

Stylebook for the English Treebank in VERBMOBIL

Valia Kordoni

Seminar für Sprachwissenschaft
Universität Tübingen

September 2000

Valia Kordoni

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

Tel.: 07071 - 29 74279

Fax: 07071 - 55 13 35

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

Gehört zum Antragsabschnitt: 6.7 Baumbanken

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

Abstract

This report describes the design principles and the annotation scheme for the VERBMOBIL treebank of English developed at the Eberhard-Karls-Universität Tübingen. It is intended as a stylebook for the construction or use of treebank data for English. The guidelines focus on the syntactic annotation of spoken language data with its characteristics (e.g. repetitions, hesitations, “false starts”, and so forth).

Acknowledgements

We would like to thank Prof. Dr. E. W. Hinrichs, Prof. Carl Pollard, Paul C. Davis, and last but not least Shравan Vasishth for their helpful comments and support in form of actual work for the English treebank, encouraging, and critical discussions, from which we could strongly benefit for the challenging task of developing an annotation scheme for an amount of 30 000 entries.

We would also like to thank our Tübingen colleagues Dale Gerdemann, John Griffith, Sandra Kübler, and Uli Schatz for their assistance with the Part-of-Speech Tagging of the data and with data conversion.

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

For assistance with the Part-of-Speech Tagging and data conversion with the transcribed Verbmobil data we are indebted to our Verbmobil colleagues at Siemens (Munich), particularly to Tobias Ruland, and at the IMS Stuttgart, particularly to Martin Emele.

*Verb*mobil Report 241

Contents

1	Introduction	1
2	Background Assumptions	3
2.1	Spontaneous Speech	3
3	The Theoretical Basis of the Annotation Scheme	5
3.1	General Annotation Principles	6
3.1.1	Flat Clustering Principle	6
3.1.2	Longest Match Principle	6
3.1.3	High Attachment Principle	6
3.2	The Structure of an Annotated Tree	7
3.2.1	The Levels of Annotation	7
3.2.2	Where Does a Tree End?	9
3.2.3	Speech Errors and Repetitions	10
3.2.4	Isolated Phrases and Sentence Fragments	12
3.2.5	Empty Categories and Crossing Branches	15
3.2.6	Empty Edge Labels	17
4	The Annotation of Phrases	19
4.1	Internal Structure of Phrases	19
4.1.1	Noun Phrases (NPs)	19
4.1.2	Determiner Phrases	23
4.1.3	Degree Phrases	24
4.1.4	Prepositional Phrases	27
4.1.5	Adjectival Phrases	31
4.1.6	Adverbial Phrases	32
4.1.7	Verb Phrases	32
5	At the Sentential Level	39
5.1	Combining Phrases into Sentences	39

5.1.1	Grammatical Functions	41
5.1.2	Head of the Sentence	47
5.2	Relative Clauses	50
5.3	Copula Constructions	51
5.4	SUGG(estion)s	54
5.5	The Periphery of the Sentence	56
5.6	Coordination	56
5.6.1	Isolated Conjuncts	59
5.6.2	Unequal Conjuncts	60
5.7	Parentheses	61
5.8	Discourse Markers	62
6	Conclusion	65
	References	66
	Appendix: The Labels Used in the Treebank	68

Chapter 1

Introduction

This report describes the design principles and the annotation scheme for a treebank of English that has been developed at the Eberhard-Karls-Universität Tübingen as part of the Verbmobil project.

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

The overriding goal of the Verbmobil project was to develop a speaker-independent system for translating spontaneous speech. The English treebank described here provides linguistic annotations for the Verbmobil-I and Verbmobil-II dialogue corpus of spontaneous speech in the scenarios of appointment negotiations, travel arrangements and personal computer maintenance. In order to obtain realistic and quantitatively significant data for the relevant scenarios, the Verbmobil project launched a major data collection initiative for spoken-language dialogues. The dialogues were recorded in a variety of settings and were transcribed according to mutually agreed upon standards. The transcribed data were then further annotated for the purposes of signal processing and linguistic analysis.

The treebank project, carried out by the Division of Computational Linguistics at the Eberhard-Karls-Universität Tübingen (Lehrstuhl Prof. Hinrichs), constituted part of the overall effort of linguistic annotation within the Verbmobil project. Apart from the English treebank described in the report at hand, treebanks for German and Japanese have also been developed. The parallel development of the Tübingen treebank for German is described in (Stegmann et al. 2000), the one for Japanese in (Kawata and Bartels 2000).

The size of the English treebank has reached 30 000 fully annotated trees at

the conclusion of the Verbmobil-II project phase. The coverage of the English treebank includes the complete set of dialogues that were collected during the Verbmobil-I and Verbmobil-II project phases. The overall annotation scheme for the English treebank was negotiated with all the relevant partners in the Verbmobil-II consortium.

The linguistic annotations of the English treebank pertain to the levels of morpho-syntax (part-of-speech tagging), syntactic phrase structure and function-argument structure. In the Verbmobil context, the treebanks were utilized as a data resource for a variety of processing modules in the Verbmobil system, including as a training data of stochastic parsers, as an on-line resource for the tree construction algorithm of the chunk parser (cf., (Hinrichs et al. 2000b)), and as a resource for the development of semantic construction rules and translation transfer rules. In order to ensure reusability of the data for purposes beyond the Verbmobil project, the treebank annotations follow accepted guidelines for corpus annotation (cf., (Leech 1993)).

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

Treebank annotation greatly benefitted from the use of the annotation tool *Annotate V2.3* ((Brants and Skut 1998), and (Plaehn 1998)). The probabilistic parser included in the *Annotate* package allows semi-automatic treebank construction in conjunction with manual annotation. This semi-automatic mode of annotation greatly facilitates annotation consistency, which is of utmost importance in producing a large-scale language data resource. In addition, the tool provides human annotators with a graphical, user-friendly interface for annotating and editing trees and with database support for maintaining large treebanks. The sample trees contained in this report are all rendered in the *Annotate* interface format.

