex:

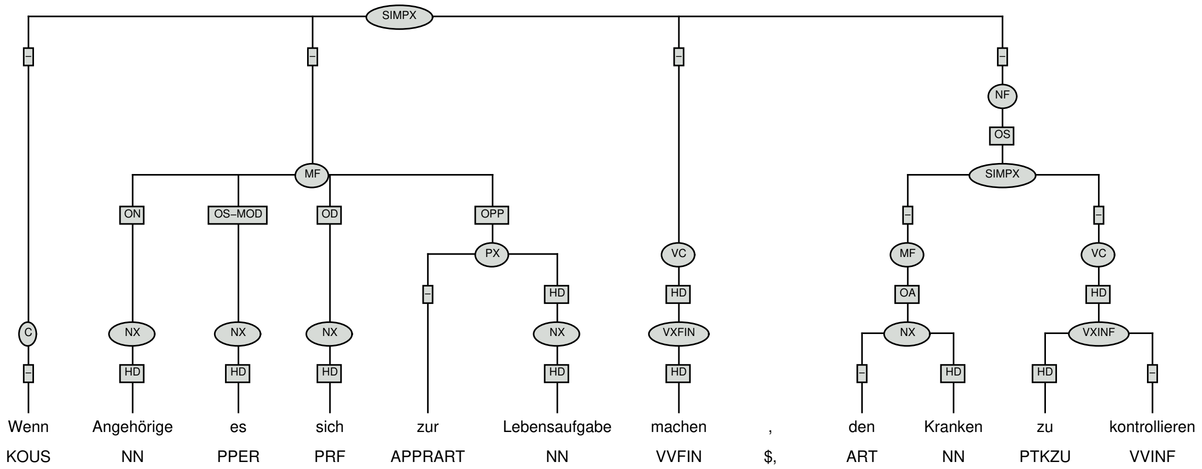


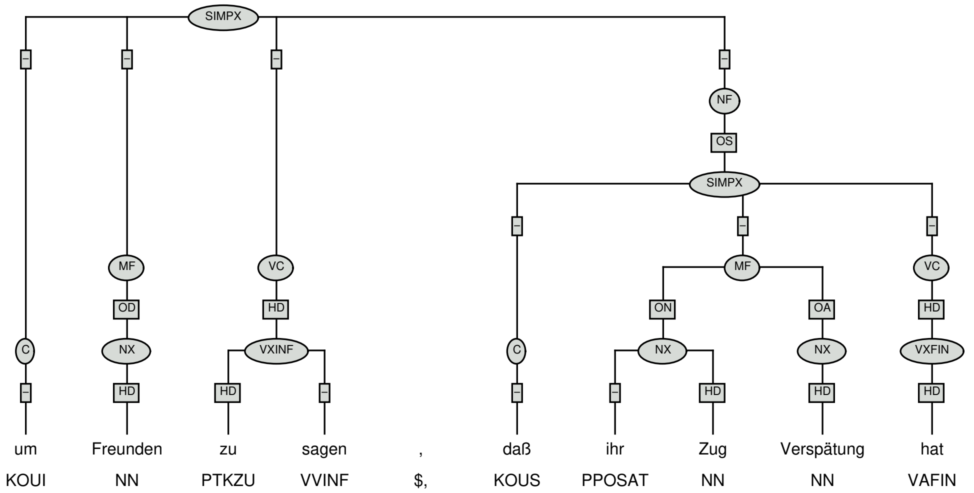


The infinitive with *zu* can also be realized as an infix of the verb. In this case, the verb is tagged as VVIZU and projected to VXINF with the grammatical function OV:

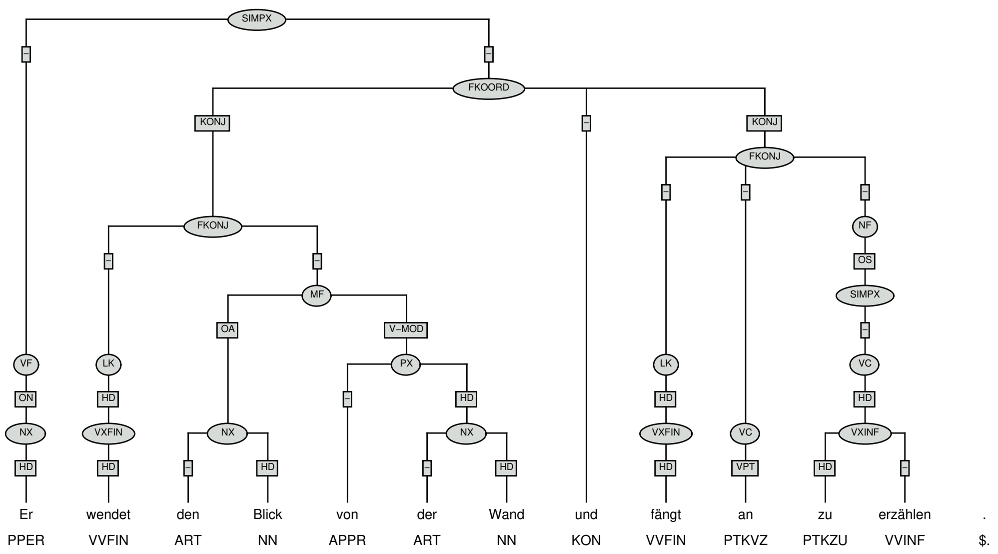


Besides the examples above, the infinitive with *zu* occurs in optional (in most cases with *um zu*) and obligatory infinitive clauses.





Infinitive clauses can consist of only one verb complex:



4.7.5 Coherency and Incoherency of Verbal Constructions

The notion of coherency attributed to Bech (1955-57) covers the relation of dependency between adjacent verbal elements, i.e. the relation of subcategorization between a verb and a non-finite verbal complement. Kiss (1995) calls this relation *infinitive Komplementation* (non-finite complementation). Bech (1955-57) distinguishes between three different modi of obligatory and optional coherency:

1. verbs constructing coherently and incoherently, e.g. *versprechen, versuchen*

   coherent, extraposition possible:
   a. [*wie er mit kritischen politischen Gegenpositionen umzugehen versteht*]
   incoherent, extraposition:
   b. [*wie er versteht,*][*mit kritischen politischen Gegenpositionen umzugehen*]

2. verbs constructing only coherently, e.g. *wollen, möchten*

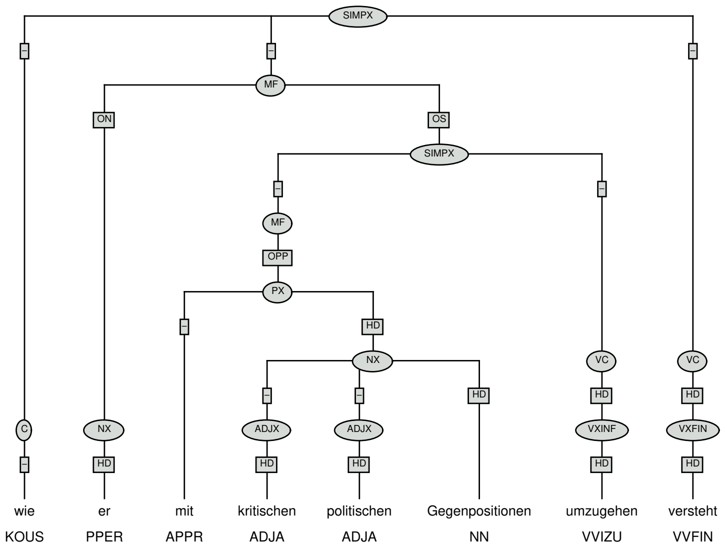
   coherent, no extrapostion possible:
   a. [*wie er mit kritischen politischen Gegenpositionen umgehen will*]
   b.*[*wie er will mit kritischen politischen Gegenpositionen umgehen*]

3. verbs constructing only incoherently, e.g. *überreden, überzeugen*

   incoherent, extraposition obligatory:
   a. [*wie er sie überredet,*][*mit kritischen politischen Gegenpositionen umzugehen*]
   b.*[*wie er sie* [*mit kritischen politischen Gegenpositionen umzugehen*] *überredet*]

Coherent and incoherent constructions of verbs are annotated differently. In case of coherency, the verbal complement is part of the verb complex. In the clause *wie er mit kritischen politischen Gegenpositionen umzugehen versteht*, for instance, the infinitive with *zu* is the verbal object of the finite verb. While in case of incoherency, the verbal complement is annotated as a sentential complement, i.e., *mit kritischen politischen Gegenpositionen umzugehen* in the clause *wie er sie überredet, mit kritischen politischen Gegenpositionen umzugehen* is a sentential object in NF.

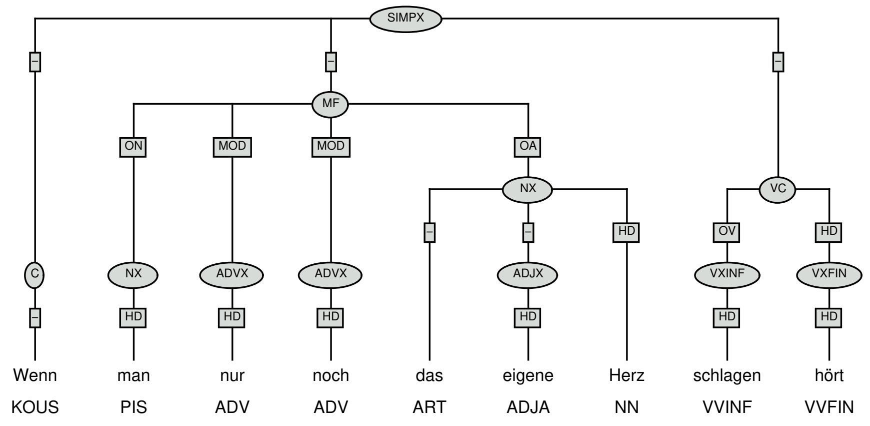
We define that a construction is incoherent, if extraposition in NF is possible. This means, whenever it is possible to shift the infinitive complement together with a constituent of MF, which it subcategorizes for, into NF, these elements are annotated as sentential objects. Therefore, the coherent example above (*wie er mit kritischen politischen Gegenpositionen umzugehen versteht*) is annotated with a sentential object in MF since extraposition is possible (cf. the incoherent example 1.b.).



If a complement of the verb in the sentential object is located out of the sentence boundaries, e.g. in the C-field, the secondary edge label REFCONTR gives additional information about the dependency relation (cf. 3.4.6).

4.7.6 AcI Constructions

AcI (accusativus cum infinitivo) verbs are a small group of verba sentiendi (e.g. *sehen, hören, fühlen, spüren*) which subcategorize for an accusative and an infinitive. The infinitive itself subcategorizes for complements in respect to its valency but its subject position is occupied by an accusative which is the direct object of the AcI verb. Like in all other infinitive constructions the direct subject is missing in the AcI. Since AcI constructions are coherent infinitive constructions in which extraposition is not possible (cf. (Eisenberg 1999-2001), p.355), the AcI is not annotated as a sentential object (* *wenn man nur noch hört das eigene Herz schlagen*). The infinitive as the verbal object of the AcI verb is located in the verb complex and the accusative is realized as OA in MF.



4.7.7 Imperatives

Imparatives have only one singular and one plural form and are not inflected concerning the grammatical category person. Their form corresponds to second person singular and plural verbs which are tagged as VVIMP or VAIMP.

Imperatives can be identified, for instance, by different morphological verb forms:

Warte du mal!

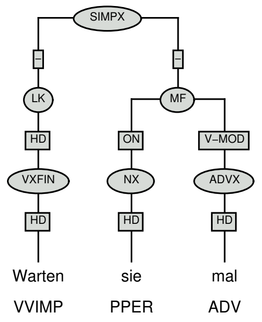
instead of

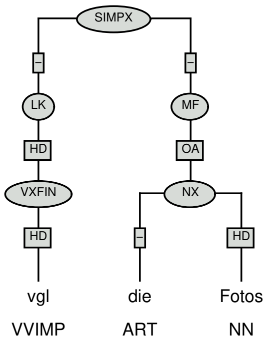
Wartest du mal?

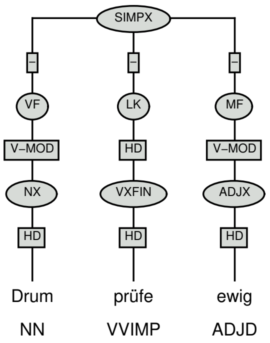
or by the co-occurrence of particles that exclude a non-imperative meaning of the sentence:

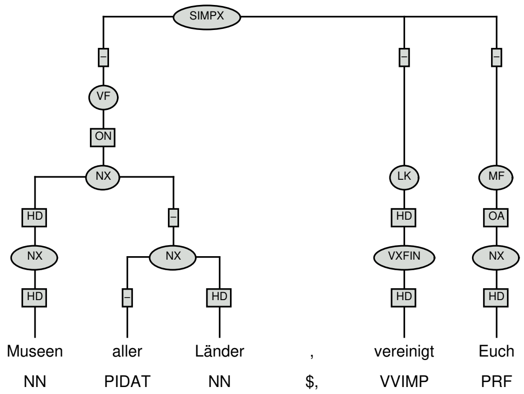
Warten Sie mal bitte!

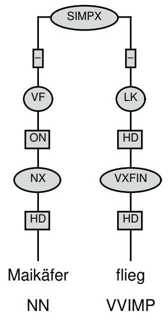
Imperatives usually occur as verb-initial but also as verb-second clauses. Often, they are lacking the subject, but the addressed person can also be mentioned to stress the utterance:




SIMPX
LK
MF
HD
OA
VXFIN
NX
HD
HD
vgl
die
Fotos
VVIMP
ART
NN

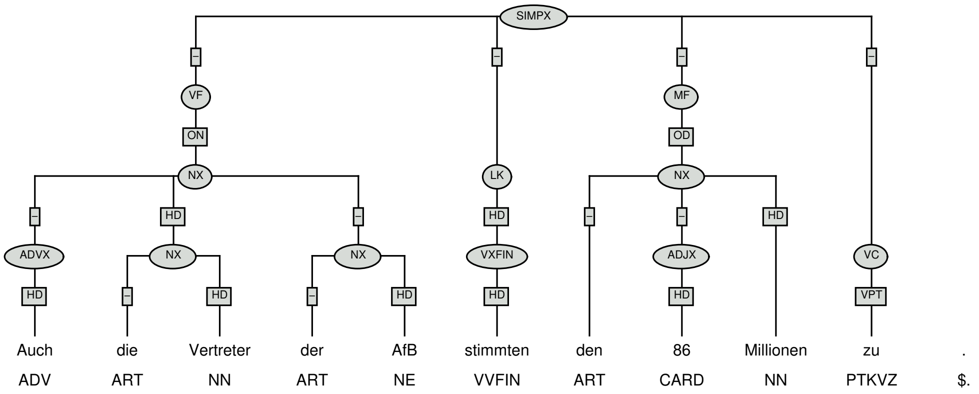

SIMPX
VF
LK
MF
V–MOD
HD
V–MOD
NX
VXFIN
ADJX
HD
HD
HD
Drum
prüfe
ewig
NN
VVIMP
ADJD


SIMPX
VF
ON
NX
LK
MF
HD
HD
OA
NX
NX
VXFIN
NX
HD
HD
HD
HD
Museen
aller
Länder
,
vereinigt
Euch
NN
PIDAT
NN
$,
VVIMP
PRF

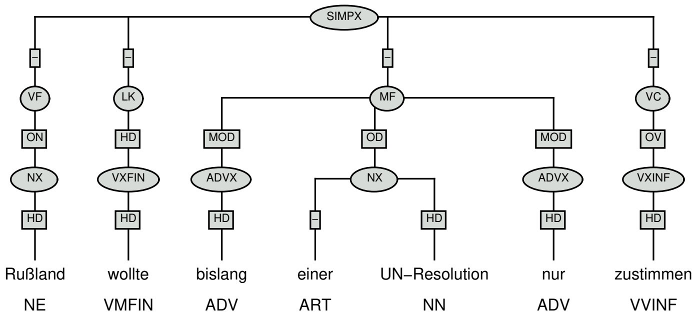

SIMPX
VF
LK
ON
HD
NX
VXFIN
HD
HD
Maikäfer
flieg
NN
VVIMP

4.7.8 Particle Verbs

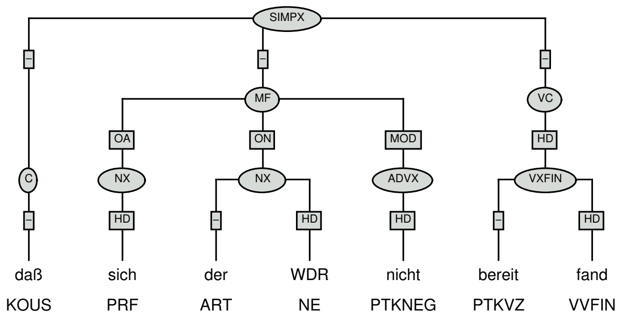
Separable verb particles are tagged as PTKVZ and annotated with the edge label VPT:



In verb-final clauses, the particle verb occurs unseparated within the verb complex:



In some cases, the particle verb precedes the infinite verb in the verb complex. It is annoted as part of the verb complex, directly attached to the VXINF node.



4.7.9 Verbs with Predicate

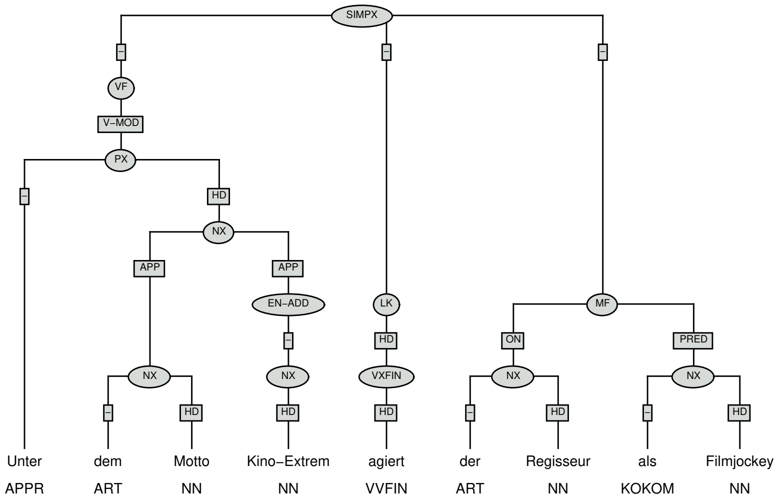
Typically, the complement type PRED (predicate) occurs with verbs like *sein, haben, scheinen, aussehen, sich anhören, klingen* etc. PRED is annotated, if the following conditions apply:

- if it is not possible to determine 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 *Die Ursache war unklar. Die Ursache* is characterized by the property of being unclear)

- many PRED verbs are raising-verbs (subject without theta-role)

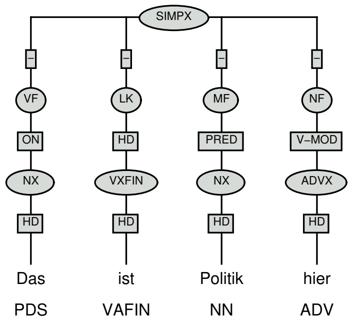
- if *als*-phrases are selected by the verb they are labeled as PRED (e.g. *Unter dem Motto Kino-Extrem agiert der Regisseur als Filmjockey.*)

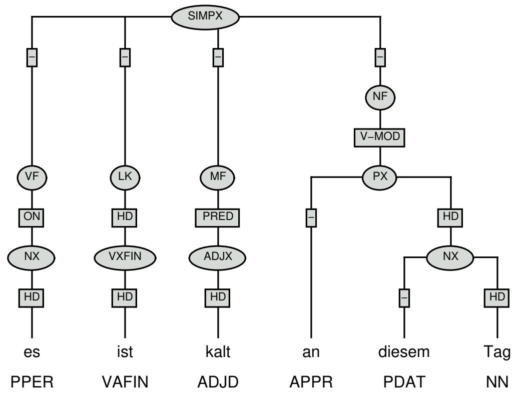


Some examples for verbs that take predicates: *recht sein, recht haben, leid tun, frei sein, fertig sein, sich gut/schlecht treffen, gut/schlecht finden* etc.

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

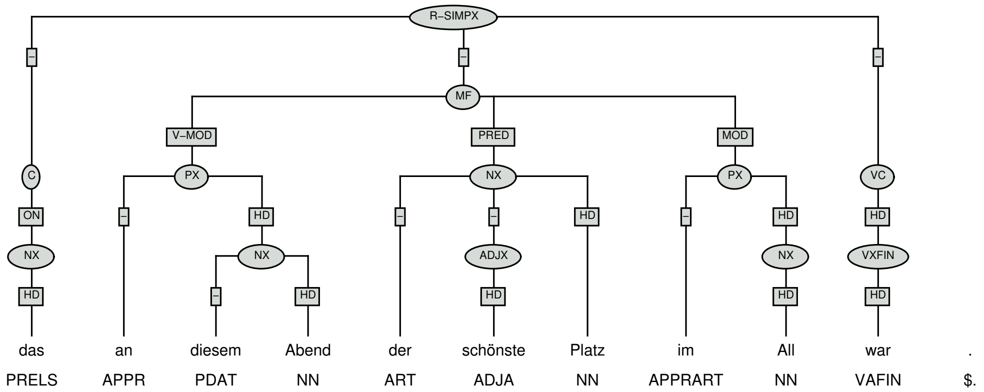
With respect to topological fields, note that PRED usually marks the border between MF and NF, i.e., whatever constituent occurs on the right hand side of PRED belongs to NF. In general, this constituent is an adjunct which PRED does not subcategorize for:



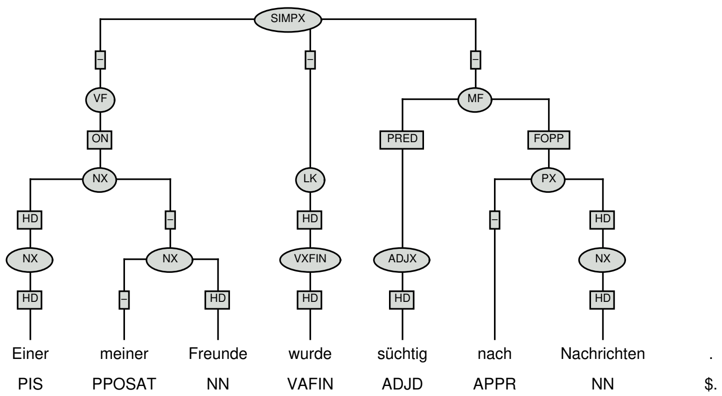


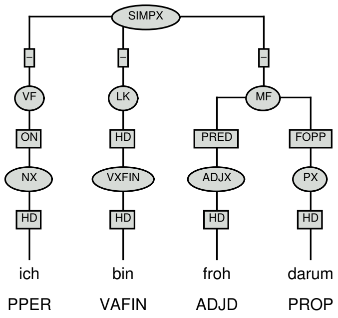
But there are exceptions in which PRED does not necessarily constitute the border between MF and NF:

- Another constituent may occur between PRED and VC, for instance, if an ambiguous modifier follows PRED.

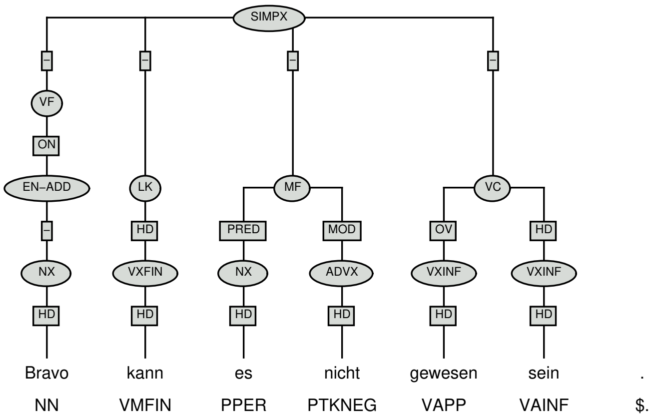


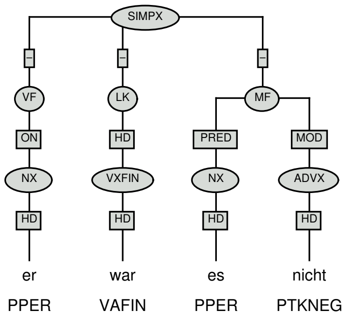
- PRED subcategorizes for the constituent that follows it. Complements of PREDs are always attached to a field since they are assigned a grammatical function within the sentence structure (cf. 8.1):





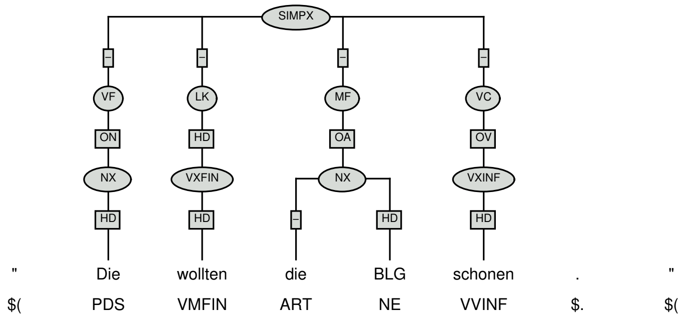
- Because of the word order rule that pronouns in MF have to precede other constituents, PRED might not be the last element in MF if it is a pronoun:

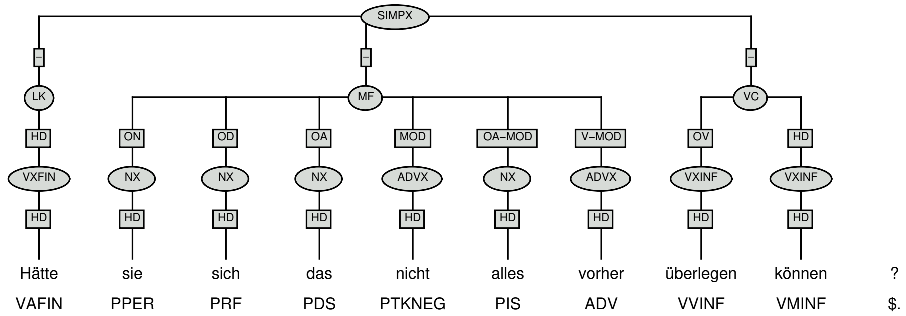


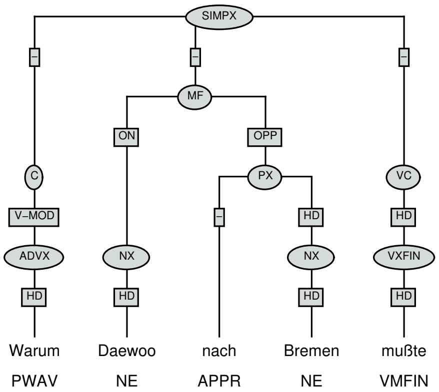


4.7.10 Modal Verbs

Modal verbs are always tagged as VMFIN or VMINF regardless of their use as an auxiliary or a main verb. If a modal verb functions as an auxiliary verb, it is projected like any other auxiliary verb. If a modal verb is the main verb of a sentences, verbal modifiers refer to the modal verb in the same way as they refer to other main verbs:


SIMPX
VF LK MF VC
ON HD OA OV
NX VXFIN NX VXINF
HD HD HD HD
" Die wollten die BLG schonen . "
$(PDS VMFIN ART NE VVINF $. $(


SIMPX
LK MF VC
HD ON OD OA MOD OA–MOD V–MOD OV HD
VXFIN NX NX NX ADVX NX ADVX VXINF VXINF
HD HD HD HD HD HD HD HD HD
Hätte sie sich das nicht alles vorher überlegen können ?
VAFIN PPER PRF PDS PTKNEG PIS ADV VVINF VMINF $.


SIMPX
MF
ON OPP
C PX VC
V–MOD HD HD
ADVX NX NX VXFIN
HD HD HD HD
Warum Daewoo nach Bremen mußte
PWAV NE APPR NE VMFIN

Chapter 5

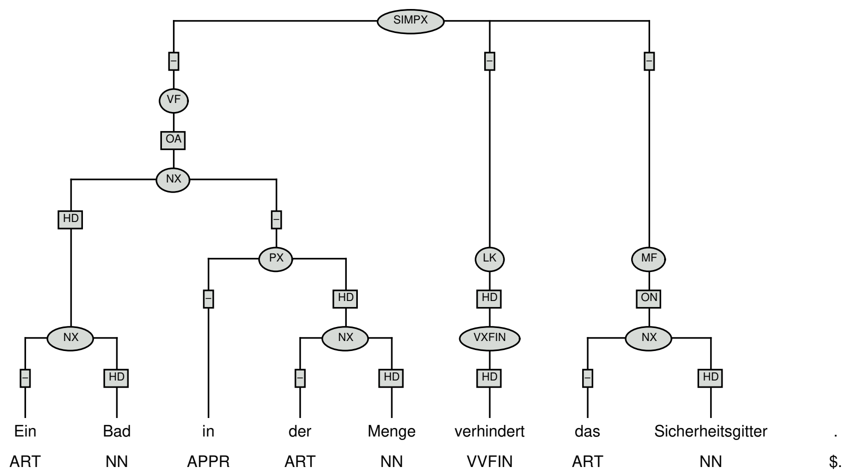
Attachment Principles for Phrases

5.1 Attachment to Fields

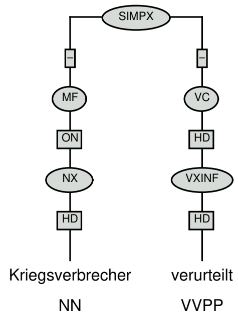
Phrases are attached to the topological field in which they occur. Their edge labels denote their grammatical function within the sentence structure. In LK and VC there can only occur verb forms, separable verbal prefixes, or infinitive particles. LK and VC mark the beginning and the end of MF (cf. 3.2).

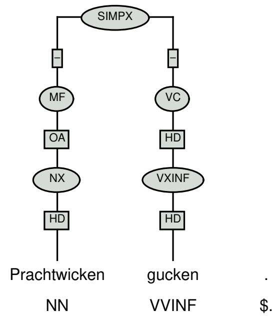
5.2 Attachment of Ambiguous Complements

The partially free word order and the morphological properties of German can cause ambiguity concerning the grammatical function of a constituent. In the following example, the syntactic structure does not give any information about case assignment. Both noun phrases can be identified as ON or OA:



Headlines like the following are lacking the finite verb. Therefore, in the first example it cannot be decided if it is an active or a passive construction, i.e., if the noun phrase is ON or OA. The second example is an active construction, but again the noun phrase can be both, ON or OA:





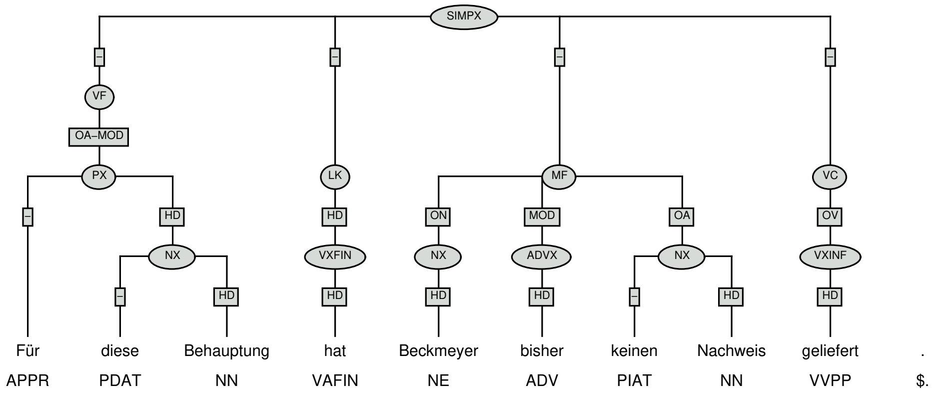
Since we do not assign specific edge labels for ambiguous complements, we formulate the following *preference principle for case assignment*:

Preference principle for case assignment:
If case assignment is ambiguous, we decide on the more plausible grammatical function respectively on the more plausible sequence of grammatical functions. The main criteria for the decision are the unmarked word order and the semantic content.

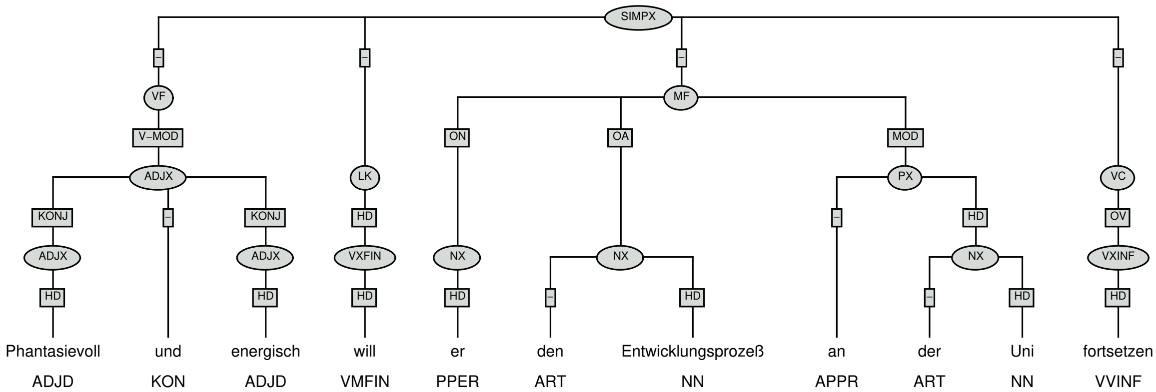
Therefore, in the first example above, OA appears in VF whereas ON has its position in MF. For elliptical headlines, we assume a passive construction if the verb in VC is a past participle and an active construction if the verb in VC is an infinitive (cf. 4.7.2 and 7.4).

5.3 Modifier Attachment

Modifiers either modify one specific constituent or more than one constituent. The scope of modification can even range over the whole sentence structure. Therefore, they are either unambiguous or ambiguous. An unambiguous constituent that modifies just one other 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 X-MOD edge labels, which 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:



If a modifying constituent is ambiguous (i.e. it modifies more than one constituent, the entire sentence, or a constituent that occurred in previous sentences), it is attached to its topological field and given the ambiguous edge label MOD to preserve ambiguity. In the following example *an der Uni* either modifies the accusative object *den Entwicklungsprozeß* or the verb *fortsetzen*:



We formulate the following definitions for MOD and X-MOD:

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). Sometimes it is difficult to determine whether a modifier is definite or not. In cases of doubt, modifiers are marked as ambiguous (MOD) rather than as definite modifiers.

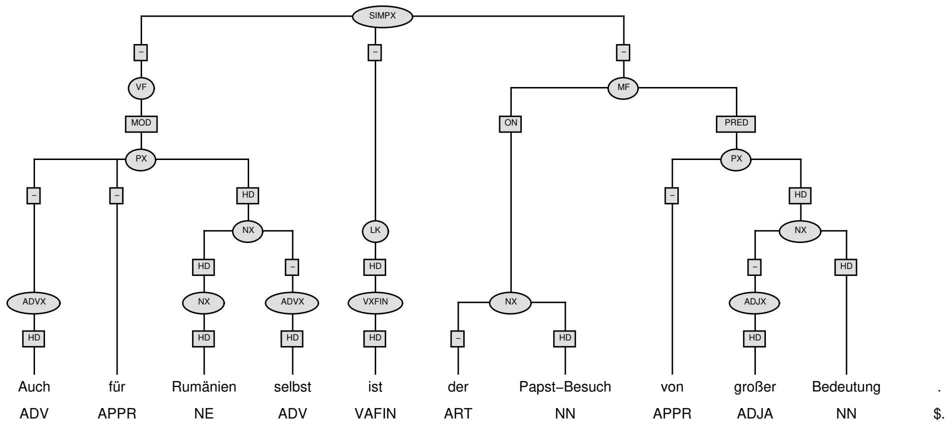
Definition of X-MOD:
X is a variable that can be replaced by labels for syntactic categories like OA, OPP, MOD, V. X-MOD marks long-distance modification which is unambiguous, e.g. relative clauses (*Aber es gäbe (intelligente Lösungen OA), (die kein Geld kosten OA-MOD)*).

Typical MODs and V-MODs:

Generally, modifying subclauses (e.g. *Katastrophenstimmung herrscht erst, [wenn nichts mehr zu verheimlichen ist] (MOD).*) are MOD because they modify the complete main clause. Modifying particles and adverbs like *da, dann, auch, eigentlich, ja, vielleicht, auch, natürlich* usually show attachment ambiguity and therefore are annotated as MOD. Only if they unambiguously express the modification of the verb (e.g. *Das Buch liegt da.* or *Er geht auch.*) they carry the edge label V-MOD. Pronominal adverbs (PROP) like *dabei, dafür, trotzdem, deswegen, hierauf, etc.* are either ambiguous (e.g. *Dabei* (MOD) *erscheinen Sie in anderen Verlagen.*) or unambiguous [e.g. *Er achtet dabei* (V-MOD) *auf alles.*). Non-pronominal adverbs such as *vorher, später*, etc. in most cases give temporal or local information. Therefore, they are rather V-MOD than MOD.

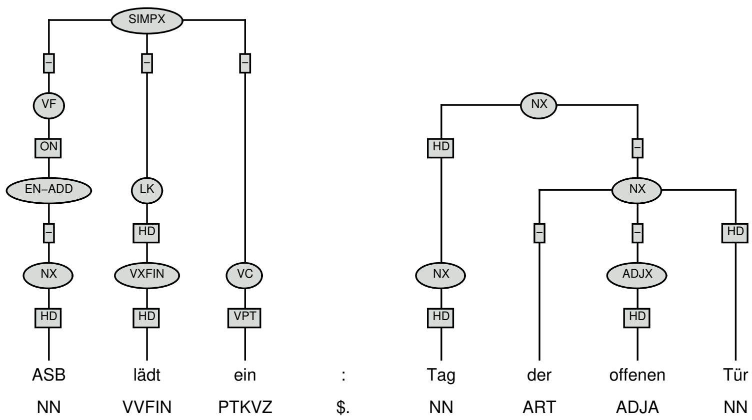
5.3.1 Modifier Attachment in the Initial Field

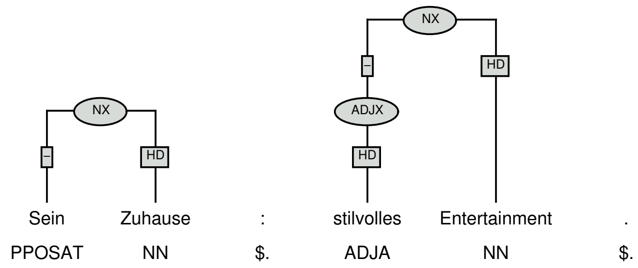
Since only one constituent is allowed in the initial field, all elements preceding and following the head are attached as premodifiers (low attachement) or postmodifiers (high attachment) concerning the attachment rules explained in 4.1.



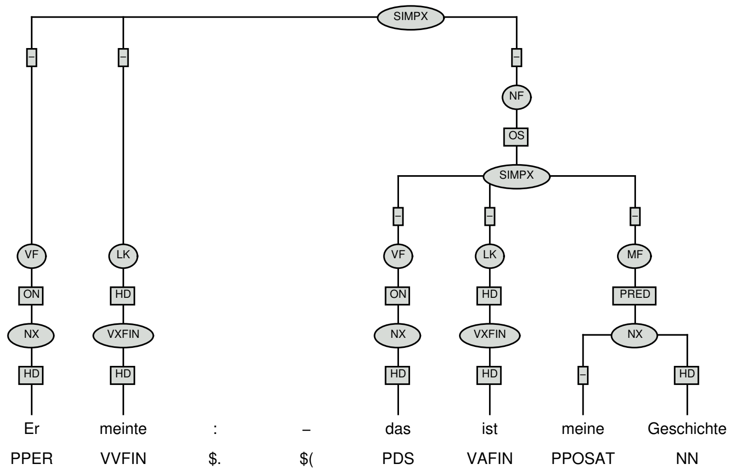
5.3.2 Attachment across Punctuation Marks

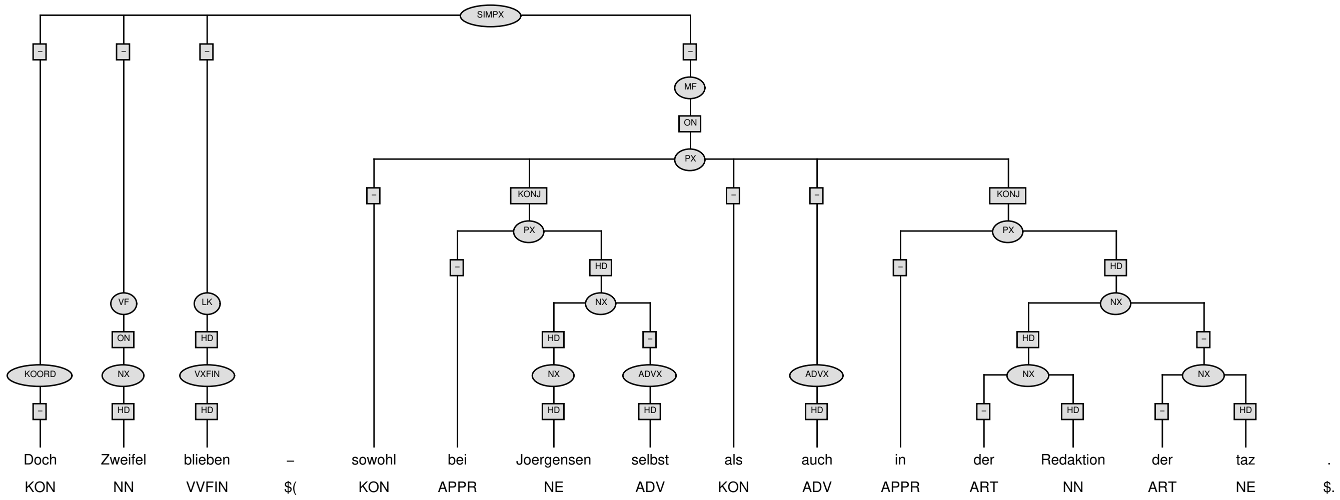
The punctuation marks : and - and ... separate a syntactic construction within a unit unless there is no syntactic dependency relation between the two parts (cf. 3.4.5) like in the following:





Attachment is necessary if the part following the punctuation mark has a grammatical function within the sentence structure:



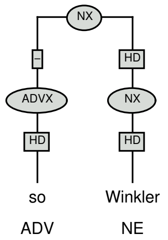


5.3.3 Ambiguous Modifiers in Isolated Phrases

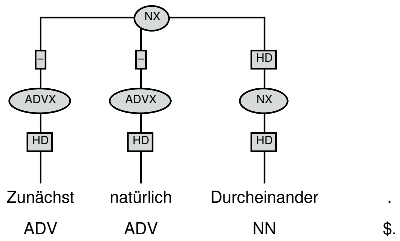
Since isolated phrases (cf. 3.4.5) do not consist of topological fields, ambiguous modifiers (MOD) have to be attached to the phrase itself. The isolated phrase is projected one level higher and the modifier is attached on this higher level. Thus, the information about ambiguity can be preserved even without topological fields or explicit MOD labelling, just by the existence of yet another projection level of the phrase.

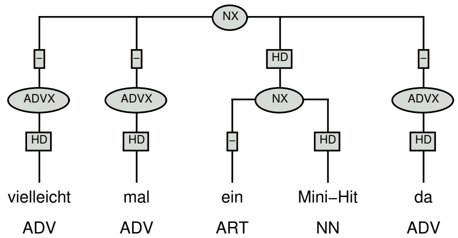
The overall attachment strategy has been chosen in order to keep syntactic structure flat and to be able to preserve attachment ambiguity where necessary.

In the following examples, *so* may refer to something that is implicit or has been mentioned before:



If there is more than one ambiguous modifier in an isolated phrase, all of them are attached on the next higher level. The mother node of this isolated phrase is marked with the node label of the modified phrase.





Chapter 6

The Annotation of Sentences

The approach of topological fields supports the *flat clustering principle* inasmuch MF and 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.

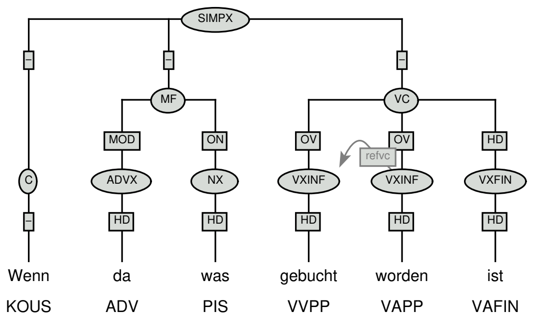
In the following sections, the annotation of sentence structures will be demonstrated.

6.1 Sentence Initial Fields

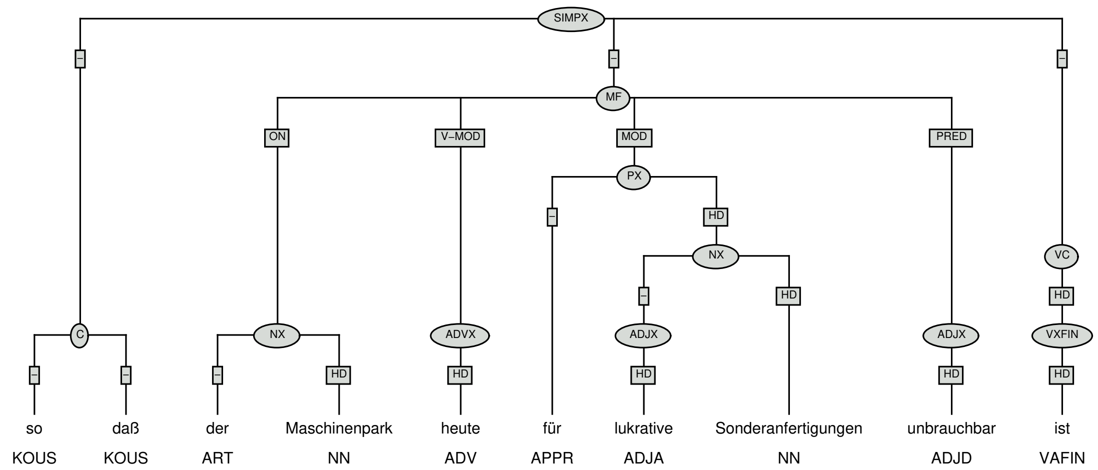
6.1.1 The C-Field in Verb-Final Clauses

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 pronouns (PRELS), interrogative (PWAV) pronouns and (complex) interrogative or relative phrases. Thus, it only occurs in verb-final clauses, except for comparison clauses with the conjunction *als*.