Chapter 2

Background Assumptions

In order to ensure high-quality linguistic annotations, two prerequisites are indispensable:

1. **linguistically adequate annotation schemes:** the inventory of classificatory labels at each level of annotation needs to be based on sound linguistic principles. For the Verbmobil English treebank, the linguistic annotations were based on existing, widely accepted labelling schemes;
2. **consistency of annotation:** the annotation scheme has to be made as explicit as possible to all human annotators that participate in the treebank annotations. To this end, complementary to the present report, detailed stylebooks (cf., (Stegmann et al. 2000), (Kawata and Bartels 2000)) were developed at the outset of the project for the Verbmobil German and Japanese treebanks, respectively. Consistency was further aided by automatic consistency checks that were conducted at regular intervals throughout the annotation phase and by the semi-automatic annotation mode supplied by the *Annotate* tool.

2.1 Spontaneous Speech

As has already been mentioned above, the Tübingen English treebank is based exclusively on spontaneous speech data. The focus on spoken language immediately raises a number of research questions that do not arise when the input data is taken from newspaper corpora or other sources of written data.

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

of syntactic analysis. For corpora of spontaneous speech utterances, though, such an immediate fit does not exist. In the case at hand, i.e., the corpora of Verbmobil spoken language dialogues, the primary segmentation is that of the dialogue turn, which

1. consists of a single, typically uninterrupted contribution to the dialogue by one of the dialogue participants, and
2. exhibits all the properties characteristic of spontaneous speech, which include speech errors, fragmentary utterances, false starts, repetitions, interruptions, and hesitation noises.

The dialogue turns are automatically preprocessed into units, delimited by full stops and question marks, thus forming a secondary domain of analysis. These units themselves may consist of one or more sentences in the grammatical sense, and/or phrases.

At the annotation level, the tree boundaries are defined by the so-called “longest match” principle, which requires that the tree structure includes as many constituents as possible, provided that the sentence/tree remains syntactically, as well as semantically, well-formed (cf., also Section (3.1) below).

Speech errors, repetitions, corrections, and “hesitations” are structured as much as possible (mostly up to the level of phrasal categories), but are not typically connected to surrounding constituents as a whole.

Chapter 3

The Theoretical Basis of the Annotation Scheme

The annotation scheme for the English treebank is HPSG-oriented¹, in accordance with the HPSG grammar of English for use in Verbmobil developed in the CSLI LinGO (Linguistic Grammars Online) Project (see (Flickinger et al. 2000)). The main advantage of such an annotation scheme is that it provides both fine-grained syntax and fine-grained semantics that are necessary in order to treat the challenging phenomena of the spoken English data in Verbmobil.

The output of the annotation scheme we have developed is relatively flat. At the same time, these flat representations are supplemented by predicate-argument annotations that obviate the need for expressing such dependencies in terms of tree configurations. Such enriched flat structures have a number of advantages, resulting in a simpler and more systematic treatment of sentential adverbs, coordination, scope ambiguities, and long-distance dependencies. A second advantage of the modular encoding of syntactic and predicate-argument dependencies is that it is easy to acquire and use for human annotators. As a result, the error rates of incorrect or inconsistent annotations are reduced.

Coming to the scheme in question itself, the following general observations are due here:

- The S(entence) node of the tree is considered to be the root node of the whole sentence.
- One level below the sentence node, the grammatical functions of phrases are represented by means of edge labels.

¹This is clear at the level of edge labels (grammatical functions).

- One level below the grammatical functions level, the syntactic categories of phrases are represented by means of node labels.

3.1 General Annotation Principles

In every annotation project two questions have to be resolved:

1. how contoured should the tree structures be, and
2. how are syntactic and semantic ambiguities, in particular attachment ambiguities, resolved.

To resolve these questions in a principled way for the English treebank we adopted the following three annotation principles: the *flat clustering principle*, the *longest match principle*, and the *high attachment principle*.

3.1.1 Flat Clustering Principle

The *flat clustering principle* dictates that the number of hierarchy levels in a syntactic structure should be as small as possible; that means that as many constituents as possible are clustered on the same level in order to form a higher order constituent. As a consequence, any degree of branching is allowed.

3.1.2 Longest Match Principle

The *longest match principle* demands that as many constituents as possible be included into a syntactic structure, as long as the whole structure remains syntactically, as well as semantically, well-formed.

3.1.3 High Attachment Principle

The *high attachment principle* dictates that in cases of syntactic and semantic ambiguity in the attachment of modifiers these should be attached to the highest possible level in the tree.

Table 3.1: Phrasal categories.

Phrasal Categories	Description
AP	adjectival phrase
APS	adjectival phrase heading a small clause
ADVP	adverbial phrase
DGP	degree phrase
DTP	determiner phrase
NP	noun phrase
NPS	noun phrase heading a small clause
PP	prepositional phrase
PPS	prepositional phrase heading a small clause
VP	verb phrase

3.2 The Structure of an Annotated Tree

3.2.1 The Levels of Annotation

3.2.1.1 What is a Well-Formed Tree in the English treebank?

Bottom-up viewed, the trees in the English treebank consist of the following levels of representation:

1. Part-of-Speech (POS) tags. The tagset used in the Verbmobil English treebank is that of the Penn treebank (cf., (Santorini 1990)).
2. Phrasal Categories, indicated by means of node labels.
3. Grammatical functions, indicated by means of edge labels, and
4. Root Labels (see below).

3.2.1.1.1 Phrasal Categories. Table 3.1 shows the list of phrasal categories used in the English treebank.

Taking a closer look at the phrasal categories employed in the English treebank, one can easily detect that

1. Determiner Phrases (DTPs) can be further distinguished into *definite articles* (DT-ARTS), *demonstrative determiners* (DT-DMS), *quantifiers* (DT-QNTS), and *interrogative determiners* (DT-WH).

2. Moreover, Degree Phrases (DGPS) can be further defined as *non-wh*-DGPS (that is, DGs), and *wh*-DGPS (that is, DG-WHS).
3. In addition, numbers are further distinguished in *cardinal numbers* (CNUMS), and *ordinal numbers* (ONUMS).
4. Noun phrases can also subsume *demonstrative pronouns* (PR-DM), *interrogative pronouns* (PR-WH), and *relative pronouns* (PR-R).
5. Finally, verbs are further distinguished in *gerunds* (V-GS), *present participles* (V-PRPS), and *passive participles* (V-PSS).