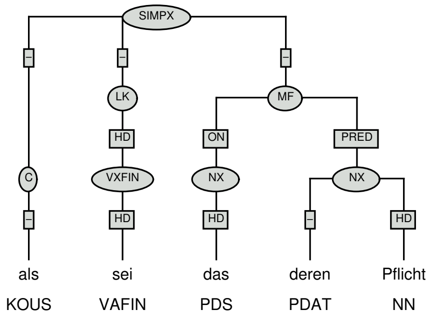
In case of a conjunction, we directly project to the C-field:



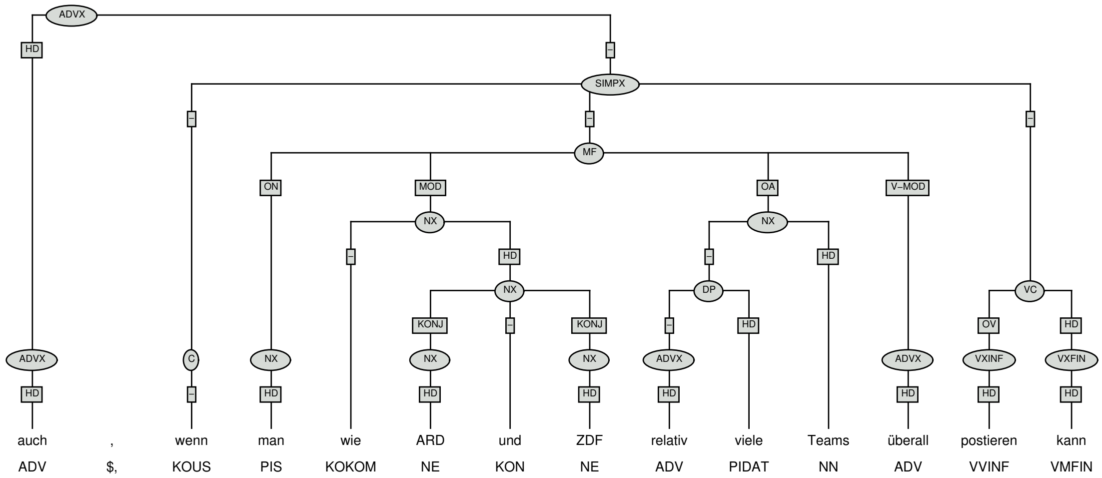
There are conjunctions in German which consist of two elements (e.g. *so daß* and *als ob*). Both of them are also directly attached to the C-field, while none of them carries a head label.



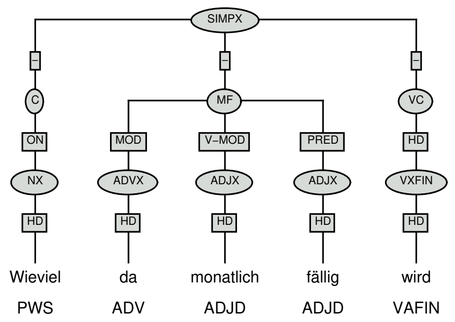
Only comparison clauses with *als* allow for a C-field and a left sentence bracket in the same clause:

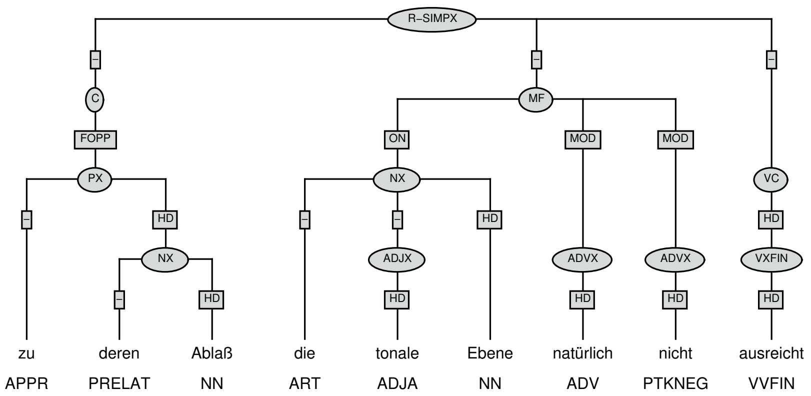


Since C generally does not contain more than one constituent, the adverb *auch* in the following example is not supposed to occur in the C-field together with the conjunction *wenn*. The *wenn*-clause is annotated as the modifier of the adverbial phrase *auch*, ie., the adverbial phrase subcategorizes for the verb-final clause.



If the constituent in the C-field is a pronoun or a complex phrase, it is first projected to the phrase level and then projected to the C-field. The edge label below the C-Field denotes the grammatical function of this constituent.

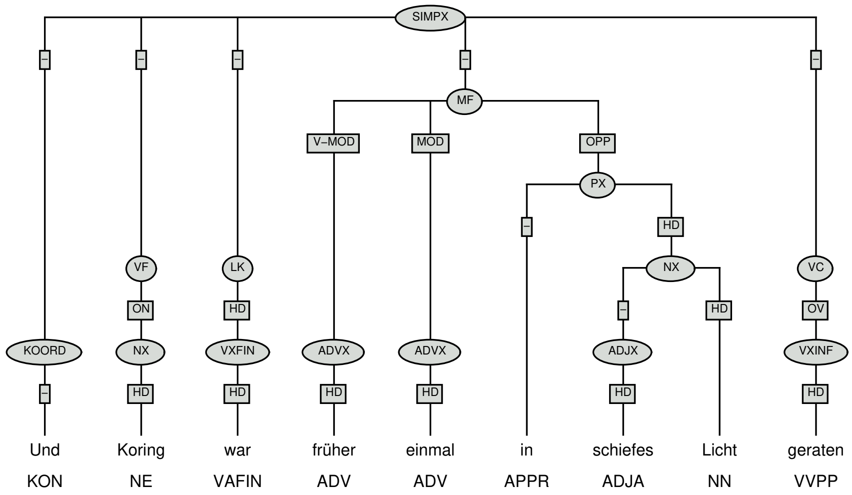


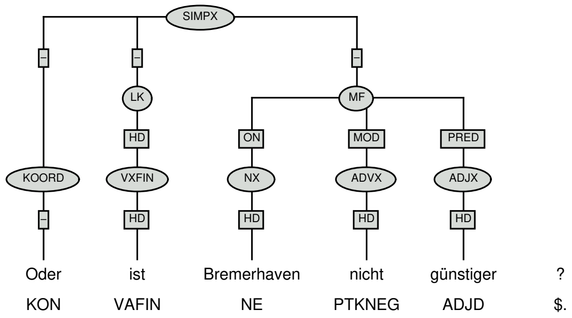


6.1.2 The KOORD-Field in all Clause Types

The KOORD-field is optionally the left-most field of all clause types (V-1, V-2, V-end). Therefore, it can only occur at the beginning of a syntactic unit (cf. 3.4.3).

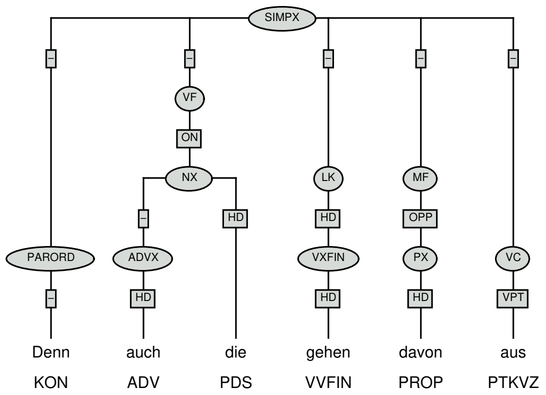
For verb-second clauses, it can be regarded as an alternative field to the PARORD-field. The KOORD-field contains coordinative particles like *und, oder, aber*, etc. (cf. Höhle (1986)). Here are two examples of different clause types:





6.1.3 The PARORD-Field in Verb-Second Clauses

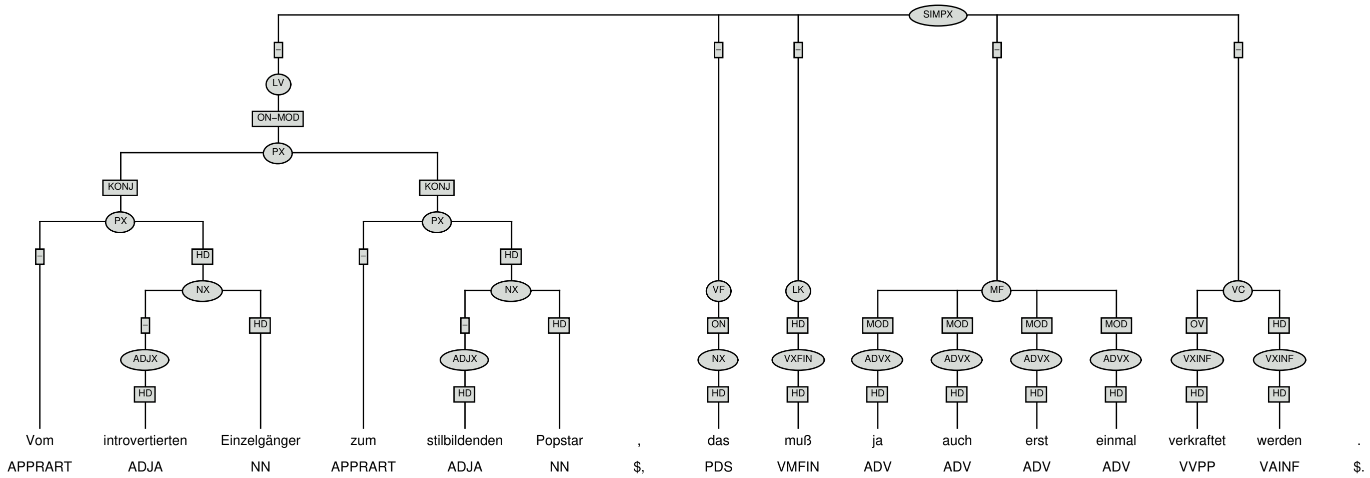
PARORD is an alternative field to KOORD for verb-second clauses only. Typical PARORD expressions are *denn, weil*[1]:

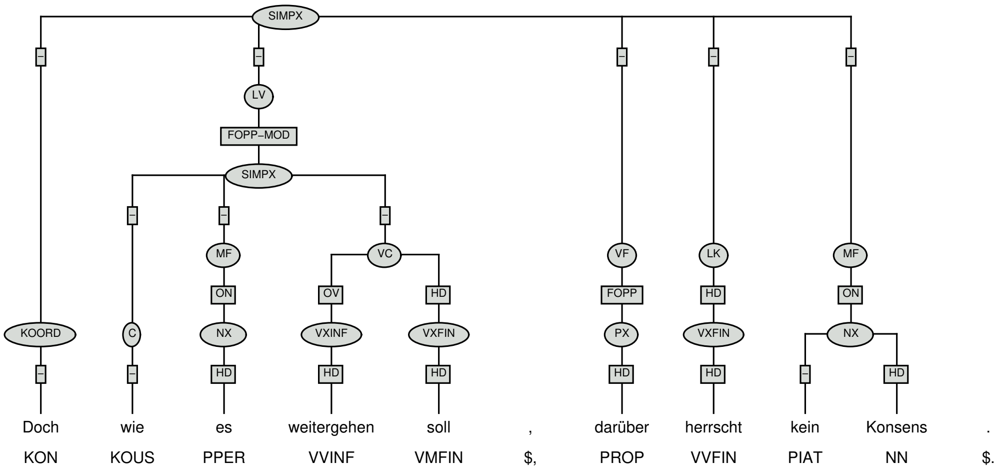


[1] *weil* can occur in verb-second and in verb-final clauses. In the first case, it is in the PARORD-field, in the latter case, it belongs to the C-field.

6.1.4 Resumptive Constructions: The LV-Field

Resumptive constructions are analysed as suggested by Höhle (1986) and Kathol (1995), by using the field LV (*Linksversetzung*) which is located on the left side of VF. In general, the LV-field is not restricted to one constituent. The typical feature of a resumptive construction is that there is a (pronominal) constituent somewhere in the sentence, on the right hand side of the LV-field, which refers back to the expression within the LV-field. Therefore, we use the X-MOD label to indicate this kind of long-distance dependency.



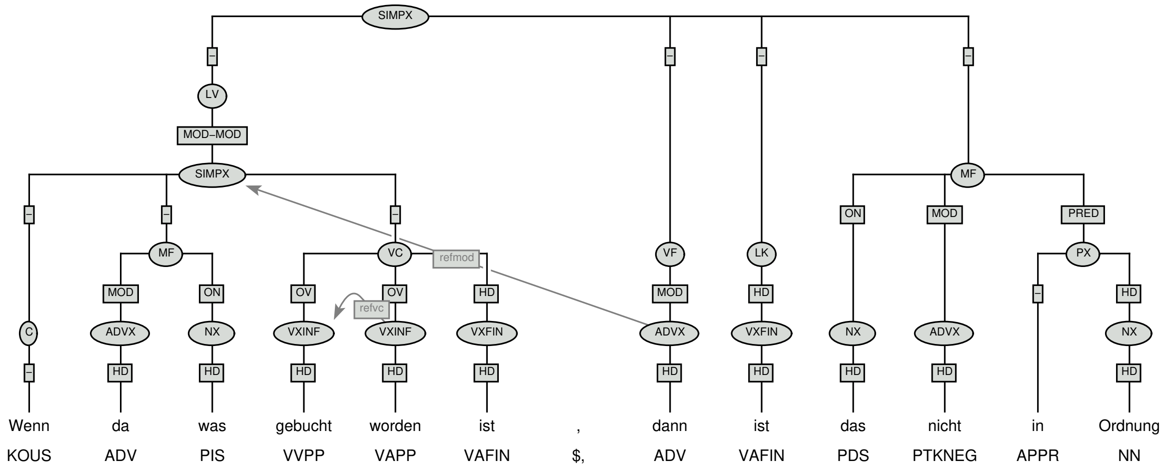


Grammatical functions within a LV-construction are assigned according to the following principle:

- 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 like in the first example above, which is also in strong accordance with the assumption that the original position of the subject in verb-second clauses is VF.

In constructions with *wenn ... dann ...*, the *wenn*-clause, which is semantically a precondition to the *dann*-clause, is in the LV-field in correlation with *dann*. Therefore, *dann* (MOD) refers back to the *wenn*-clause (MOD-MOD):



If *dann* is not present in the matrix clause , the *wenn*-clause occurs in VF. In this case, the *wenn*-clause is labeled as MOD because there is no explizit corelating constituent. It rather refers to the whole matrix clause, e.g.(*Wenn da was gebucht worden ist* (MOD), *ist das nicht in Ordnung.*)

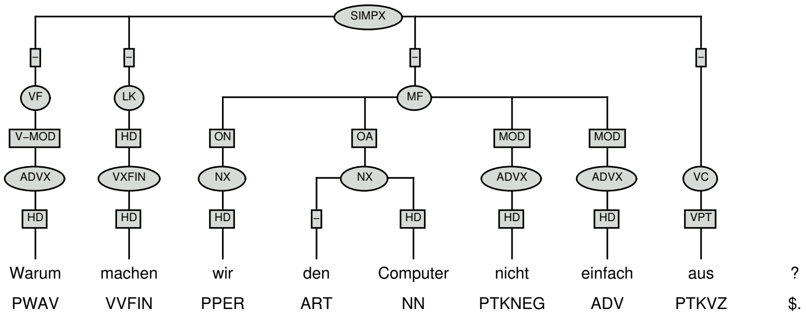
6.2 Questions

6.2.1 W-Questions

In general, w-questions are verb-second clauses with interrogative pronouns in 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:

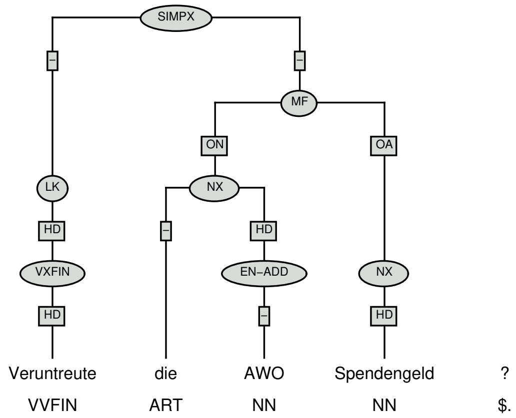
[PX wann] marschieren wir auf Peking? → *[PX im Juni]*
[ADJX wie] wirke ich? → *[ADJX gut], [ADJX schlecht]*
[NX welche Verbindungen] hat er → *[NX wichtige Verbindungen]*

warum is an exception in so far as 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 in order to be consistent with its part-of-speech tag information (PWAV):

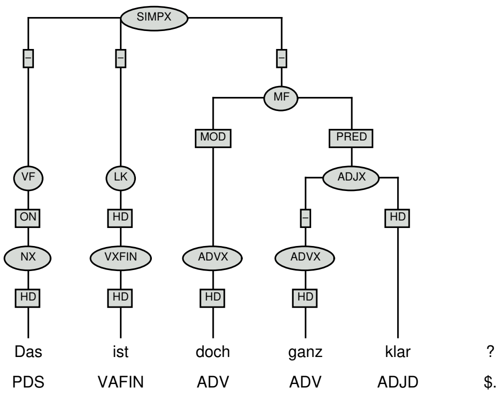


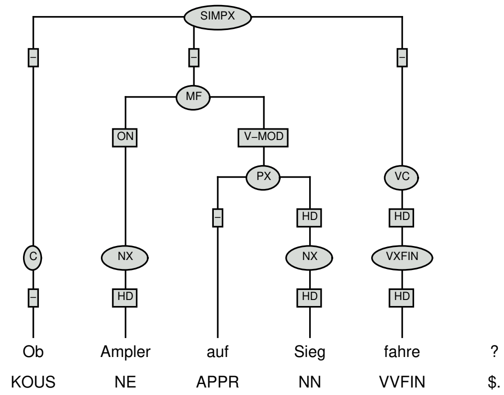
6.2.2 Yes - No Questions

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



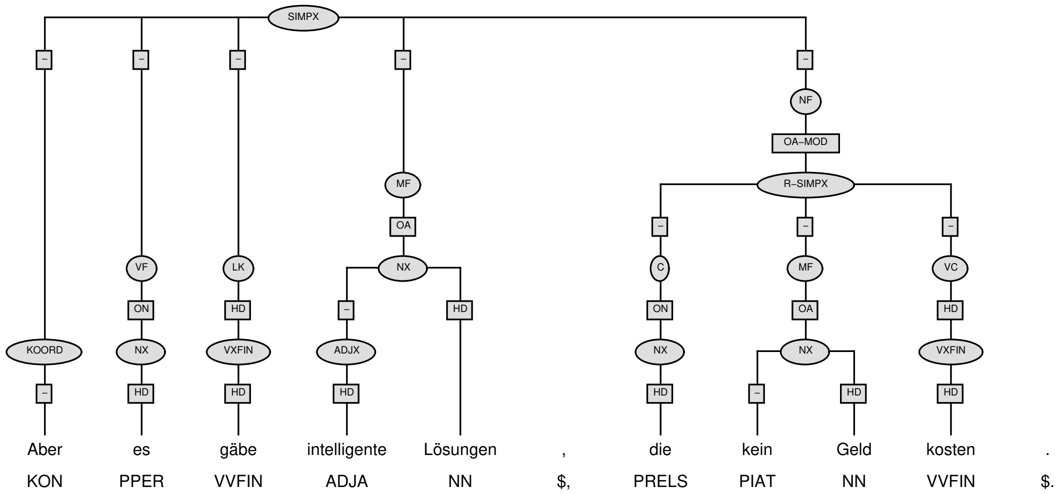
Otherwise, a question mark at the end of a verb-second or verb-final clause indicates that it is actually meant as a question:



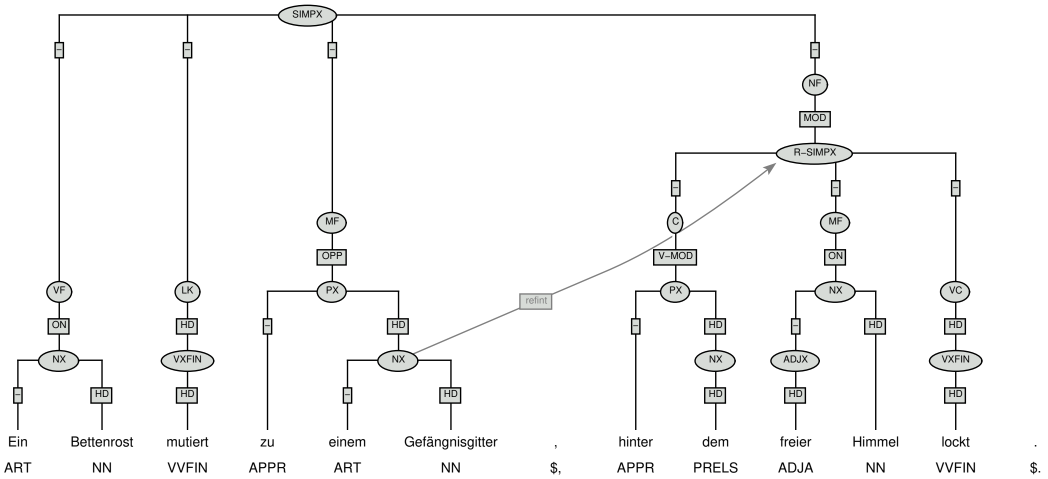


6.3 Relative Clauses

For relative clauses (R-SIMPX), the relative pronoun occurs in the C-field. It is first projected to the phrase level before it is attached to the C node. The relative clause itself is located in 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:

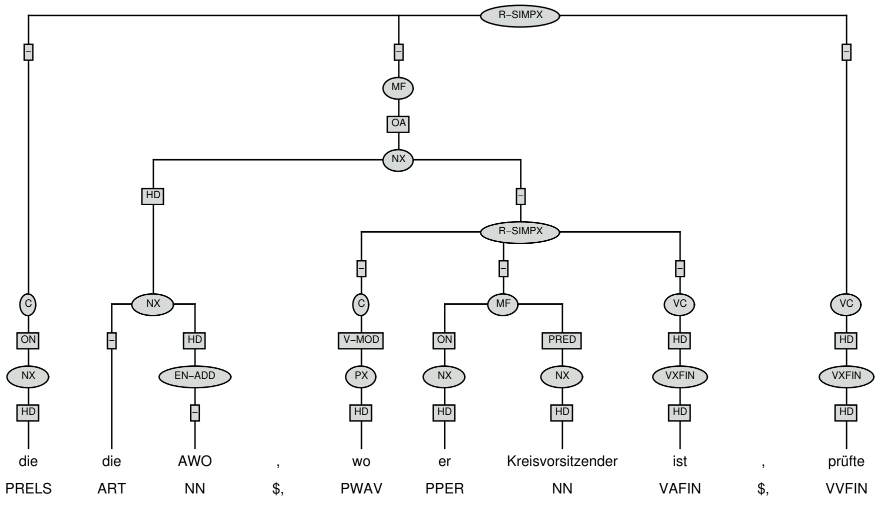


If the head noun phrase of the relative clause is the noun phrase of a prepositional phrase or a postmodifier within a complex phrase, the relative clause is labeled as MOD plus the secondary edge label REFINT (cf. 3.4.6):



The position of the relative clause in 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 to which it refers. For example, a verb complex can occur between the noun phrase and the relative clause (*Der Bettenrost ist zu einem Gefängnisgitter mutiert, hinter dem freier Himmel lockt.*). 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 NF. Unless there is strong evidence for a position in MF, the relative clause is located in NF.