There are two reasons why these intermediate node labels between the POS tags' level of representation and the Phrasal Categories' level of representation exist in the English treebank:

1. The first and most important reason is to disambiguate the Penn treebank POS tagset used in the English treebank, which in all the cases mentioned above is too general, and thus unable to capture the fine-grained syntactic and semantic analysis we opt for in the Verbmobil English treebank.
2. One additional, though longer term, aim is that the disambiguation at this intermediate level of representation can lead to a new tagset.

3.2.1.1.2 Grammatical Functions. Table 3.2 shows the list of grammatical functions employed in the English treebank.

As mentioned above, grammatical functions are technically represented by means of edge labels. The only empty edge labels in the English treebank are the edge labels found below determiners, degree words, complementizers (that is, CMPS, and CMP-WHS), some pronouns, and conjunctions (that is, CNJS). Note that these have intentionally been left empty because of lack of any semantic content at the level of syntactic analysis, where the empty edge labels occur.

3.2.1.1.3 Root Labels. In the English treebank, three different root labels are used:

1. Sentence: S.
2. Suggestion: SUGG, for utterances starting with *how about*, *what about*, and *what if*.
3. Sentence external, isolated phrases occurring without verb: NP, VP, PP, ADVP, and so forth.

Table 3.2: Grammatical functions.

Grammatical Functions	Description
HD	head
COMP	complement
SPR	specifier
SBJ	subject
SBQ	subject,WH
SBR	subject,REL
ADJ	adjunct
ADJ?	adjunct? (serves ambiguous attachment of modifiers)
FLL	filler
FLQ	filler,WH
FLR	filler,REL
MRK	marker
-	intentionally empty edge label.

3.2.2 Where Does a Tree End?

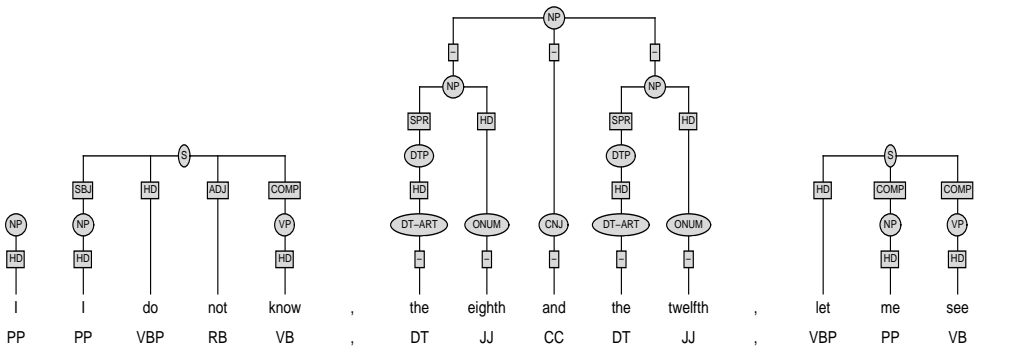
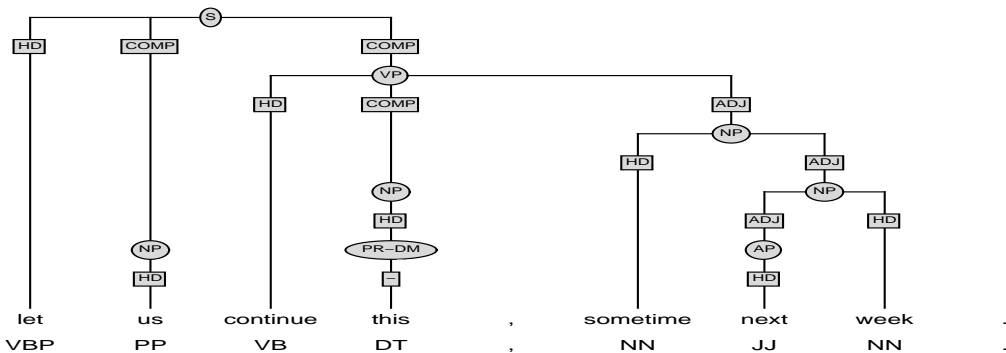
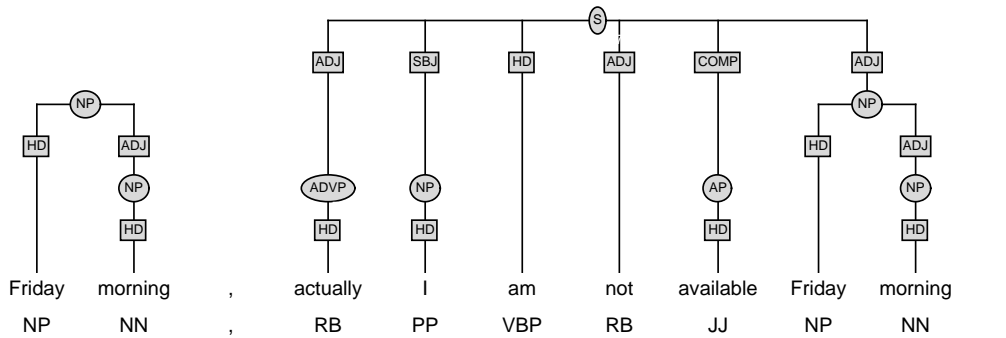
A *turn* can consist of several trees. In principle, the tree boundaries are defined by the so-called “longest match” strategy (cf., also Section (3.1) above), which requires that the tree structure includes as many constituents as possible, provided that the tree (sentence) remains syntactically, as well as semantically, **well-formed**.²

Consider, for instance, the following turn, which consists of three trees:

[The twenty sixth and twenty seventh are out] [twenty eighth I am available after one PM] [twenty ninth I could only meet you for an hour].

In the following examples, there are constituents that cannot be embedded, because they cannot be assigned reasonable grammatical functions and any potential attachment structure would cause problems to existing attachment strategies:

²Note that in Verbmobil punctuation is not always a reliable criterion to detect sentence boundaries.



Discourse markers, tagged and syntactically treated as UH (Interjections) in the English treebank, are isolated units occurring most of the times in the beginning, but sometimes also in the middle or on the left of a well-formed sentence, without being connected to the tree.