If the relative clause and its head noun phrase are adjacent constituents in VF or MF, the relative clause modifies the noun phrase directly as a postmodifier.



6.4 Coordination

Coordination is a syntactic phenomenon that occurs on the following annotation levels: phrase level, field level, and sentence level. Within coordinations, the conjuncts are first projected to their phrase, field, or clause level. In a second step, they are attached to their mother node which is n-ary branching (conjunctions between the conjuncts). This scheme is the same for all syntactic categories.

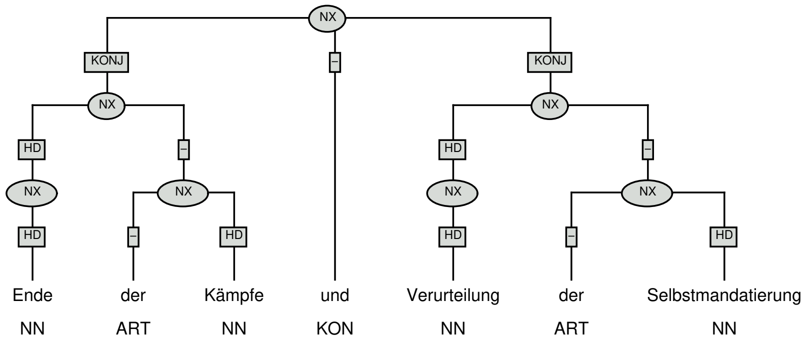
The edge labels between the mother node and the conjuncts of the coordination are labeled as KONJ. This edge label supports the distinction between conjuncts, modifiers, and conjunctions within complex conjunctions (cf. 6.4.3), as well as the distinction between coordinations and elliptical constructions (cf. 6.5).

In contrast to coordinating conjunctions in the KOORD-field, coordinating conjunctions in coordinations (*und, oder,* etc.) are directly attached to the mother node of the conjuncts. The class of coordinating conjunctions consists of single, e.g. *und, oder, aber, als,* as well as of complex conjunctions, e.g. *entweder oder, weder noch, sowohl als.* Generally, coordinating conjunctions may coordinate constituents of any category. Moreover, they can form asymmetric coordinations in which the conjuncts belong to different syntactic categories (cf. 6.4.2).[2] In order to distinguish conjunctions from conjuncts within a coordination, their edge labels are empty.

In the following, coordination on all annotation levels as well as specific cases of coordination, e.g. split-up coordinations, will be demonstrated.

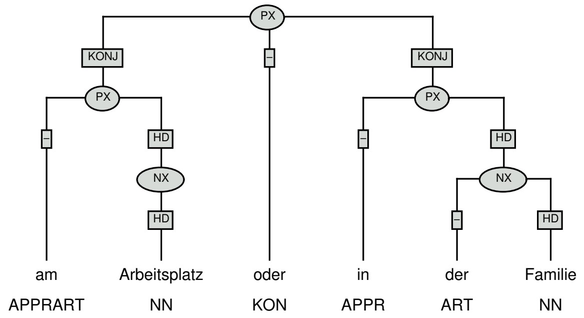
6.4.1 Coordination of Phrases

Noun Phrases

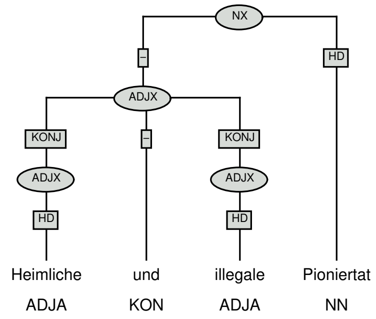


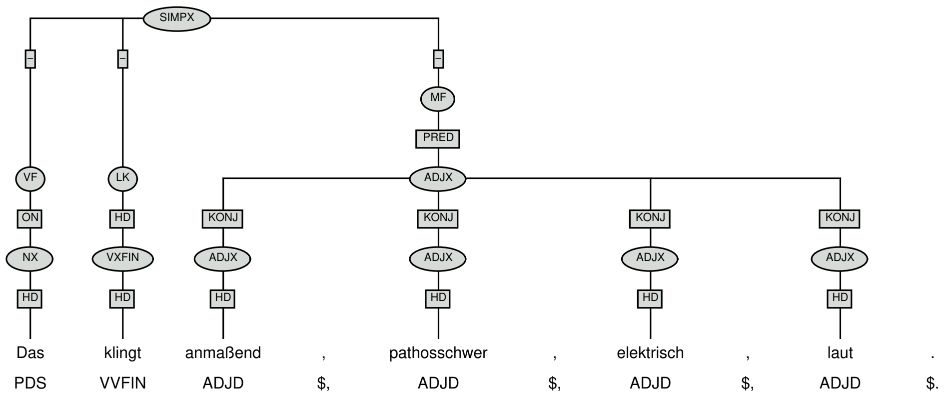
[2]If *bis* is used as a conjunction like in *10.000 bis (KON) 20.000 koreanischen Daewoo PKW* it is tagged as KON. But remember that *von ... bis ...* phrases are treated differently (cf. 4.4.1).

Prepositional Phrases

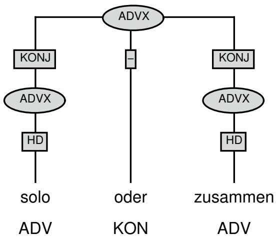


Adjectival Phrases



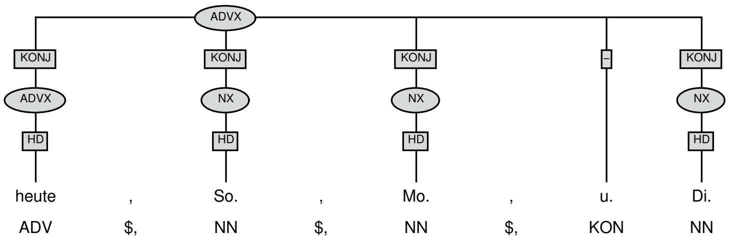


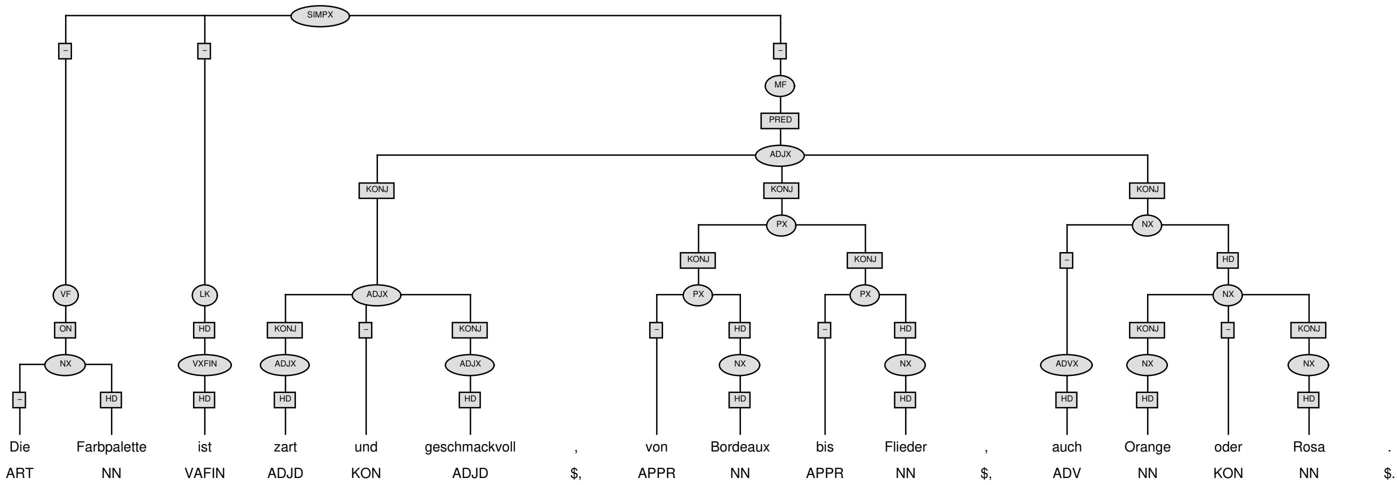
Adverbial Phrases

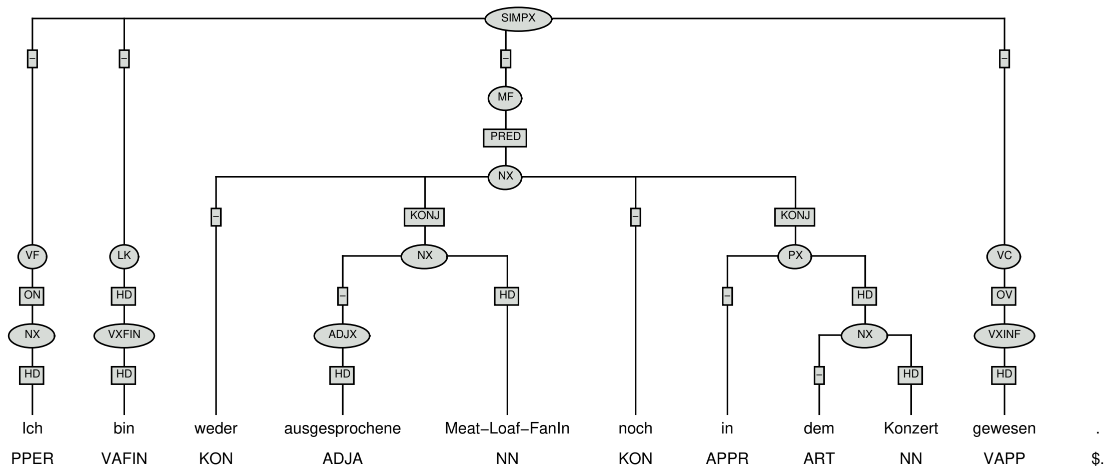


6.4.2 Asymmetric Coordination

Since constituents of different syntactic categories can be coordinated, it has to be decided on a label for the mother node of the coordination. 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:

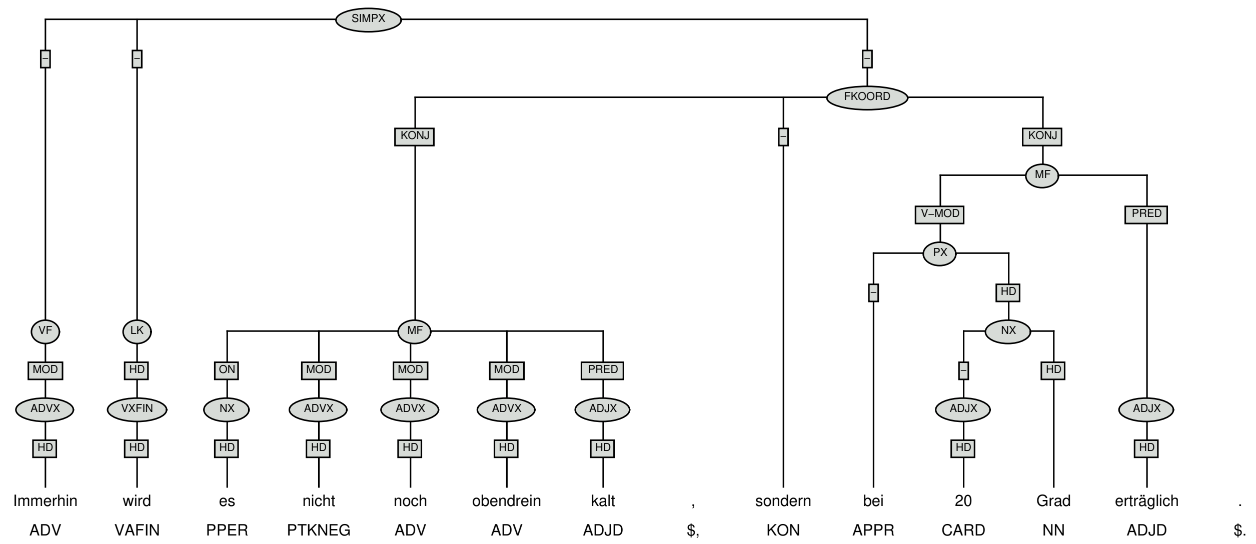


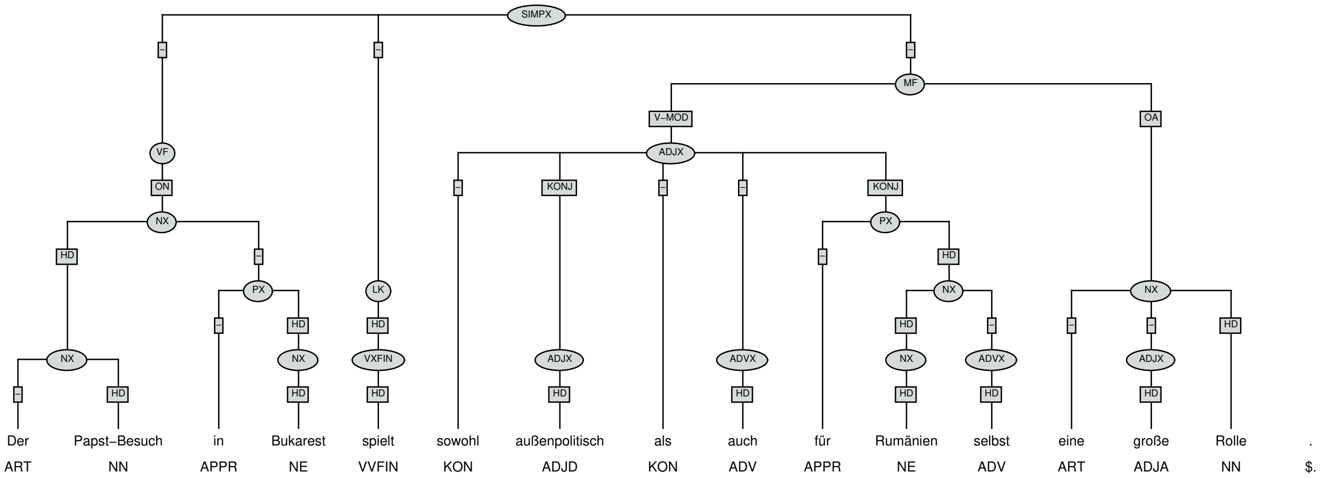


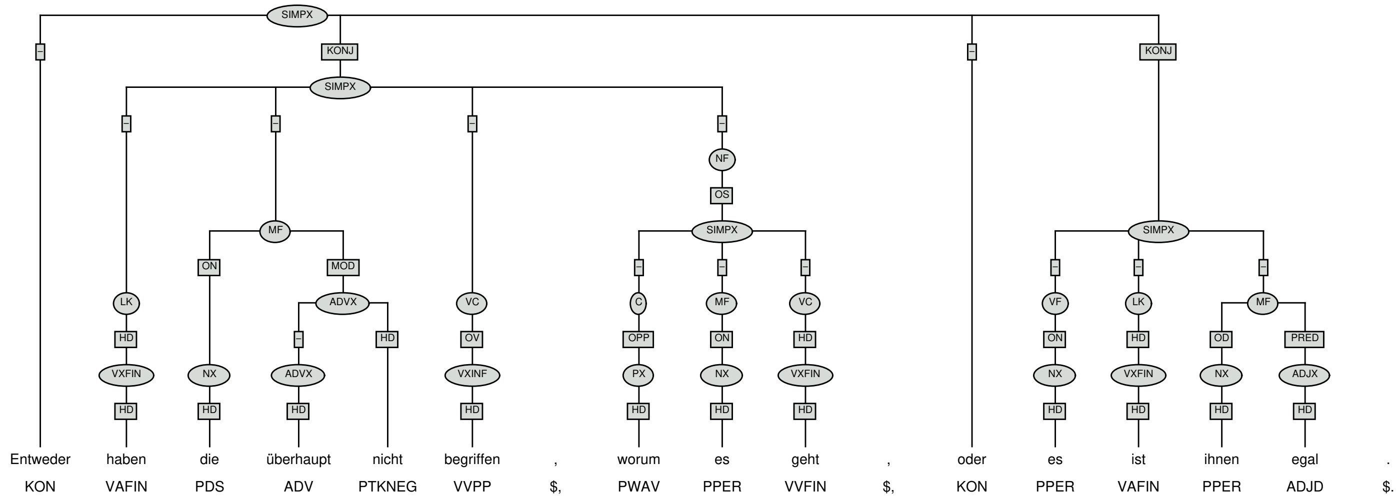


6.4.3 Coordinations with Complex Conjunctions

The conjuncts and conjunctions of a coordination with complex conjunctions are also attached on the same level following the above mentioned rules for coordination. Both parts of complex conjunctions like *entweder oder* and *sowohl als* are tagged as KON. The latter one usually occurs together with the adverb *auch*, which is tagged as ADV, projected to the phrase level, and then attached to the mother node of the coordination. The same applies for *nicht* in coordinations with *sondern*. *sondern* is tagged as KON, whereas *nicht* is always tagged as PTKNEG:

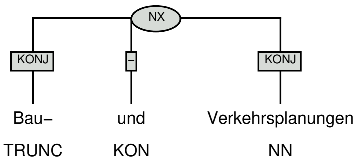


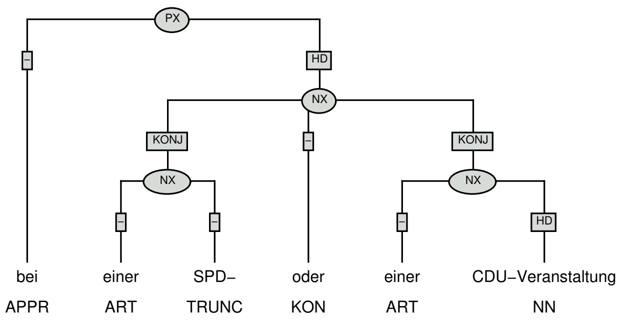


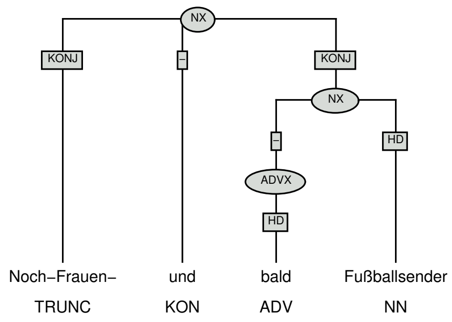


6.4.4 Coordinations with Truncated Words

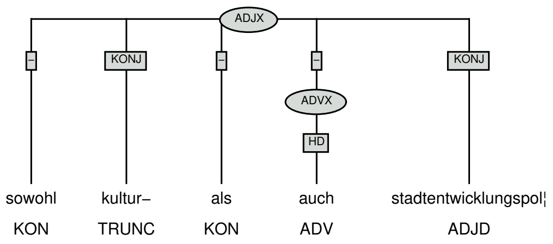
In contrast to complete lexical entries, a truncated word is directly attached to the conjunct and the conjunction. Neither the truncated word nor the second conjunct are first projected to the phrase level. Their edge label is KONJ. If both conjuncts have a determiner or are premodified, only the complete lexical entry carries a head.

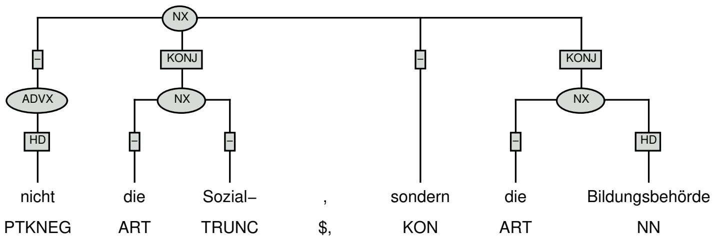






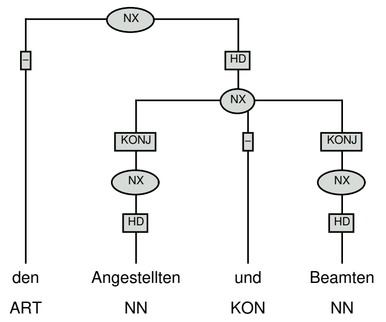
With complex conjunctions, also direct attachment of the conjuncts is applied.

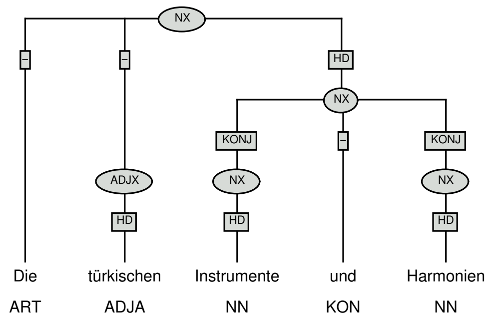




6.4.5 Attachment Principles of Coordination within Phrases

If two or more nominal conjuncts occur together with a common determiner and/or adjectival phrase, first the conjuncts are projected to their phrase level and than the determiner or the adjectival phrase is attached to the coordination on a higher level according to the *high attachment principle*. Thus, the modification scope comprises the entire coordination. The coordinated part is assigned the head function.



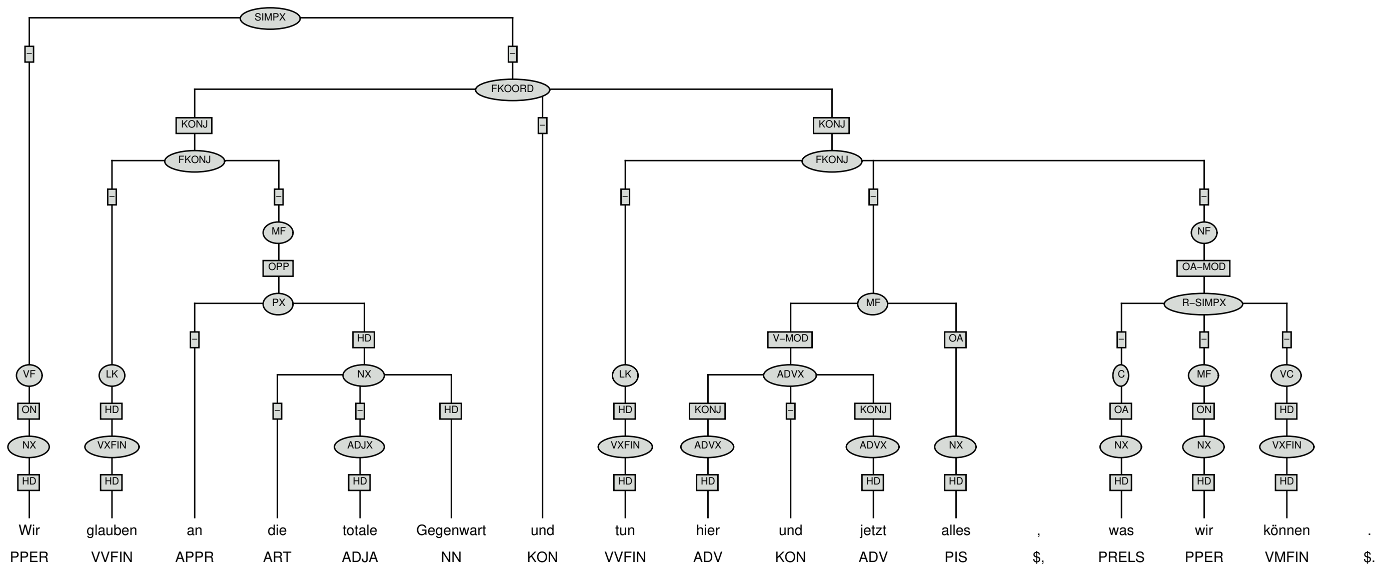


6.4.6 Coordination of Topological Fields

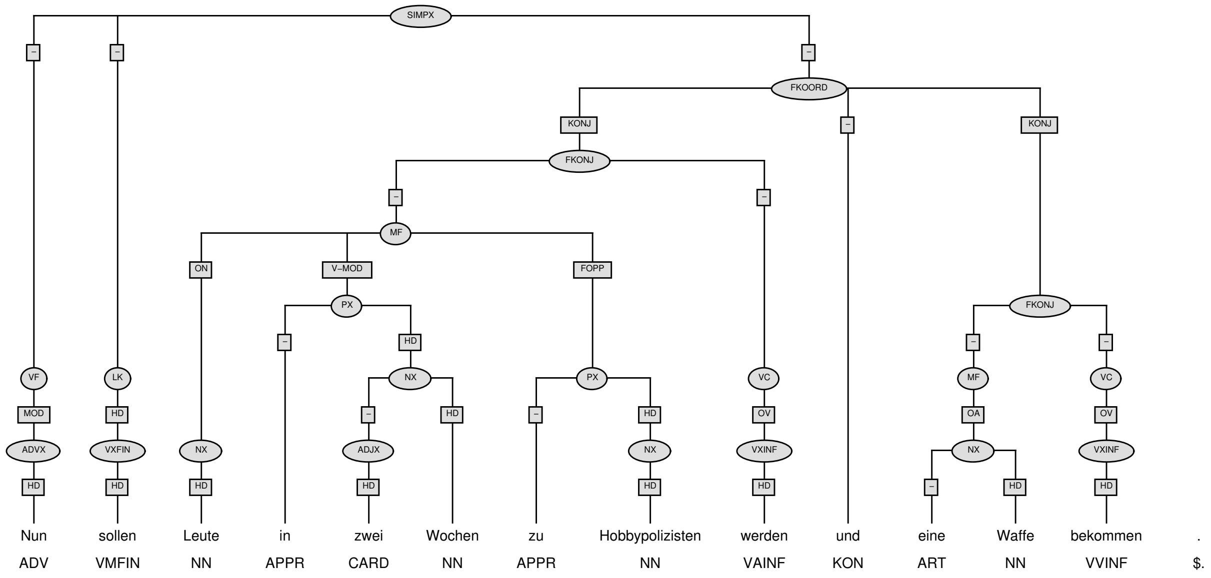
The conjuncts of a coordination of topological fields are either single fields (cf. 6.4.4) or a combination of fields. Possible combinations are, for instance, (MF + VC), (LK + MF), (LK + MF + VC). The node label for these conjuncts is FKONJ (conjunct consisting of fields) and the mother node of a coordination of conjuncts of fields is FKOORD.

In a coordination of conjuncts of fields, the following annotation steps are involved:

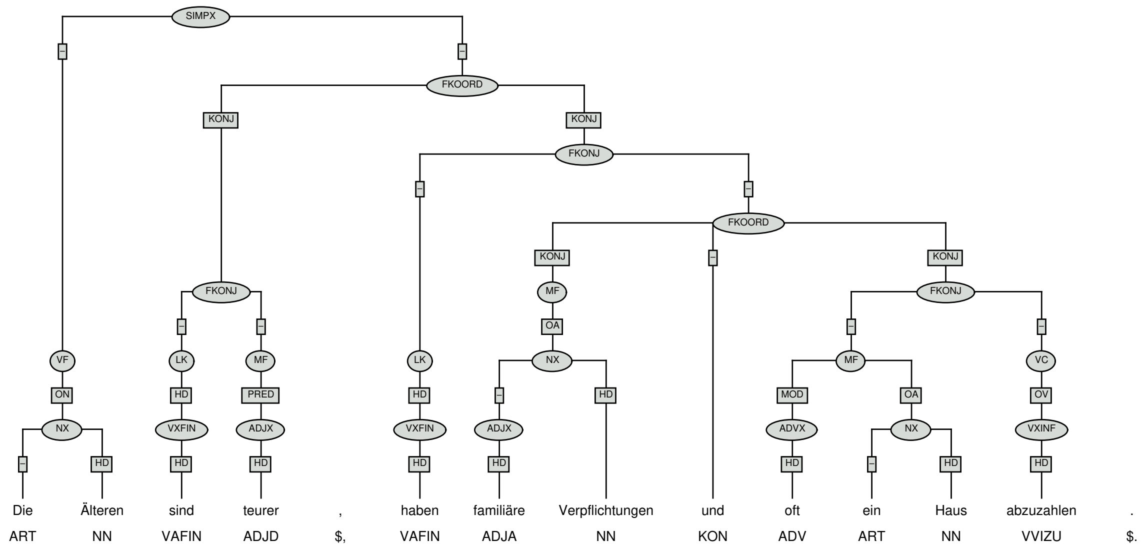
1. The constituents are attached to the fields in which they occur in (MF, VC, NF etc.).
2. Each conjunct (concatenation of fields or single field) is labeled as FKONJ.
3. The conjuncts are attached to the general coordination field FKOORD.



Often, the subject of the sentence occurs only in the left field conjunct:

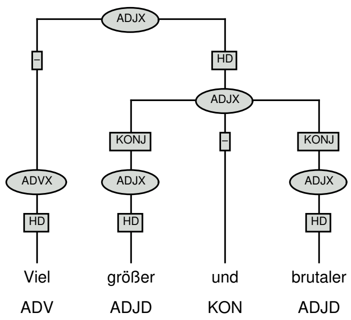


A coordination of fields may also be an embedded structure. In this case, FKOORD functions also as conjunct label:

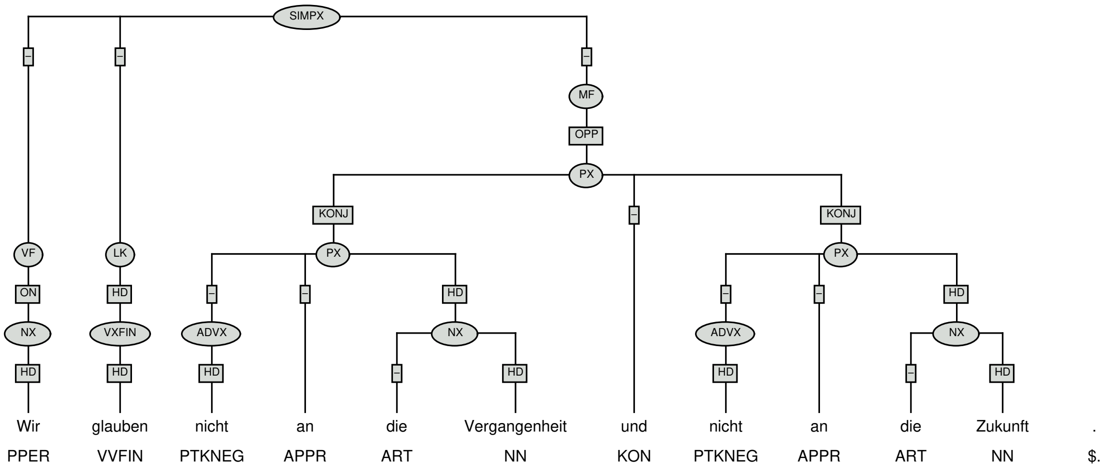


6.4.7 Attachment of Ambiguous Modifiers in Coordination

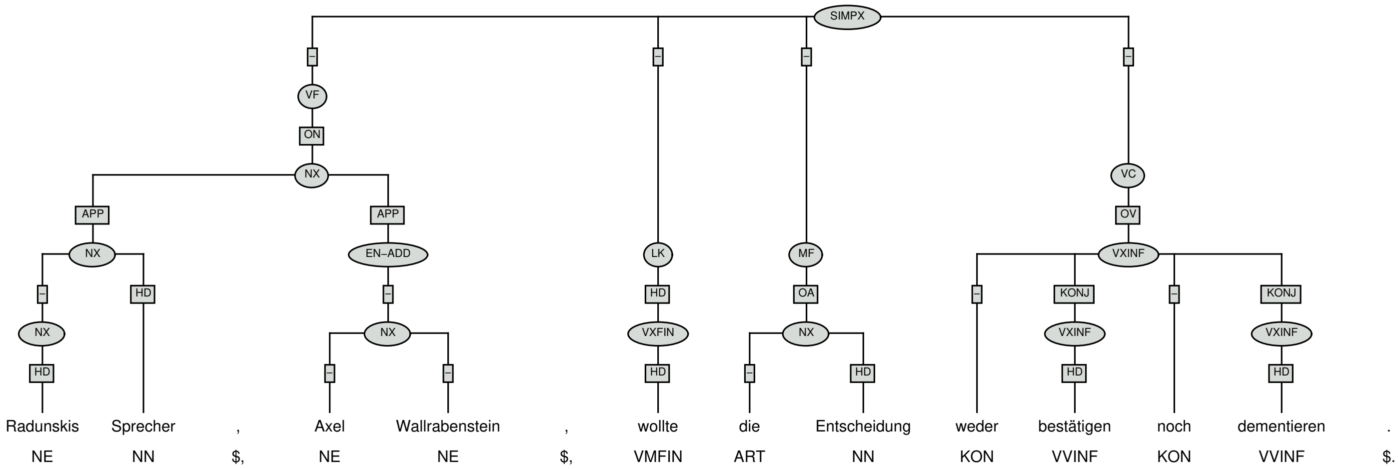
Within phrases, the modification scope of a premodifier can be ambiguous. Therefore, high attachment is applied to preserve ambiguity. In the following example, the adverb modifies the coordination of adjectives rather than only the first adjective:

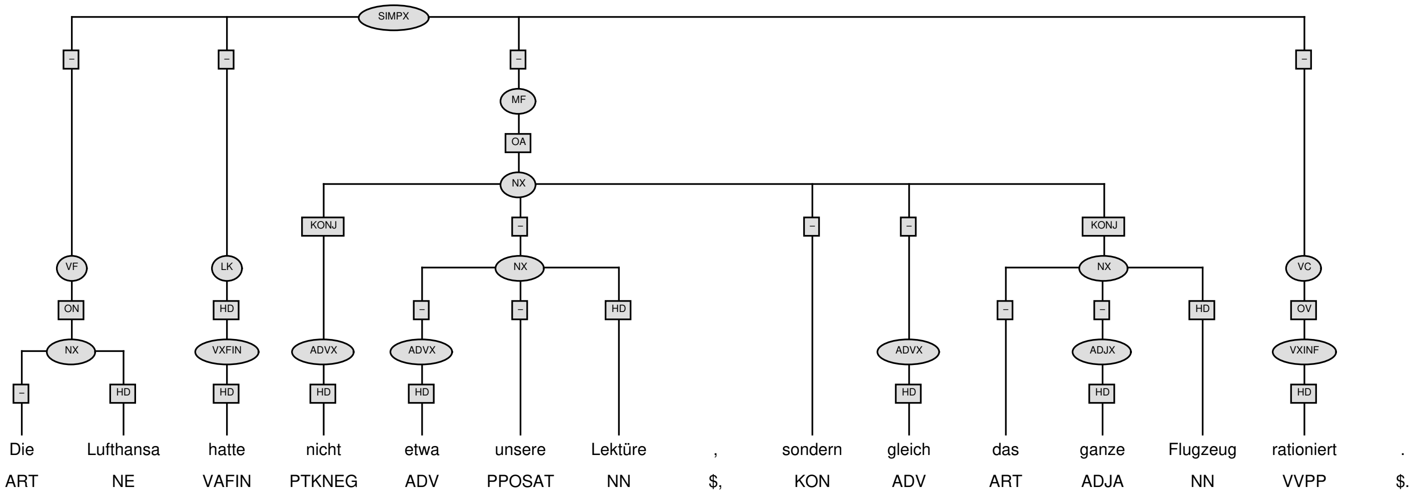


Modifying constituents are attached to a conjunct rather than to a field if their modification scope is limited to the conjunct.

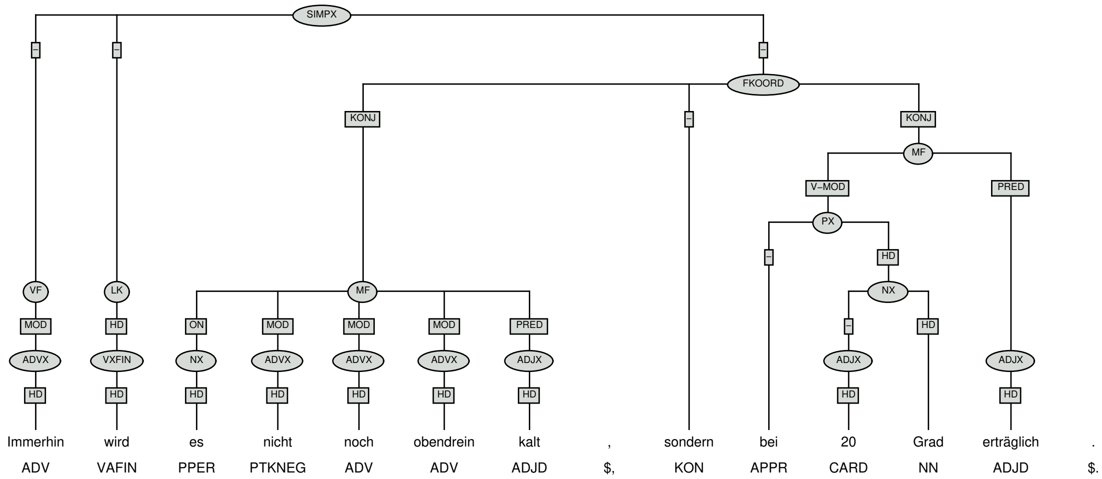


Also in coordinations with complex conjunctions, attachment on the phrase level is applied if possible.



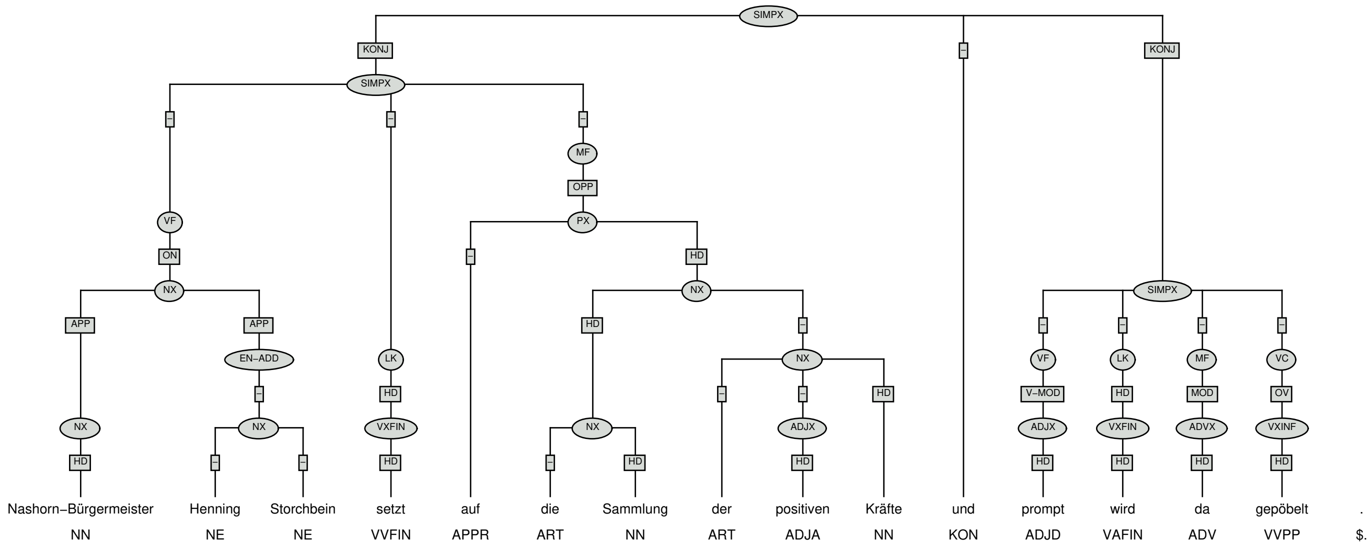


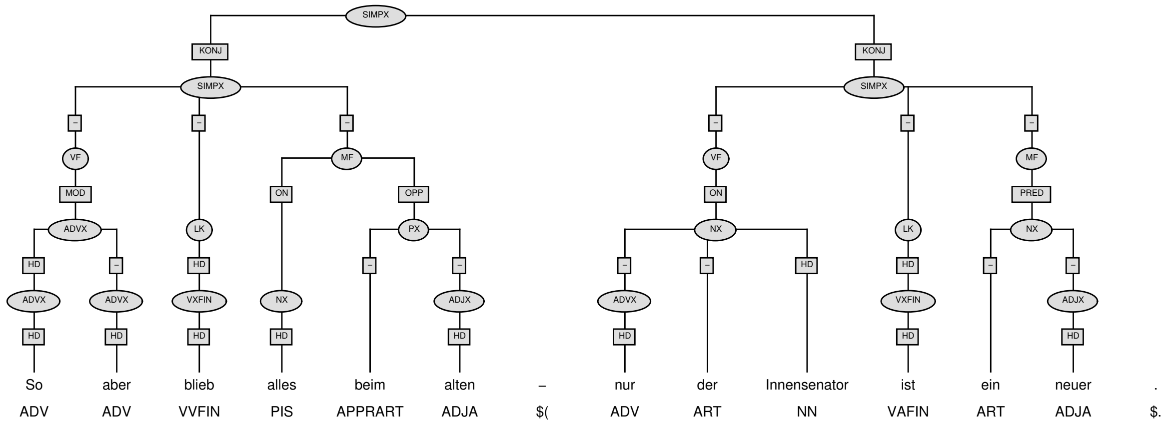
If there is more than one constituent within a conjunct, each with its own grammatical function, these constituents are first attached to the respective field node. Then, the fields are coordinated:



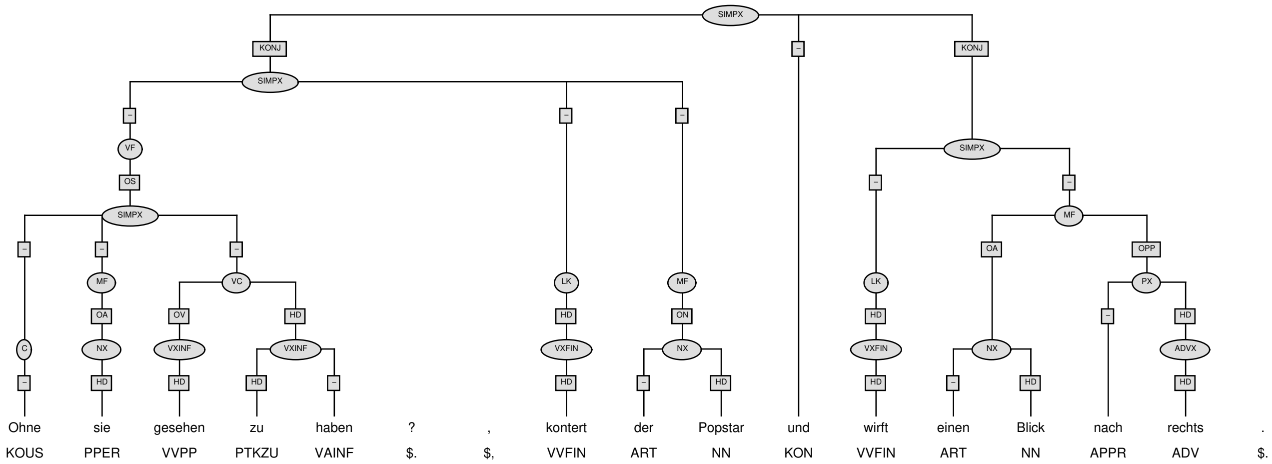
6.4.8 Coordination of Sentences

In accordance with the *longest match principle*, complete sentences are coordinated as paratactic constructions when they belong to the same syntactic unit (cf. 3.4.3), i.e., they are coordinated by a conjunction, a comma, or a dash:

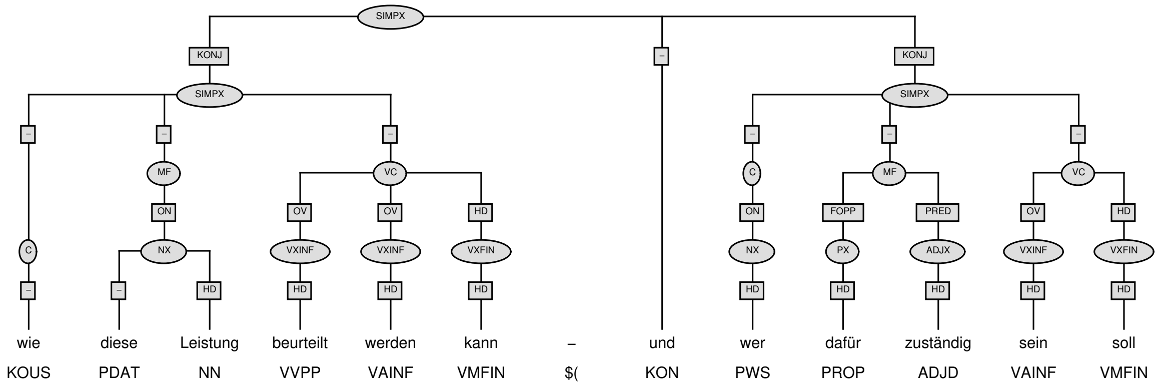




A coordination may also consist of two sentences with the subject of the whole construction only occurring in the left conjunct of the coordination.

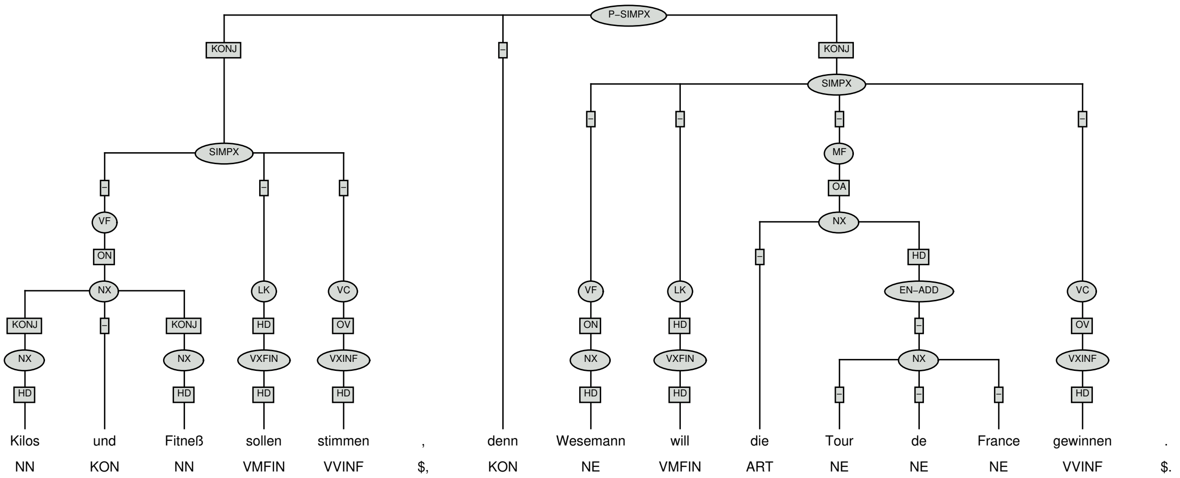


Subclauses (either in VF or in NF) with or even without a conjunction can also be coordinated.