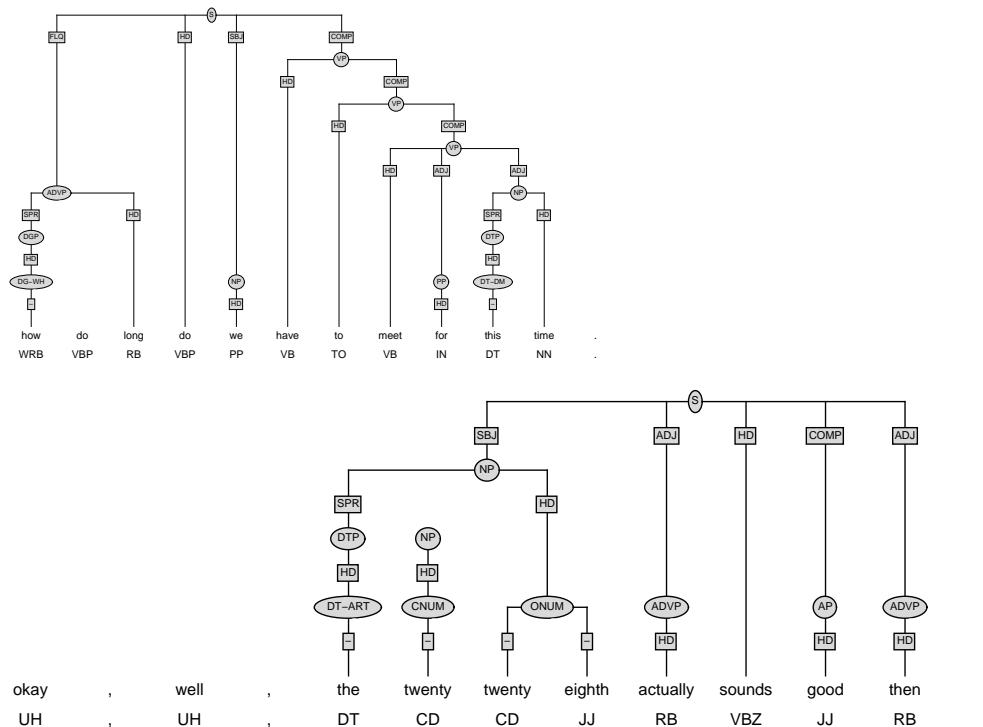
3.2.3 Speech Errors and Repetitions

Similar to interjections (i.e., discourse markers), speech errors, repetitions, corrections, and “hesitations” are isolated elements in the utterance, and consequently they are not bound to any node in the tree. Thus, they do not affect in any

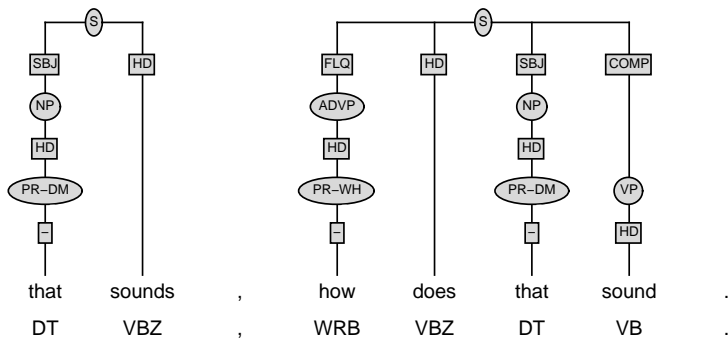
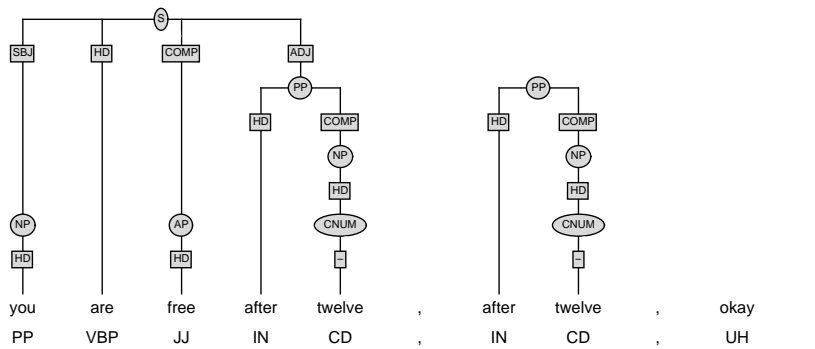
way the well-formedness of the sentence. This treatment is the consequence of a strategic decision concerning the annotation scheme adopted for the English treebank, which has been met mainly because in most cases the elements in question (i.e. discourse markers, along with speech errors, repetitions, corrections, and “hesitations”) either could not be assigned a specific grammatical function within the well-formed sentence, or the appropriate for them grammatical function has already been assigned to some other constituent:

1. *How [is] does that sound to you?*
 (*is* conflicts with *does* and cannot be included in the tree, since the S(entence) can only have one head (HD))

Speech errors, repetitions, corrections, and “hesitations” are structured as much as possible (mostly up to the level of phrasal categories), but are still not embedded in the well-formed tree for the reasons already mentioned above:

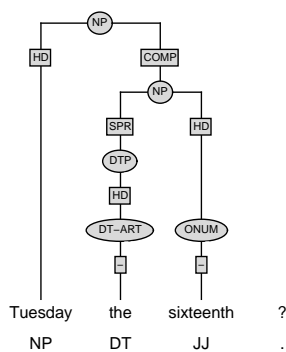


Verbmobil Report 241

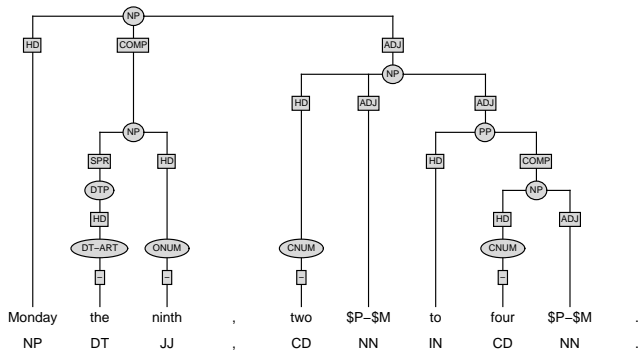
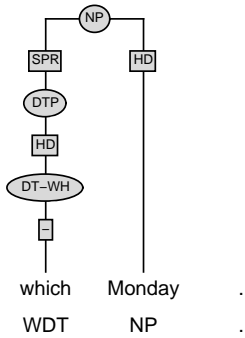
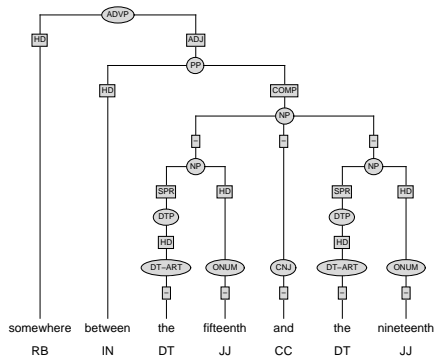
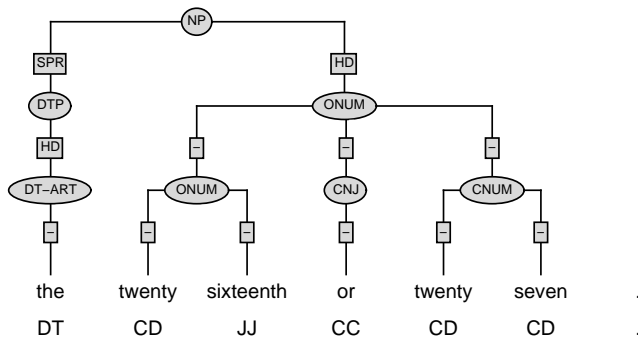


3.2.4 Isolated Phrases and Sentence Fragments

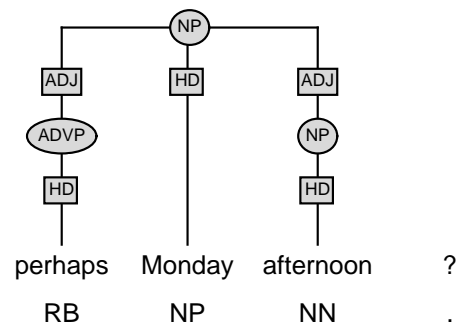
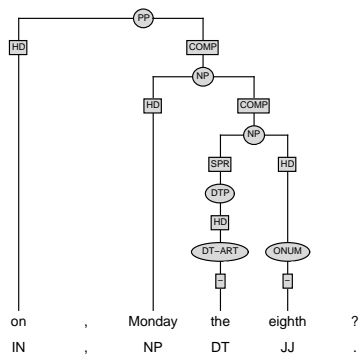
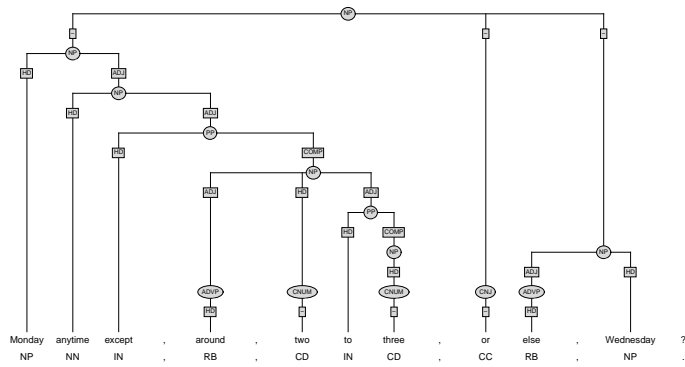
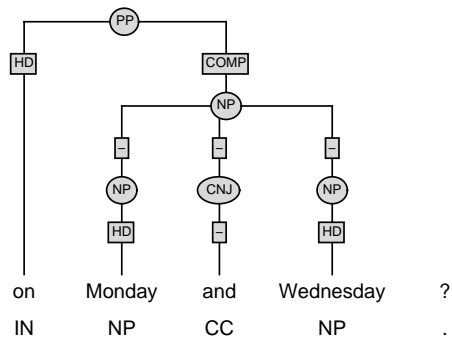
There are utterances which can be analysed neither as S(entences), since they lack a verbal constituent, nor as discourse markers (interjections). These constituents are projected only up to the level of phrasal category, as the following example shows:



Stylebook for the English Treebank

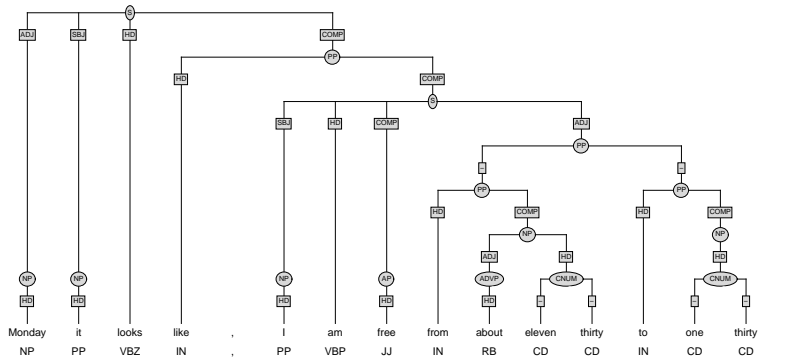


Verbmobil Report 241

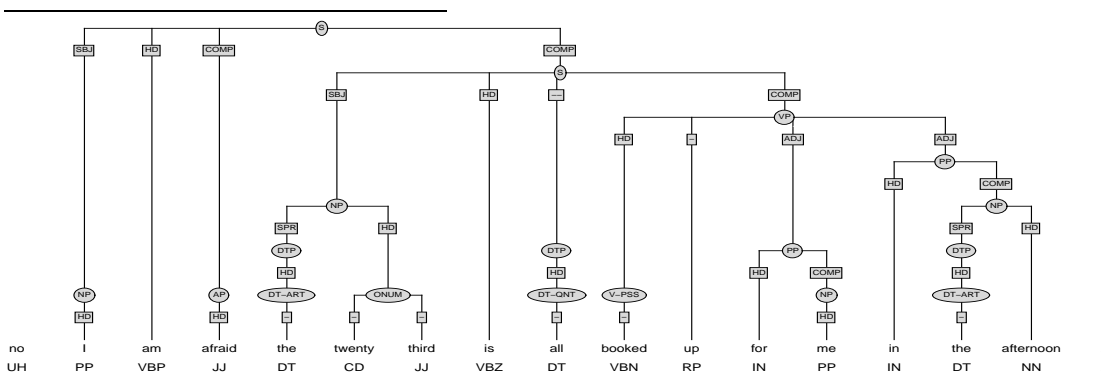
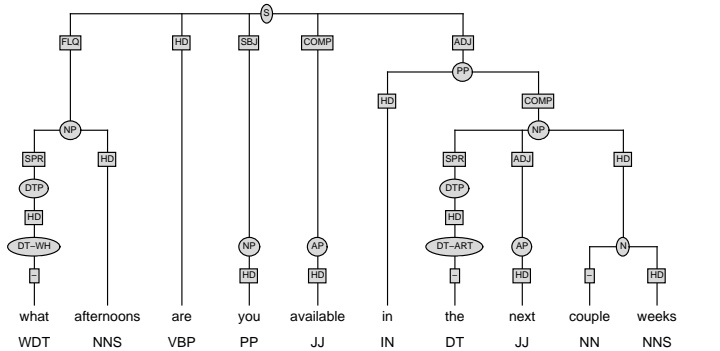