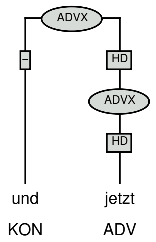
6.4.9 Paratactic Constructions with ***denn*** and ***weil***

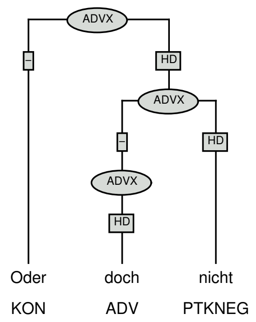
Paratactic constructions consisting of verb-second clauses conjoined by the conjunctions *denn* and *weil*, which also occur in the PARORD-field in the beginning of a sentence, are treated as equal conjuncts (verb-second instead of verb-final in *weil*-clause). In order to distinguish coordination of sentences with conjunct of the PARORD field from the above mentioned coordinations of sentences, these paratactic constructions are labeled as P-SIMPX instead of SIMPX.



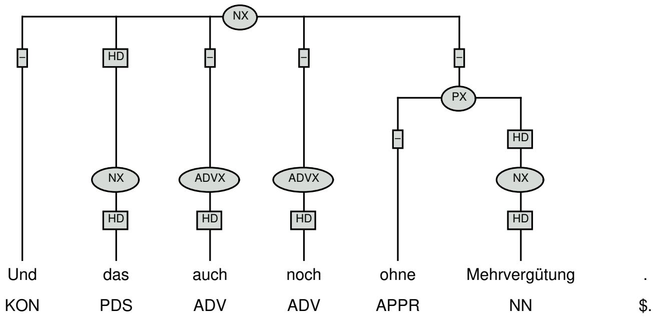
6.4.10 Conjunctions Occurring with Isolated Phrases

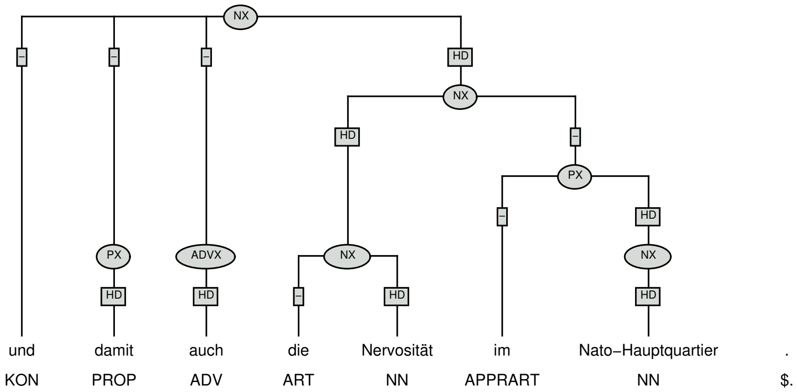
If a conjunct occurs isolated with a conjunction, high attachment is applied like in complete coordinations. But for isolated conjuncts, the conjunct is annotated as the head of the construction (HD instead of KONJ).





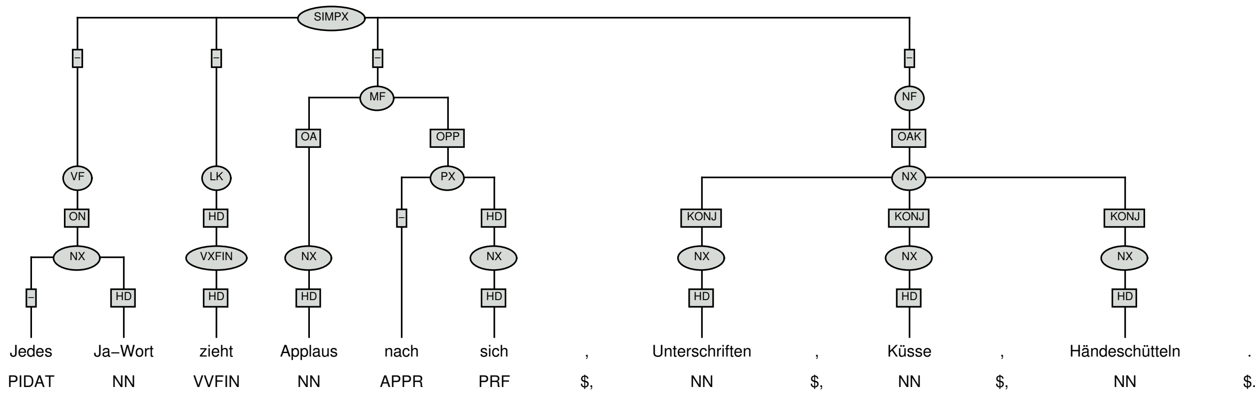
If there are modifiers which do not modify the conjunct itself because they are ambiguous or might modify something else rather than the conjunct, they are attached on the same (high) level as the conjunction:

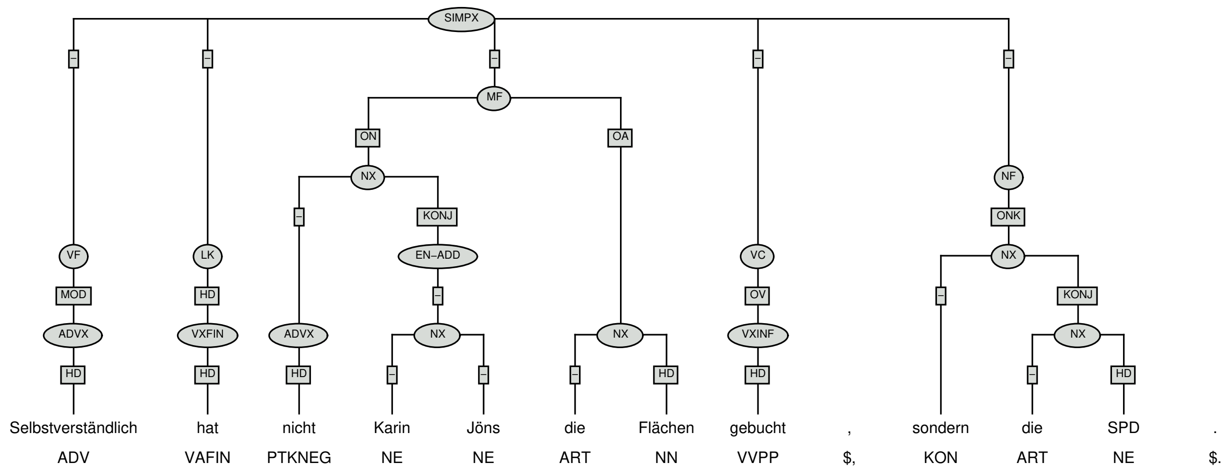


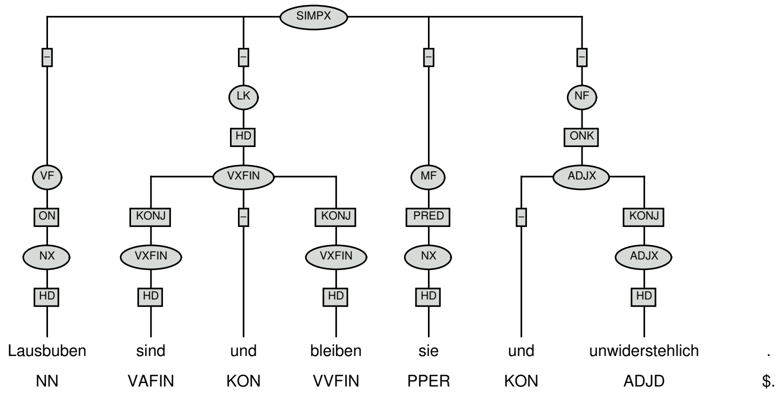


6.4.11 Split-up Coordinations

Closely related to isolated conjuncts are *split-up* coordinations. Generally, the left conjunct of a split-up coordination is located in MF, in rare cases in VF, and the right conjunct occurs in NF. In order 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.). In asymmetric coordination, the syntactic category of the second split-up conjunct determines the syntactic category one level higher up:





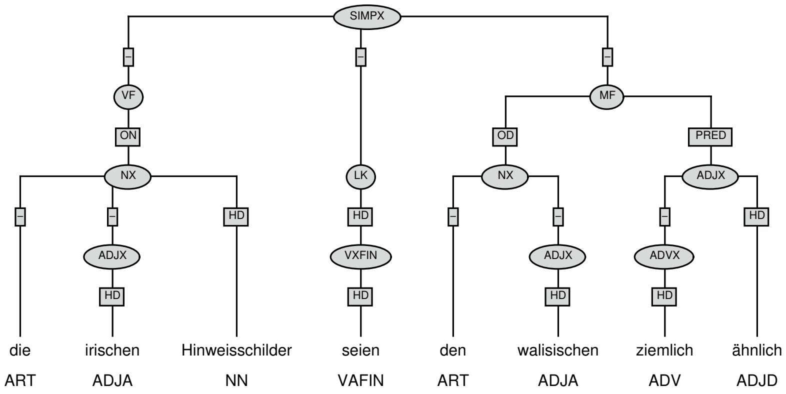


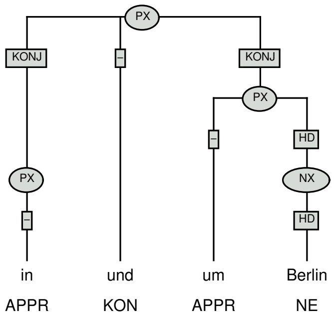
6.5 Elliptical Constructions

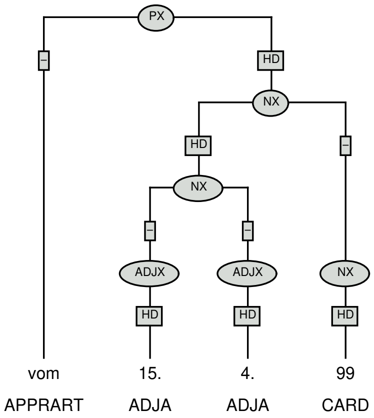
In elliptical constructions, syntactically necessary linguistic elements are missing which can be reconstructed from the context or the speech situation. Elliptical constructions appear on the phrase level as well as on the sentence level.

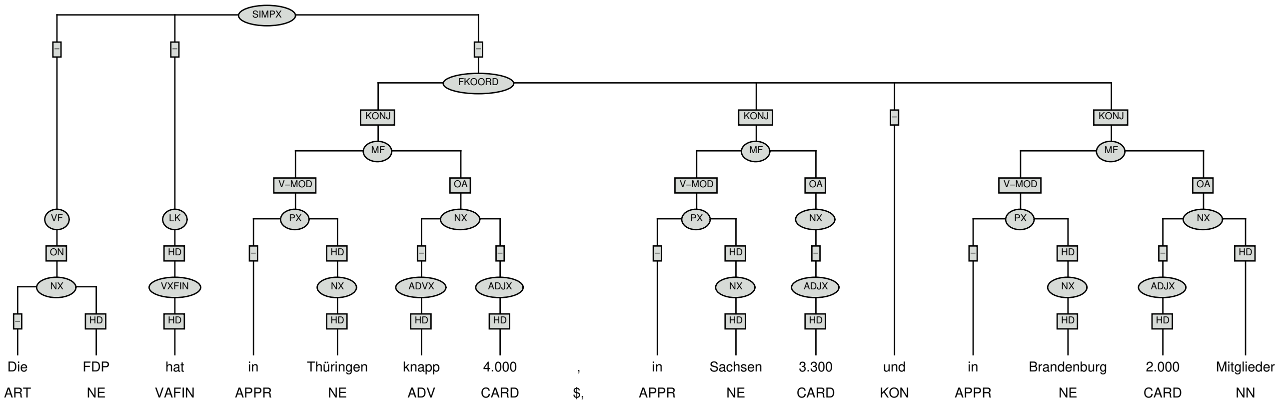
The model of topological fields does not make any assumptions about dependency relations, but it allows that topological fields may be left empty. For the description of elliptical sentence constructions, the scheme of topological fields is an appropriate model because neither crossing branches nor traces have to be used to annotate the surface structure of a sentence.

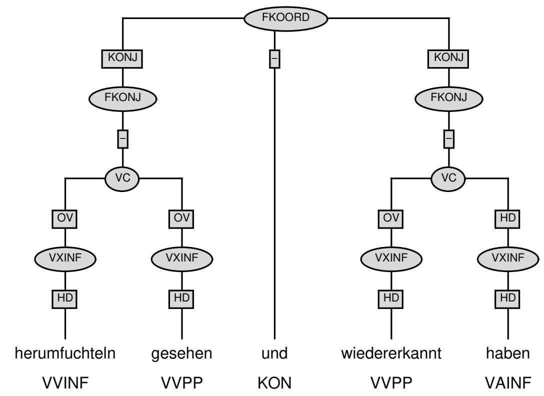
In elliptical phrases, the head noun is missing. They are annotated like phrases without a head. Therefore, the edge labels of an elliptical phrase are empty:


SIMPX
VF
ON
NX
ADJX
HD
LK
VXFIN
MF
OD
PRED
ADJX
ADVX
die irischen Hinweisschilder seien den walisischen ziemlich ähnlich
ART ADJA NN VAFIN ART ADJA ADV ADJD

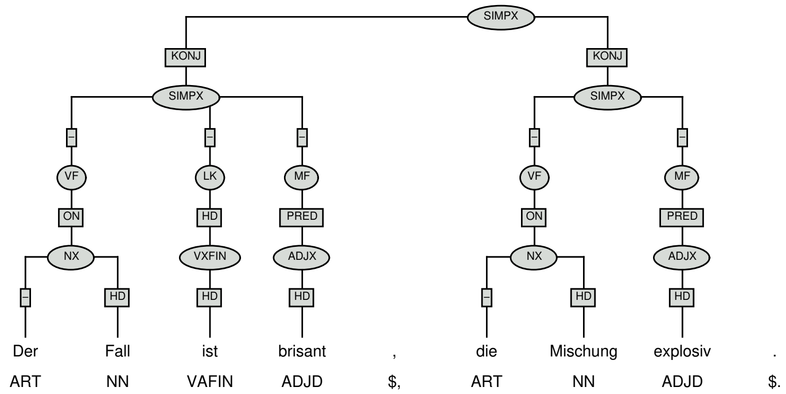

PX
KONJ
HD
NX
in und um Berlin
APPR KON APPR NE

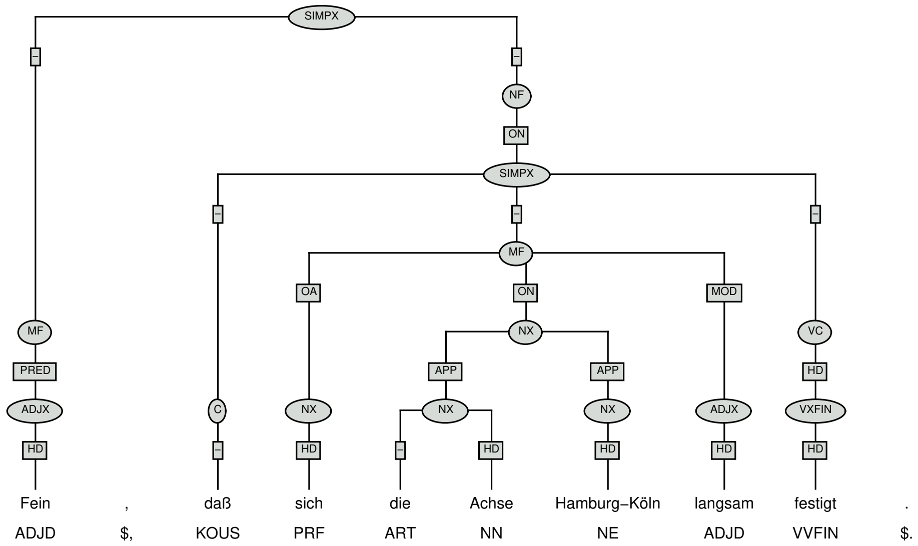

PX
HD
NX
ADJX
vom 15. 4. 99
APPRART ADJA ADJA CARD


SIMPX
FKOORD
KONJ
MF
V–MOD
OA
PX
NX
ADVX
ADJX
VF
ON
LK
HD
VXFIN
Die FDP hat in Thüringen knapp 4.000 , in Sachsen 3.300 und in Brandenburg 2.000 Mitglieder
ART NE VAFIN APPR NE ADV CARD $, APPR NE CARD KON APPR NE CARD NN



In elliptical sentence constructions, specific topological fields are not occupied. All constituents are attached to the appropriate field. In the first example, LK in the second conjunct is missing. In the second example, the subject is in NF and the main clause is lacking a verbal constituent:





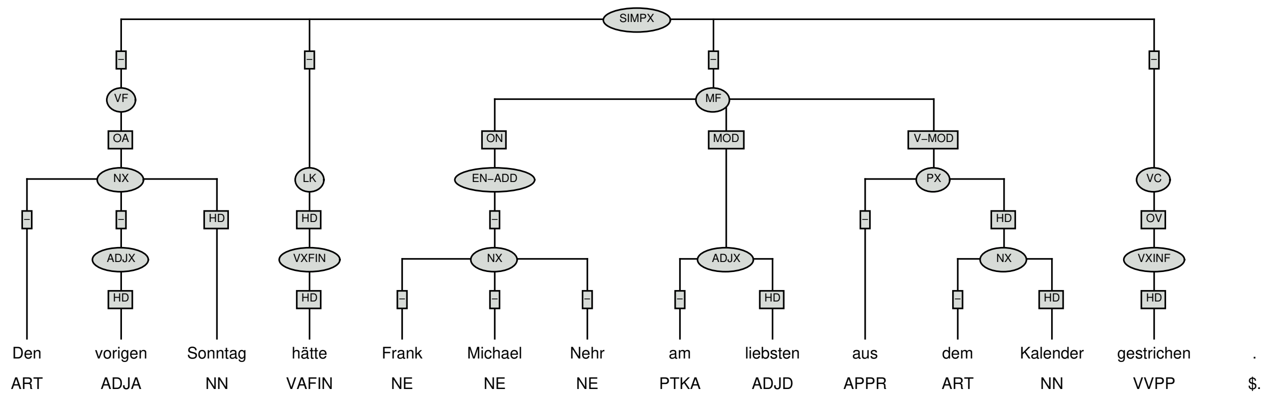
Chapter 7

The Annotation of Specific Syntactic Phenomena

7.1 Superlative and Comparative Forms

7.1.1 Superlative Forms

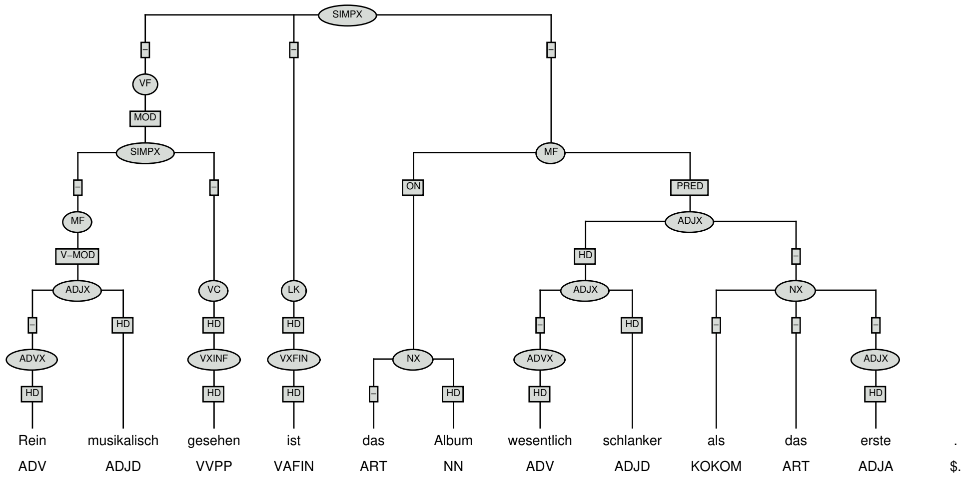
The particle *am*, which occurs as a particle with an adjective or an adverb in superlative constructions, is tagged as PTKA. Both, the particle and the adjective/adverb are attached on the same level forming an adverbial/adjectival phrase:

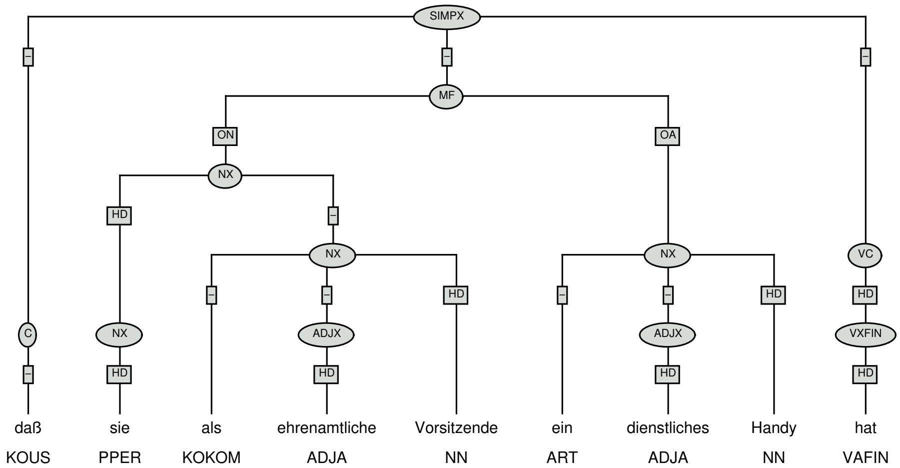


7.1.2 The Comparative Particles *wie* and *als*

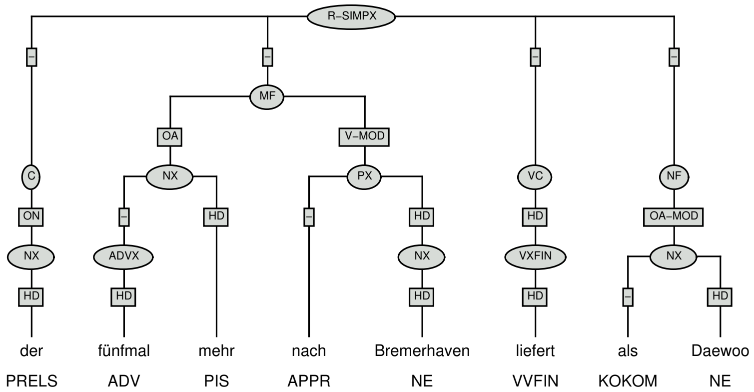
Comparative particles in German are *als* and *wie*, in rare cases also *denn* (e.g. *Die werden dort seliger schlummern denn je.*). These particles are tagged as KOKOM and occur with all types of syntactic phrases (NX, ADVX, PX, etc.). They are directly attached to an adjacent comparative phrase. In case of a comparative phrase with a postmodifier, they are directly attached to the highest node of the complex phrase.