discontinuous constituents.⁴

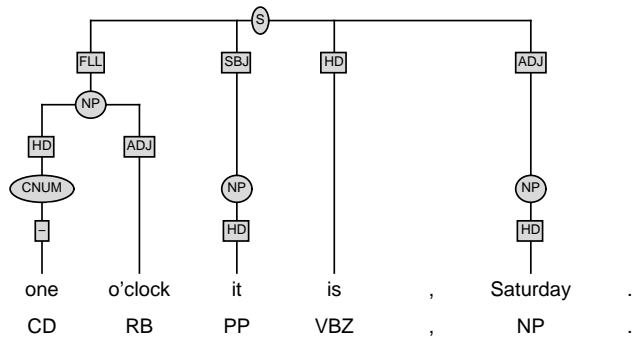
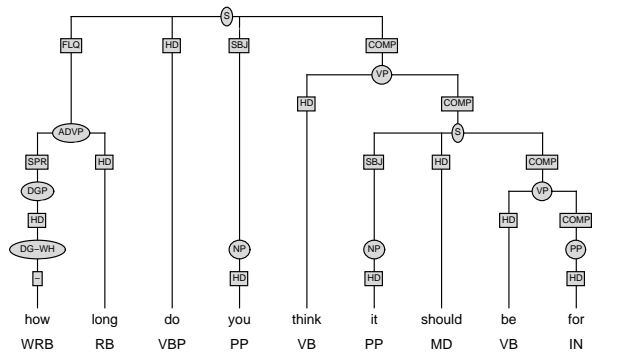
As far as discontinuous constituents in non-interrogative, non-relative clauses are concerned, their parts are separately bound flatly to the higher possible node (i.e., either to a V(erb)P(hrase), or to a S(entence)), as in the following:



To account for discontinuous constituents occurring in interrogative sentences, we adopt a filler-gap dependency syntactic analysis:



⁴The same strategy is also true as far as the German treebank is concerned (cf., (Stegmann et al. 2000)).



3.2.6 Empty Edge Labels

The only empty edge labels in the English treebank are the edge labels found below determiners, degree words, C(o)MP(lementizers), some pronouns, and C(O)NJ(unctions). Note that these have intentionally been left empty because of lack of any semantic content at the level of syntactic analysis, where the empty edge labels occur.

Chapter 4

The Annotation of Phrases

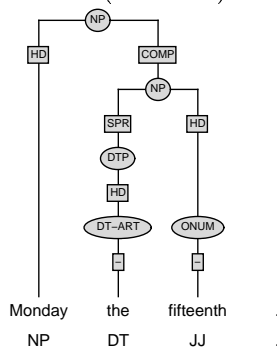
4.1 Internal Structure of Phrases

4.1.1 Noun Phrases (NPs)

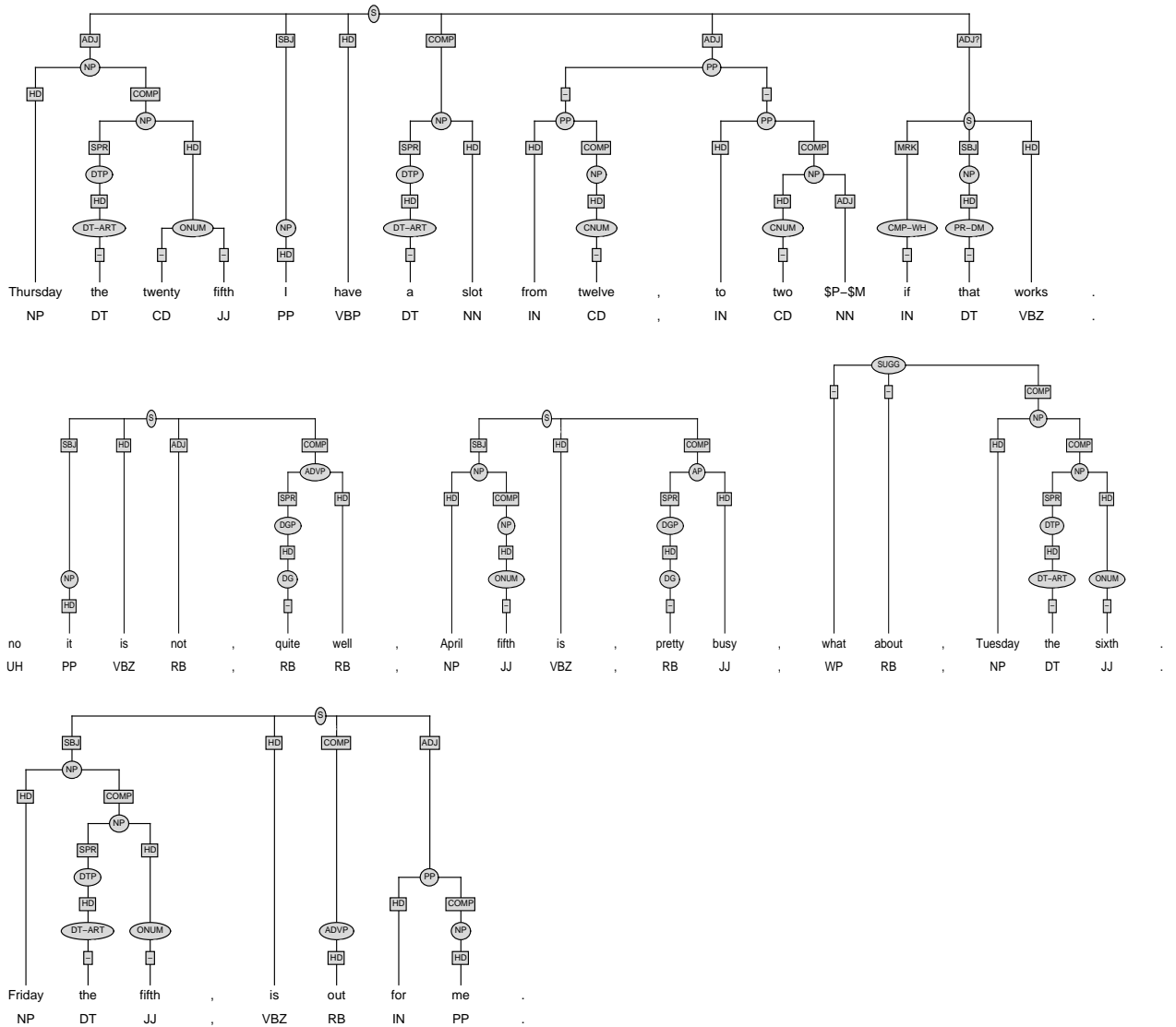
A simple noun phrase (NP) without modifiers consists of a head noun (NN, NP, or a pronoun) and (optionally) a determiner. The edge label of the head noun is HD (=head), the edge label of the determiner is SPR (Specifier).¹