A comparative phrase can occur as an adjacent postmodifier of the head phrase:

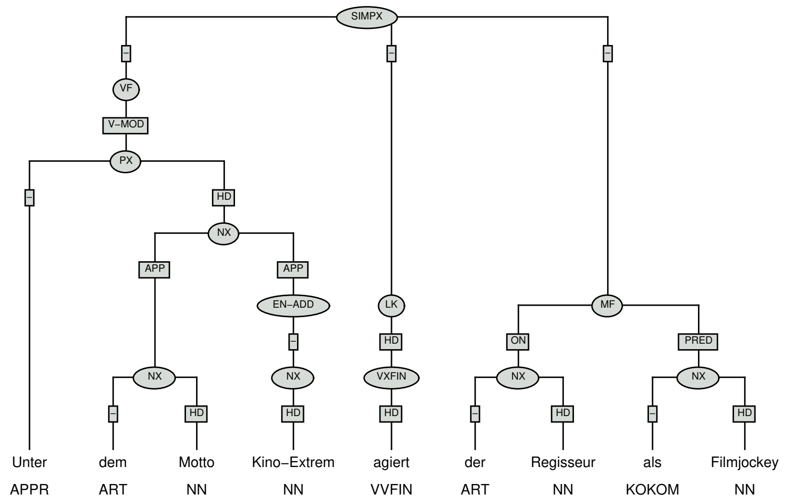


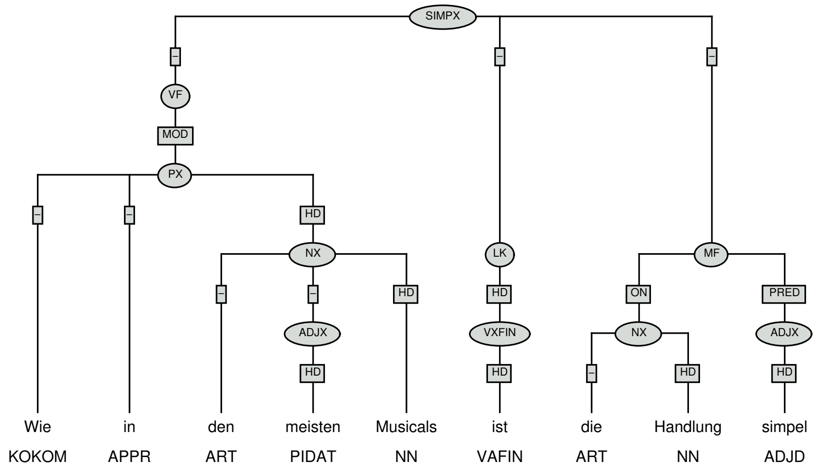


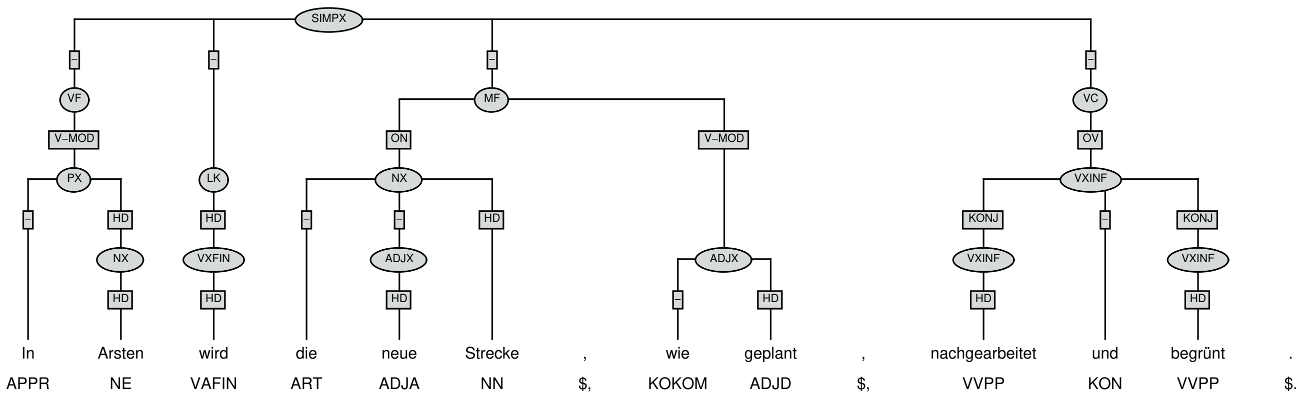
If there is a long-distance dependency between the comparative phrase and the head phrase, the dependency relation is denoted with the respective X-MOD label.



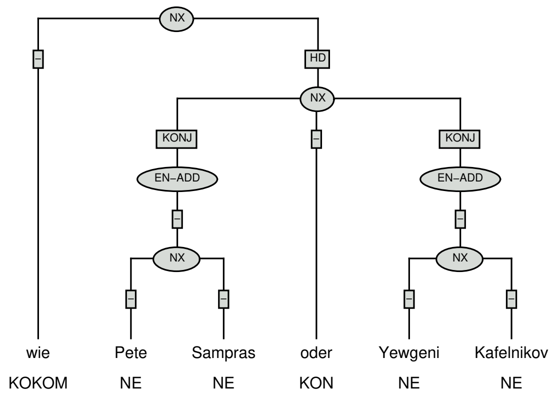
In case of a long-distance dependency between the comparative phrase and the main verb (cf. 4.7.9), the comparative phrase is either a complement (e.g. PRED) or an ambiguous or unambiguous modifier of the main verb (MOD or V-MOD).







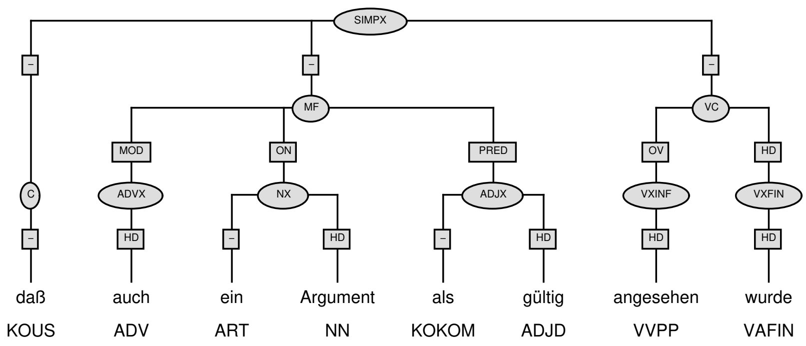
The *high attachment principle* applies when the comparative particle has scope over a coordination of phrases (cf. 6.4.5). In this case, the two conjuncts are coordinated first. Then the particle is attached on a higher level.



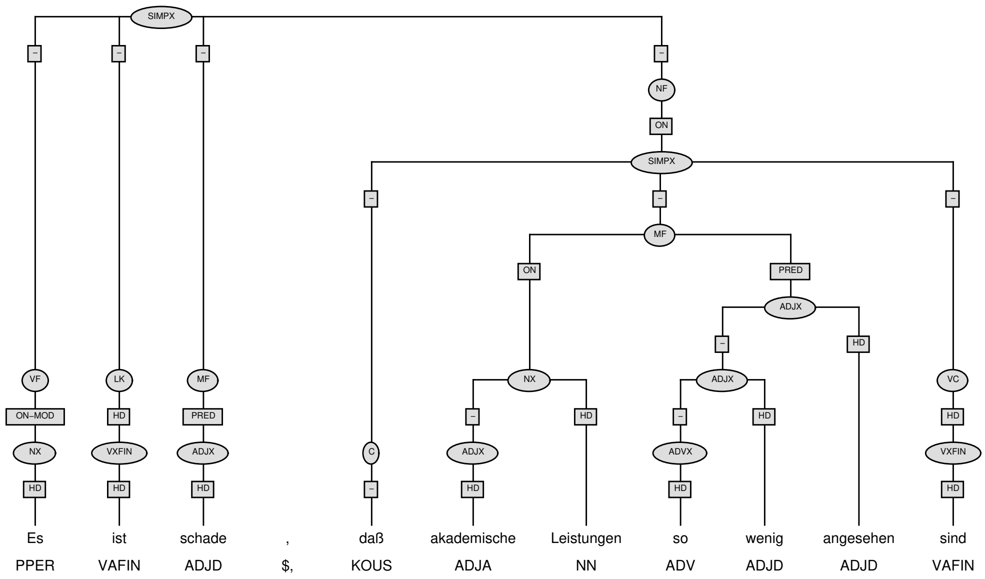
7.2 Verbal and Adjectival Use of Participles

In German, verbal participles which are passive verb forms (*Der Mensch wird angesehen*) can be used as adjectives: either as an attribute adjective (*der angesehene Mensch*) or - depending on the context - also as a predicative adjective (*der Mensch ist angesehen.*). In contrast to the auxiliary *werden* in verbal passives, the auxiliary *sein* is used in constructions with adjectival passives. Concerning the problematic distinction between verbal and adjectival passives, we follow the criteria in the Stuttgart-Tübingen tagset (STTS) (Schiller et al. 1995).[1]

The following two tree structures show the annotation of the verbal and adjectival passives of the verbal participle *angesehen*. In the first example, the verbal participle is a VVPP in VC. In the second example, the verbal participle has an adjectival reading and is annotated as an ADJD in MF.



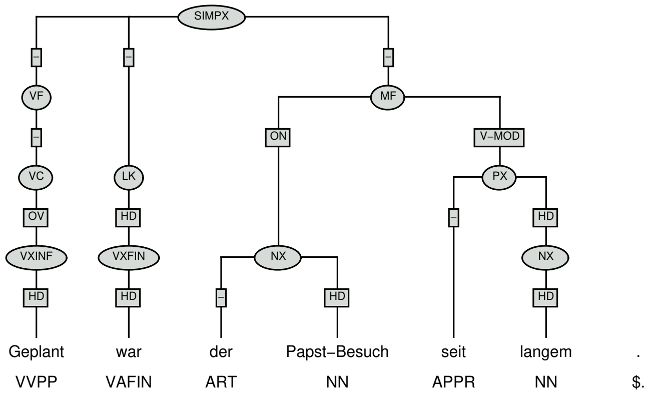
[1]Concerning the differences between verbal and adjectival passives in English cf. Bresnan (1995).

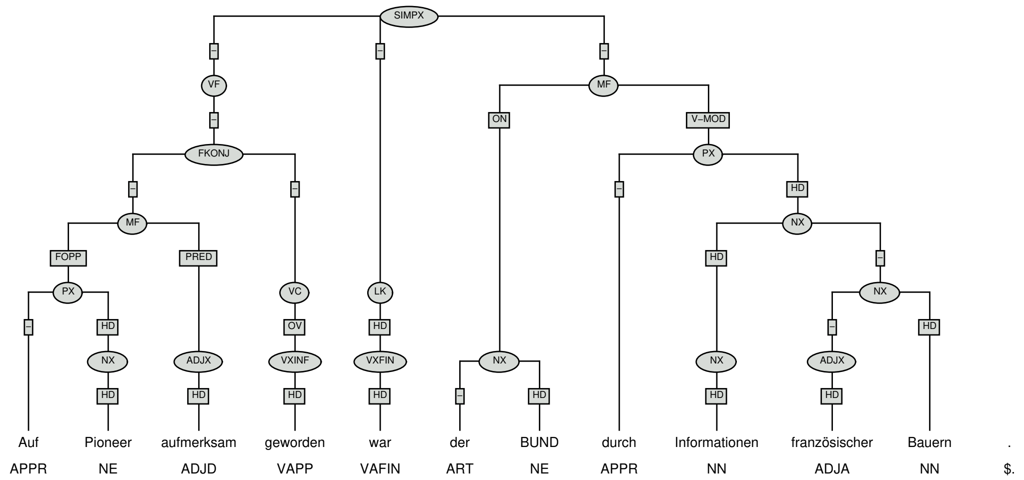


7.3 Topicalization

Topicalization is almost exclusively found in verb-second clauses. As a consequence, the subject is not in the first position of the clause. In topicalized constructions, we are confronted with word order phenomena that differ from those within MF, e.g., infinite parts of VC are not allowed in MF.

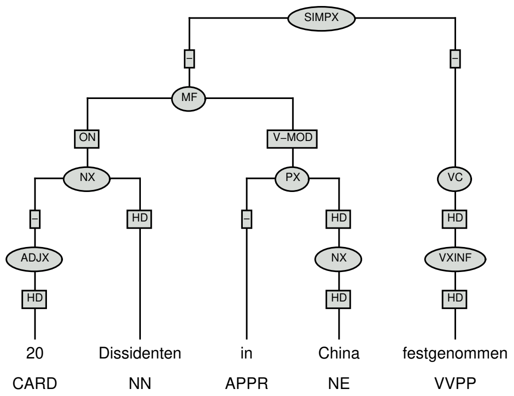
Our annotation strategy is to annotate the topicalized verb complex, respectively the infinite parts of it, as VC in the first position of the clause. VC is then attached to VF. If a part of MF is topicalized together with VC, first MF and VC are combined to form FKONJ before they are attached to VF:



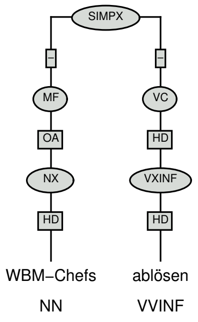


7.4 Headlines

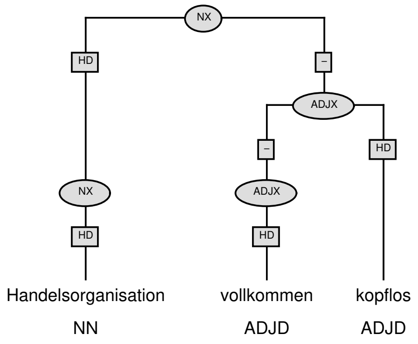
The syntax of headlines differs from other syntactic constructions in so far as headlines[2] often lack the finite verb or a verb at all. If a headline has only an infinitive, the case assigment follows the preference principle formulated in 5.2. Therefore, we assume in general the more plausible grammatical function in each case: a passive constructions with ON in MF if the verb in VC is a past participle and an active construction with OA in MF if the verb in VC is an infinitive.



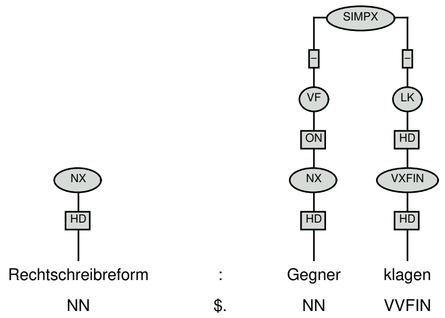
[2]The identifier "HEADLINE" is automatically inserted into the comment line above the sentence for each syntactic unit which is marked as a headline in the original data.



If there is no verb at all within a headline, it is annotated like an isolated phrase (cf. 3.4.5):



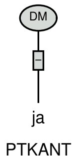
Headlines can also consist of more than one syntactic structure, for instance, separated by a colon or a dash (cf. 4.7.2 and 5.2):



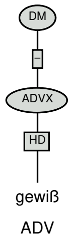
7.5 Discourse Markers

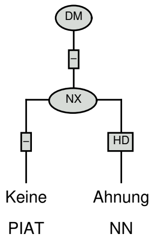
Generally, discourse markers are expressions or phrases of greeting, apologizing, thanking, short emotional utterances, and interjections. Their node label is DM. The edge label of a discourse marker is empty, i.e., it does not have a head. Typical discourse markers are: *ja, nein, hallo, oh, aha, pst, nunja, gewiß, toll, nun ja,* etc.

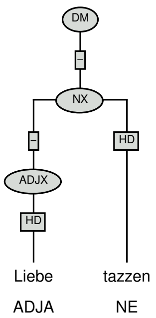
In most cases, discourse markers occur as isolated expressions. Interjections, tagged as ITJ, are directly projected to DM without internal structure. The same applies for answer particles (PTKANT):



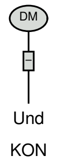
Phrases which function as discourse markers are first projected to their phrase level before they are assigned the node label DM.

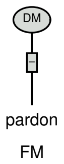




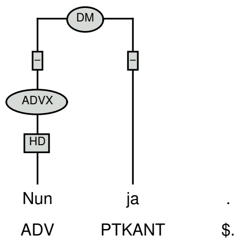


Isolated conjunctions and foreign language discourse markers are tagged according to their part of speech (KON and FM) and are projected to DM:

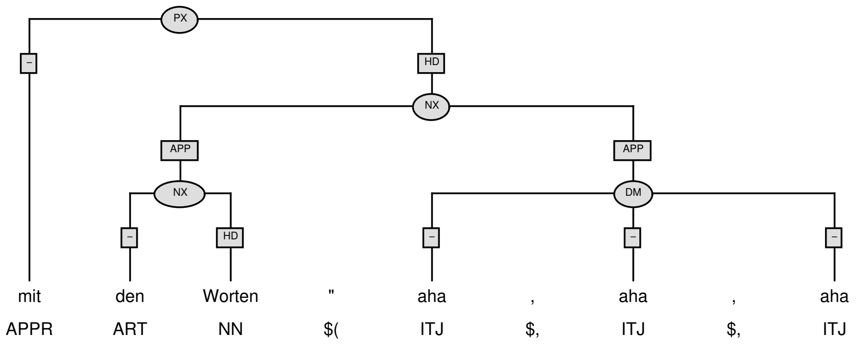


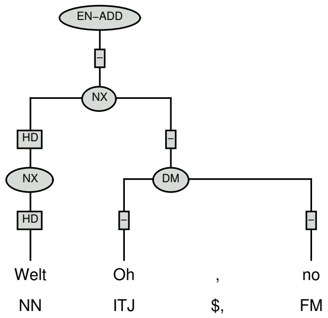


Discourse markers may also consist of an interjection or an answer particle and a phrase:



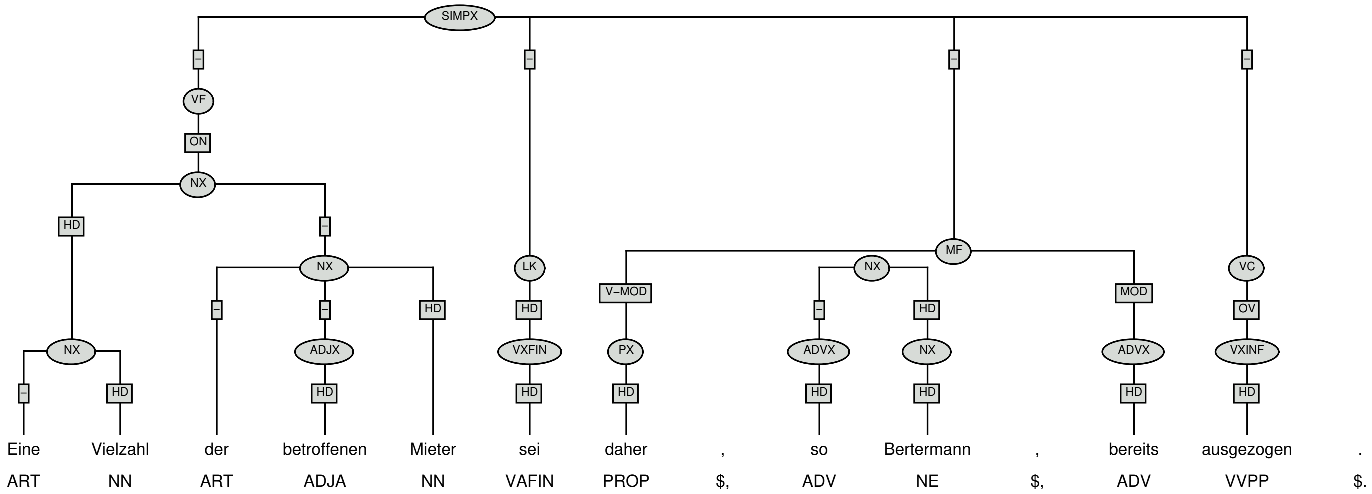
In some cases, discourse markers have a grammatical function within a phrase or a clause. Therefore, they are attached to the syntactic structure:

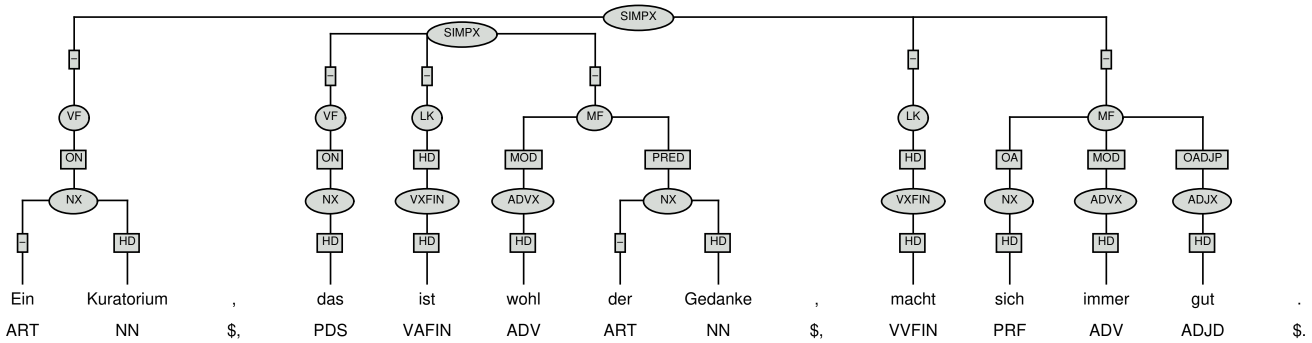




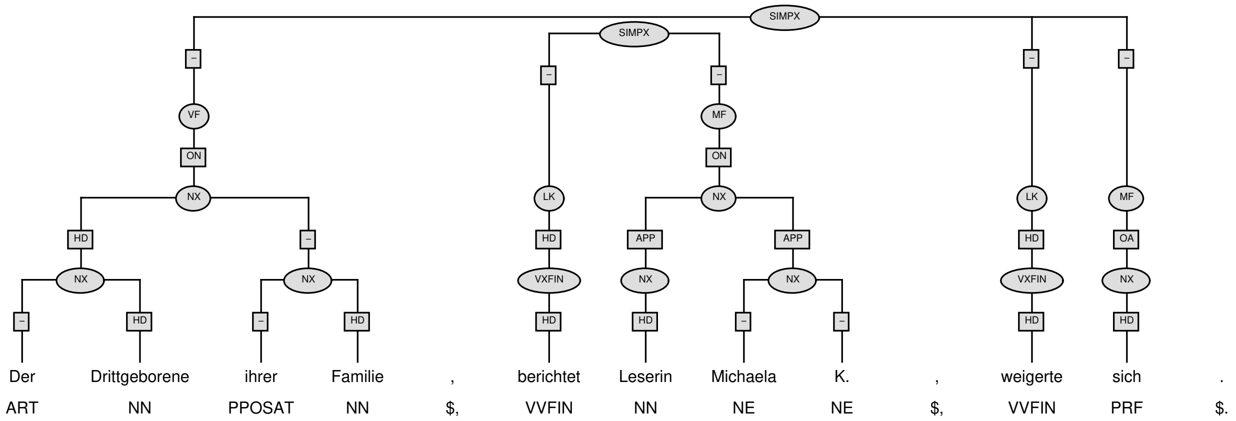
7.6 Parentheses

Parentheses occur as interjective utterances within a sentence. Since there is no dependency relation between the parenthesis and the rest of the construction, the parenthesis is not attached to the surrounding constituents. Often parentheses occur as SIMPX-clauses. Insertions like *berichtet Leserin Michaela K.* into direct speech are also annotated as parenthesis.[3]





[3] On the the TüBa-D/Z web page (`www.sfs.uni-tuebingen.de/en_tuebadz.shtml`), the treebank is also available in the Penn Treebank format. For this format, parentheses are attached to the tree structure with the edge label PAR. For further details about the Penn Treebank format cf. 8.1.3.


SIMPX
SIMPX
VF
MF
LK
MF
ON
ON
OA
NX
NX
NX
NX
NX
NX
NX
VXFIN
VXFIN
HD
APP
APP
Der Drittgeborene ihrer Familie , berichtet Leserin Michaela K. , weigerte sich .
ART NN PPOSAT NN $, VVFIN NN NE NE $, VVFIN PRF $.

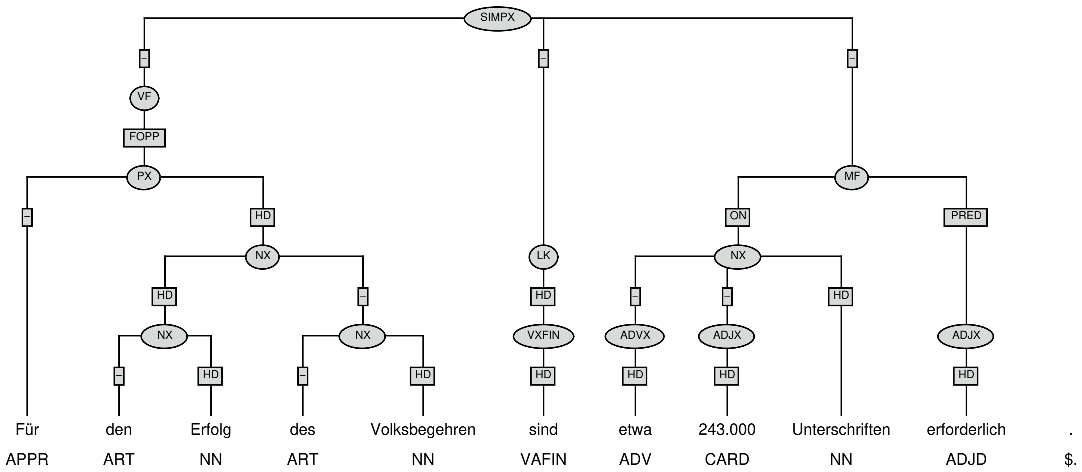
Chapter 8

Criteria for the Distinction of Grammatical Functions

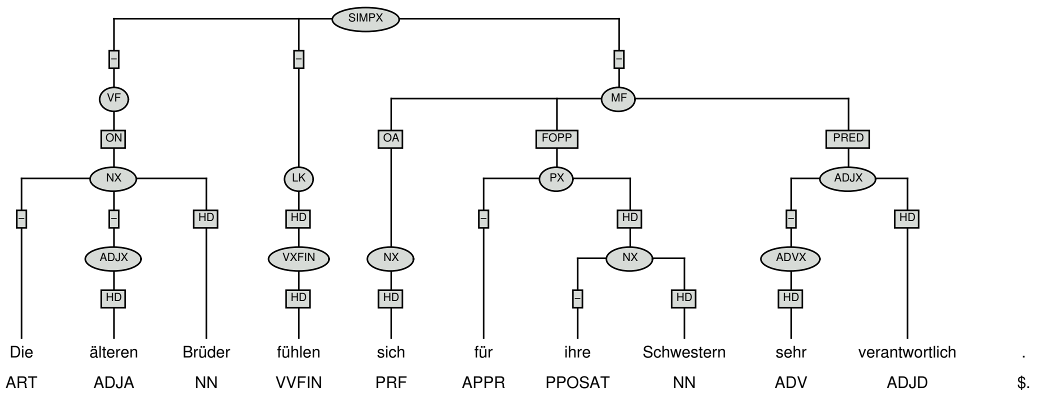
8.1 Subcategorization of Verbs

The *TüBa-D/Z-Verblist* document[1] lists all verbs occurring in the treebank with their specific subcategorization frames. This reference list guarantees the consistent annotation of grammatical functions.

Subcategorization of PREDs:
Since constituents which predicates subcategorize for have grammatical function within a sentence, they are neither marked as PRED-MOD nor attached to the predicate itself. These constituents are attached to a field and assigned the respective grammatical function like the constituent which is marked as FOPP in the following examples:



[1]In case of interest, please refer to web page (`www.sfs.uni-tuebingen.de/en_tuebadz.shtml`) for contact information.



8.1.1 Distinction of 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 *TüBa-D/Z-Verblist* 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, *Sie gehen [OPP gegen die Faschisten] vor./ Das Gesetz ist [OPP in Kraft] getreten.*

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, *Insgesamt berichtet die Polizei [FOPP von 19 Festnahmen und 98 Ingewahrsamnahmen]./Später würden wir [FOPP über Auswandern] nachdenken.* Here, the prepositions select these specific verbs and the PPs cannot be added to any arbitrary verb (which is possible for free adjuncts). In addition, in passive clauses, the subject of the original active clause, which has the form of a prepositional phrase, is marked as FOPP (*Sie wurden [FOPP von Autonomen] umringt.*).

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 (*Nur [V-MOD im griechischen Lager] gab es Probleme*). Typical V-MODs are temporal or local adjuncts specifying time and location of the action, event, or state expressed by the verb.

8.1.2 Distinction of MOD, MOD-MOD, and V-MOD

A typical case of modification of modifiers is a temporal expression (V-MOD) that further specifies another temporal expression (MOD-MOD) in the same clause:

1. *[V-MOD am Samstag] finden [MOD-MOD ab 16 Uhr] Führungen statt.*
   [MOD-MOD Wann] finden [V-MOD am Samstag] Führungen statt?

2. *[MOD da] finden [V-MOD am Samstag] Führungen statt.*
   [V-MOD wann] finden [MOD da] Führungen statt?
   [MOD dann] finden [V-MOD am Samstag] Führungen statt.
   da, dann, etc. can be either temporal, causal, consequential, or local expressions. Thus, one cannot make sure whether the following time expression *am Samstag* 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 VF is MOD/X-MOD, whereas the 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, e.g. *Wenn da was gebucht worden ist, dann ist das nicht in Ordnung.* (cf. 6.1.4).

8.1.3 Distinction of ON, PRED, ON-MOD, and PRED-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 ON and PRED:

1. Typically, PRED occurs in MF, whereas ON occurs in VF of verb-second clauses. This should be considered for annotation, if no other criterion (as described below) applies.

2. 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.

3. If there is a suitable NP that could serve as subject, then this NP is annotated as 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 distinction:

1. Have a look at the constituent in VF. If it is an NP which might serve as subject and if it agrees with the verb, annotate it as ON.

2. If it does not agree with the verb, annotate it as PRED (ADJP, ADVP, PP, etc.).

Examples:

1. *[ON neue Wortschöpfungen] sind [PRED es] nur.*
   [PRED es] sind nur [ON neue Wortschöpfungen].
   oder sind [PRED es] nur [ON neue Wortschöpfungen].
   [PRED das] sind ohnehin [ON die schwächsten Partner].
   [ON die schwächsten Partner] sind [PRED das] ohnehin.
   oder sind [PRED das] ohnehin [ON die schwächsten Partner].

Subject-verb agreement suggests that *neue Wortschöpfungen* und *die schwächsten Partner* are the subject, because of their plural form regardless in which field they occur.

2. *[ON die Ursache] war [PRED unklar].*
   [PRED unklar] war [ON die Ursache]
   [ON Candan Ercettin] ist [PRED überall].
   [PRED überall] ist [ON Candan Ercettin].

   ADJPs and ADVPs typically have PRED function when occurring together with predicative verbs and NP subjects.

3. *[PRED aus den Trauernden] wird [ON ein wütender Mop].*

   ein wütender Mop is considered the subject, because it is a noun phrase. Therefore, the prepositional phrase is PRED.

4. *[ON das] ist [PRED eine einmalige Chance].*
   [ON eine einmalige Chance] ist [PRED das].
   [ON es] ist [PRED der erste Besuch eines Papstes].
   [ON der erste Besuch eines Papstes] ist [PRED es].
   [ON Hauptauftraggeber] ist [PRED die Bremer Verwaltung].
   [ON die Bremer Verwaltung] ist [PRED Hauptauftraggeber].

   The NP in VF position agrees with the verb and therefore has subject priority. As a consequence, the constituent in MF is PRED.

5. *[PRED wer] bin [ON ich].*
   [PRED was] ist [ON das].

   In w-questions, the interrogative pronoun is always PRED because here also the agreement rule applies.

6. *[ON-MOD es] sei [PRED wichtig], [ON daß man ... .*
   [ON Aufgabe des Festspielhauses] sei [PRED-MOD es], [PRED das Haus spielfertig zu halten].

   If a sentential subject or a sentential predicate occurs with an expletive *es*, the expletive *es* is either ON-MOD or PRED-MOD (cf. 4.2.9).

References

Bech, G. 1955-57. *Studien über das deutsche Verbum infinitum. Kopenhagen. 2 Bände. 2. unveränderte auflage 1983 mit einem Vorwort von Catharine Fabricius-Hansen.* Tübingen: Max Niemeyer.

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.

Brants, T. 1997. *The NeGra Export Format for Annotated Corpora.* University of Saarbrücken, Germany.

Brants, T. 1998. *TnT–A Statistical Part-of-Speech Tagger.* Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany.

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.

Drach, E. 1937. *Grundgedanken der Deutschen Satzlehre.* Frankfurt/M.

Drosdowski, G. (Ed.). 1995. *Duden "Die Grammatik der deutschen Gegenwartssprache".* Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Eisenberg, P. 1999-2001. *Grundriß der deutschen Grammatik, Band 2: Der Satz.* Stuttgart, Weimar: J.B. Metzler.

Engel, U. 1996. *Deutsche Grammatik.* Heidelberg: Julius Groos Verlag.

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.

Helbig, G., and J. Buscha. 1998. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Leipzig, 18. Auflage.

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., J. Bartels, Y. Kawata, V. Kordoni, and H. Telljohann. 2000. The Tübingen treebanks for spoken German, English, and Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin: Springer.

Höhle, T. N. 1986. 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.

Kiss, T. 1995. *Infinitive Komplementation. Neue Studien zum deutschen Verbum infinitum.* Tübingen: Max Niemeyer.

Kübler, S., and H. Telljohann. 2002. Towards a dependency-based evaluation for partial parsing. In *Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems – (LREC 2002 Workshop), Las Palmas, Gran Canaria, June 2002.*

Marcus, M., B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

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.

Stegmann, R., H. Telljohann, and E. W. Hinrichs. 2000. Stylebook for the German Treebank in VERBMOBIL. Technical report, Verbmobil-Report 239.

Appendix: The TüBa-D/Z Data Formats

The TüBa-D/Z treebank is released in three different data formats :

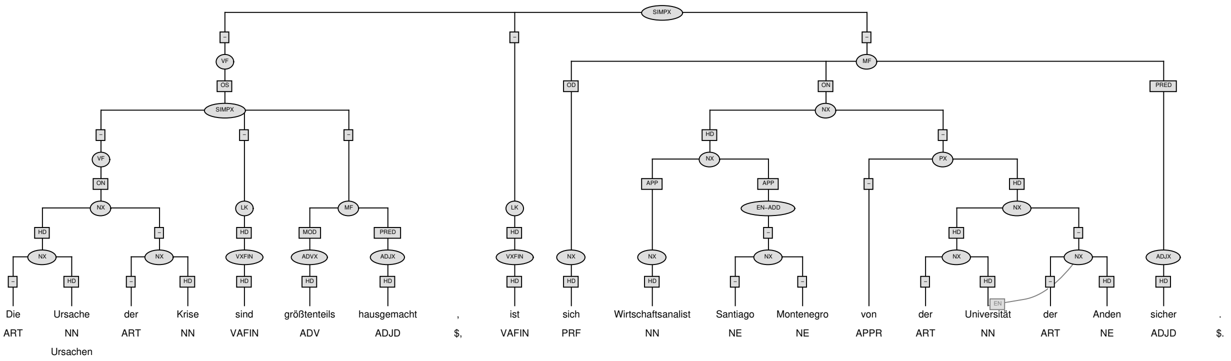
1. the Negra export format,
2. the Penn Treebank format,
3. an XML format.

1. The Negra Export Format

This format is provided by the annotation tool Annotate (Brants and Skut 1998), it is created automatically from the database underlying the annotation process in Annotate. The Negra export format is a line-oriented pointer-based representation of the syntactic annotation. It is also the most complete data format since it preserves all the information available during the manual annotation. A more complete description of the negra export format can be found in (Brants 1997).

An example of the Negra export format is given below, combined with the graphical representation of the syntactic annotation for the sentence "Die Ursache der Krise sind größtenteils hausgemacht, ist sich Wirtschaftsanalist Santiago Montenegro von der Universität der Anden sicher.".

Graphical representation:



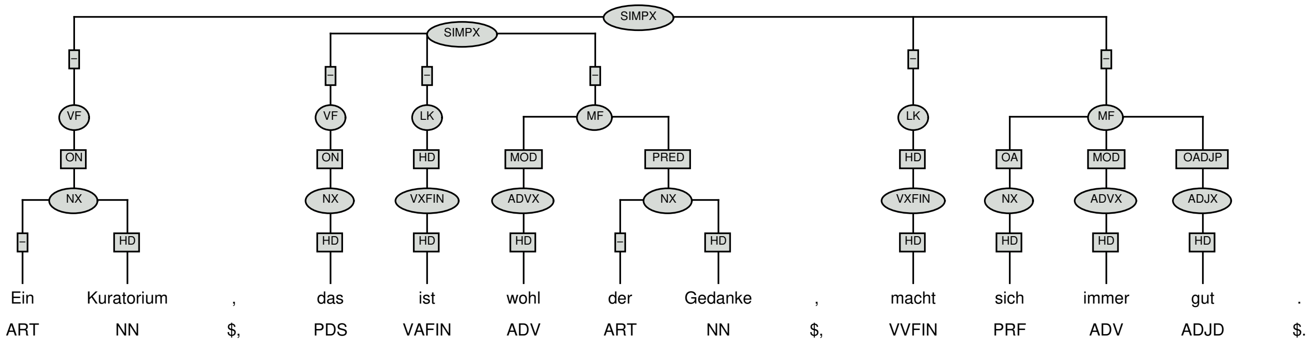
Export format:

```
#BOS 8268 1 1071659582 46
Die                     ART     --              -       500
Ursache                 NN      --              HD      500     %% Ursachen
der                     ART     --              -       501
Krise                   NN      --              HD      501
sind                    VAFIN   --              HD      502
größtenteils            ADV     --              HD      503
hausgemacht             ADJD    --              HD      504
,                       $,      --              --      0
ist                     VAFIN   --              HD      505
sich                    PRF     --              HD      506
Wirtschaftsanalist      NN      --              HD      507
Santiago                NE      --              -       508
Montenegro              NE      --              -       508
von                     APPR    --              -       520
der                     ART     --              -       509
Universität             NN      --              HD      509
der                     ART     --              -       510
Anden                   NE      --              HD      510
sicher                  ADJD    --              HD      511
.                       $.      --              --      0
#500                    NX      --              HD      512
#501                    NX      --              -       512
#502                    VXFIN   --              HD      513
#503                    ADVX    --              MOD     514
#504                    ADJX    --              PRED    514
#505                    VXFIN   --              HD      515
#506                    NX      --              OD      524
#507                    NX      --              APP     519
#508                    NX      --              -       516
#509                    NX      --              HD      517
#510                    NX      --              -       517     EN      15
#511                    ADJX    --              PRED    524
#512                    NX      --              ON      518
#513                    LK      --              -       521
#514                    MF      --              -       521
#515                    LK      --              -       525
#516                    EN-ADD  --              APP     519
#517                    NX      --              HD      520
#518                    VF      --              -       521
#519                    NX      --              HD      522
#520                    PX      --              -       522
#521                    SIMPX   --              OS      523
#522                    NX      --              ON      524
#523                    VF      --              -       525
#524                    MF      --              -       525
#525                    SIMPX   --              --      0
#EOS 8268
```

It is evident that information about misspellings and secondary edges as well as about the origin of a file, the identity of the last annotator, and the time of the last modification is present in the export format.

The only deviation from context-freeness which the annotation scheme allows concerns the annotation of parentheses. Parentheses are annotated as separate trees with no attachment to surrounding trees. The following tree gives an example for such a phenomenon (for a more complete description of the annotation cf. 7.6).

Graphical representation:



The pointer-based representation of the Negra export format separates information about the linear precedence of words from attachment information so that parentheses can be represented naturally without having to resort to explicitely marking non-attached nodes. Here, the SIMPX node dominating the parenthesis (node #515) is marked as not having a mother node.

Export format:

```
#BOS 9292 1 1068132102 93
Ein                    ART     --               -       500
Kuratorium             NN      --               HD      500
,                      $,      --               --      0
das                    PDS     --               HD      501
ist                    VAFIN   --               HD      502
wohl                   ADV     --               HD      503
der                    ART     --               -       504
Gedanke                NN      --               HD      504
,                      $,      --               --      0
macht                  VVFIN   --               HD      505
sich                   PRF     --               HD      506
immer                  ADV     --               HD      507
gut                    ADJD    --               HD      508
.                      $.      --               --      0
#500                   NX      --               ON      509
#501                   NX      --               ON      510
#502                   VXFIN   --               HD      511
#503                   ADVX    --               MOD     512
#504                   NX      --               PRED    512
#505                   VXFIN   --               HD      513
#506                   NX      --               OA      514
#507                   ADVX    --               MOD     514
#508                   ADJX    --               OADJP   514
#509                   VF      --               -       516
#510                   VF      --               -       515
#511                   LK      --               -       515
#512                   MF      --               -       515
#513                   LK      --               -       516
#514                   MF      --               -       516
#515                   SIMPX   --               --      0
#516                   SIMPX   --               --      0
#EOS 9292
```

2. The Penn Treebank Format

This format is based on the format of the Penn Treebank (Marcus et al. 1993). The attachment of constituents is shown via bracketing and indentation. Thus, all constituents which show the same level of indentation are attached on the same level. In the Penn Treebank format, grammatical functions, which are shown in the Negra export format in the column "edge label", are attached to the sntactic label via a colon. Thus, the label "NX:OA" means that the constituent is a noun phrase with the grammatical function accusative object.

The Penn Treebank format is a representation that combines the linear representation of words with their attachment to higher constituents. For this reason, this format is restricted to completely context-free tree structures, i.e. it cannot adequately represent the annotation of parentheses in TüBa-D/Z. In order to capture the original syntactic annotation as well as the original word order in the sentence, it was decided to introduce a new edge label to mark such cases: PAR. Thus, the sentence "Ein Kuratorium, das ist wohl der Gedanke, macht sich immer gut.", as shown above is represented in the Penn Treebank format by the following bracketed structure:

```
%% sent. no. 9292
(
  (SIMPX
    (VF
      (NX:ON
        (ART Ein)
        (NN:HD Kuratorium)
      )
    )
  ($, ,)
    (SIMPX:APP                      %% here starts the parenthesis!
      (VF
        (NX:ON
          (PDS:HD das)
        )
      )
      (LK
        (VXFIN:HD
          (VAFIN:HD ist)
        )
      )
      (MF
        (ADVX:MOD
          (ADV:HD wohl)
        )
        (NX:PRED
          (ART der)
          (NN:HD Gedanke)
        )
      )
    )
  ($, ,)
    (LK
      (VXFIN:HD
        (VVFIN:HD macht)
      )
    )
    (MF
      (NX:OA
        (PRF:HD sich)
      )
      (ADVX:MOD
        (ADV:HD immer)
      )
      (ADJX:OADJP
        (ADJD:HD gut)
      )
    )
  )
  ($. .)
)
```

Comments are preceded by a double '%' sign. The comment behind the structure is intended to help the reader locate the beginning of the parenthesis, it is not part of the actual data.

Commata, which are not attached to the tree, are indented on the highest level although they are included in the bracketing of the constituent surrounding them. In the sentence below, e.g., the first comma is grouped into the noun phrase NX via word order. The indentation, however, signals that the comma cannot necessarily be attached to this node. It is also conceivable that it may be attached to one of the lower nodes, NX or R-SIMPX. In the case of the second comma, there are even more possible attachment sites.

```
%% sent. no. 33
(
  (R-SIMPX
    (C
      (NX:ON
        (PRELS:HD die)
      )
    )
    (MF
      (NX:OA
        (NX:HD
          (ART die)
          (EN-ADD:HD
            (NN AWO)
          )
        )
  ($, ,)
        (R-SIMPX
          (C
            (PX:V-MOD
              (PWAV:HD wo)
            )
          )
          (MF
            (NX:ON
              (PPER:HD er)
            )
            (NX:PRED
              (NN:HD Kreisvorsitzender)
            )
          )
          (VC
            (VXFIN:HD
              (VAFIN:HD ist)
            )
          )
        )
      )
    )
  ($, ,)
    (VC
      (VXFIN:HD
        (VVFIN:HD prüfte)
      )
    )
  )
  ($. .)
)
```

3. The XML Format

The XML format is a custom-made XML format that follows the Negra export file format. It is designed to accomodate all original information provided in the export format, including e.g. comments and editor/origin information, which are resolved so that it is not necessary to consult the Negra tables. Dominance relations between nodes are represented directly within the XML tree structure. Root nodes are marked by the attribute parent="0". Thus, it is possible to represent parentheses without the use of additional labels. A more detailed description of the XML structure can be found at the web page `http://www.sfs.uni-tuebingen.de/dereko/anno-doc.html`.

The following example shows the XML structure for the sentence "Ein Kuratorium, das ist wohl der Gedanke, macht sich immer gut.". The root node of the parenthesis contains the attribute parent="0", which signifies that while the tree is part of the higher constituent where linear order is concerned, it is not attached to the surrounding tree.

```
<sentence editor="hschulz" date="2003110616:21:42" origin="T990503.140" comment="">
  <node cat="SIMPX" func="--" parent="0" comment="">
    <node cat="VF" func="-" comment="">
      <node cat="NX" func="ON" comment="">
        <word form="Ein" pos="ART" morph="--" func="-" comment=""/>
        <word form="Kuratorium" pos="NN" morph="--" func="HD" comment=""/>
      </node>
    </node>
    <word form="," pos="$," morph="--" func="--" parent="0" comment=""/>
    <node cat="SIMPX" func="--" parent="0" comment="">
      <node cat="VF" func="-" comment="">
        <node cat="NX" func="ON" comment="">
          <word form="das" pos="PDS" morph="--" func="HD" comment=""/>
        </node>
      </node>
      <node cat="LK" func="-" comment="">
        <node cat="VXFIN" func="HD" comment="">
          <word form="ist" pos="VAFIN" morph="--" func="HD" comment=""/>
        </node>
      </node>
      <node cat="MF" func="-" comment="">
        <node cat="ADVX" func="MOD" comment="">
          <word form="wohl" pos="ADV" morph="--" func="HD" comment=""/>
        </node>
        <node cat="NX" func="PRED" comment="">
          <word form="der" pos="ART" morph="--" func="-" comment=""/>
          <word form="Gedanke" pos="NN" morph="--" func="HD" comment=""/>
        </node>
      </node>
    </node>
    <word form="," pos="$," morph="--" func="--" parent="0" comment=""/>
    <node cat="LK" func="-" comment="">
      <node cat="VXFIN" func="HD" comment="">
        <word form="macht" pos="VVFIN" morph="--" func="HD" comment=""/>
      </node>
    </node>
    <node cat="MF" func="-" comment="">
      <node cat="NX" func="OA" comment="">
        <word form="sich" pos="PRF" morph="--" func="HD" comment=""/>
      </node>
      <node cat="ADVX" func="MOD" comment="">
        <word form="immer" pos="ADV" morph="--" func="HD" comment=""/>
      </node>
      <node cat="ADJX" func="OADJP" comment="">
        <word form="gut" pos="ADJD" morph="--" func="HD" comment=""/>
      </node>
    </node>
  </node>
  <word form="." pos="$." morph="--" func="--" parent="0" comment=""/>
</sentence>
```

Index