4.1.1.1 Appositions

Appositions such as *Monday, the fifteenth*, or *on Monday, the fifteenth* are immediately attached to the head noun on a “low level”, and they are treated as COMP(lements) of the head noun:

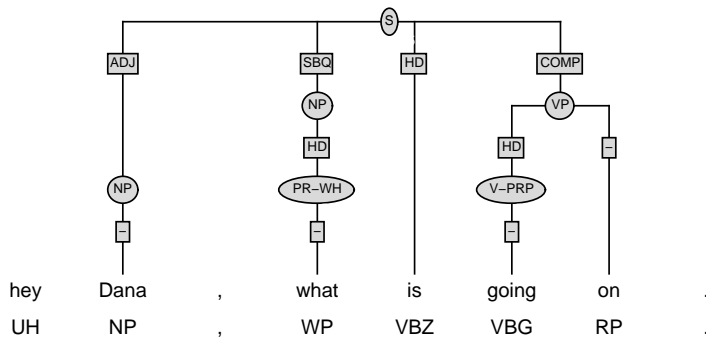
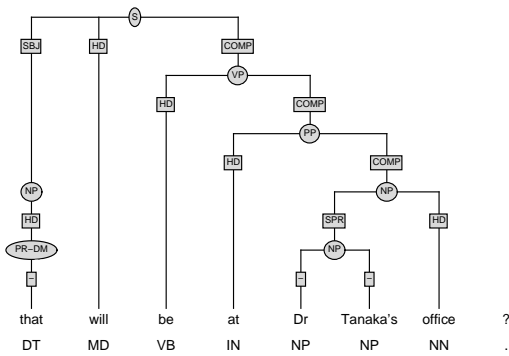
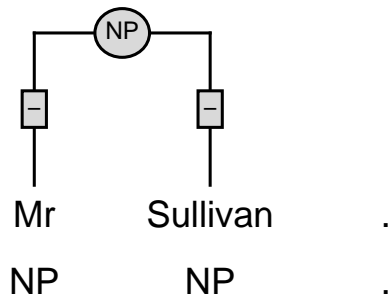


¹Note that the English treebank, unlike the German one, having adopted an HPSG-oriented annotation scheme does not support any D(eterminer)P(hrases).



4.1.1.2 Noun Phrases including Proper Nouns

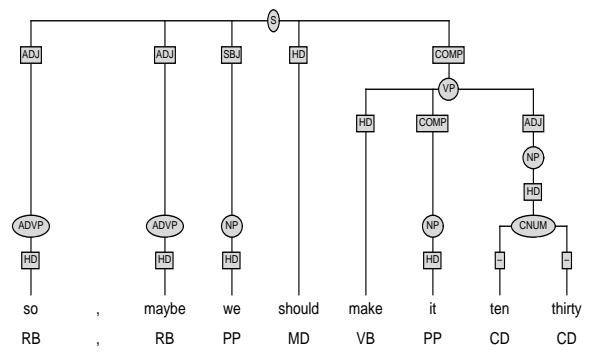
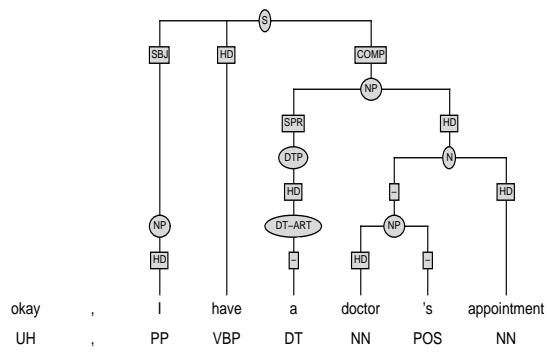
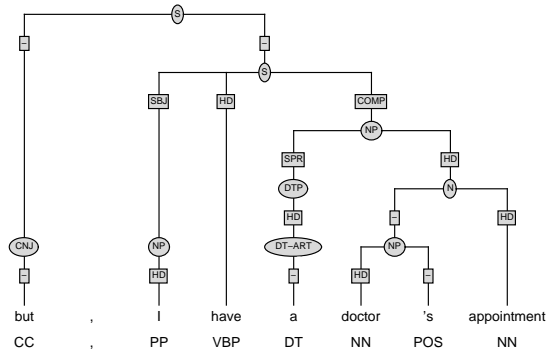
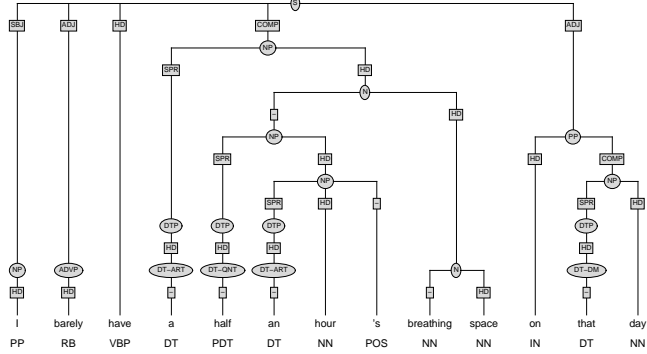
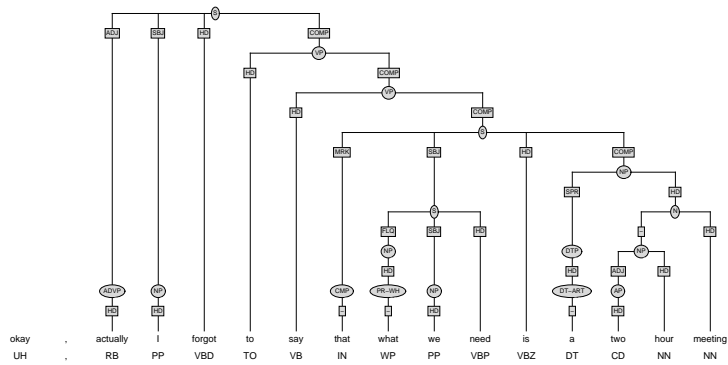
Noun phrases including proper nouns, names with titles, etc., are treated as normal NPs, though without heads:



4.1.1.3 Compound Nouns

The English treebank follows the common in all linguistic theories assumption that compounds are noun phrases (NPs) whose main semantic information comes from the last in the row noun, which is also the head of the whole noun phrase:

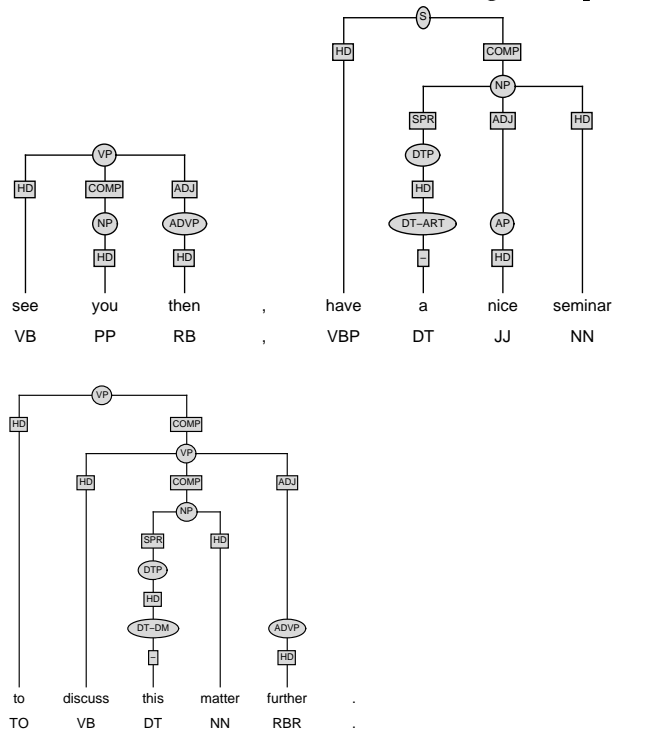
Verbmobil Report 241



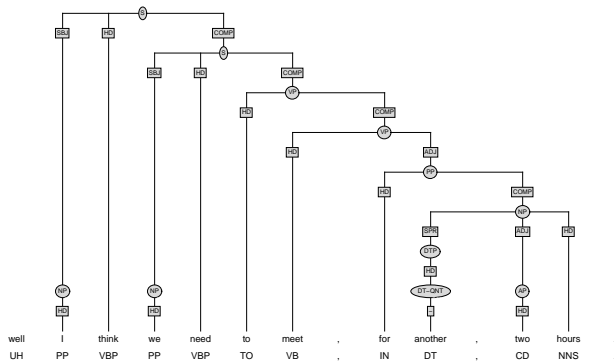
4.1.2 Determiner Phrases

There are three kinds of determiners found in the English treebank: D(e)T(erminer) ART(icle)s (*a, the*), D(e)T(erminers)D(e)M(onstrative)s (*this, that, these, those*), and D(e)T(ermines)-Q(ua)NT(ifier)s (*any, some, no, both, either, neither*). Because the Penn treebank Tagset that is being used for the pos-tagging in the English treebank does not differentiate between these three different sorts of determiners,² a decision has been met as far as the syntactic annotation scheme developed for the English treebank is concerned, which dictates that the three different determiner subsorts should be labelled accordingly in a node label immediately above the terminal categories, projecting thereafter to a normal D(e)T(erminer)P(hrase), which is ultimately treated as the SP(ecific)R of the corresponding NP.

Consider, for instance, the following examples:



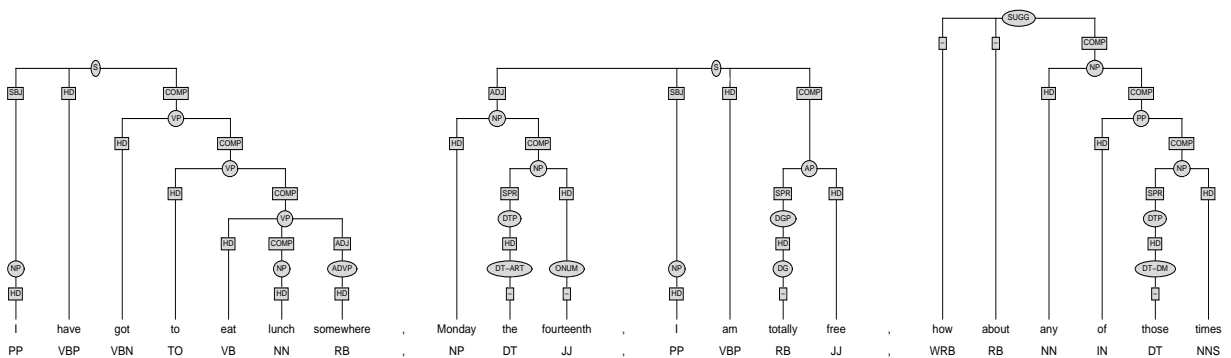
²All of them bear the Part-of-Speech tag: D(e)T(erminer).



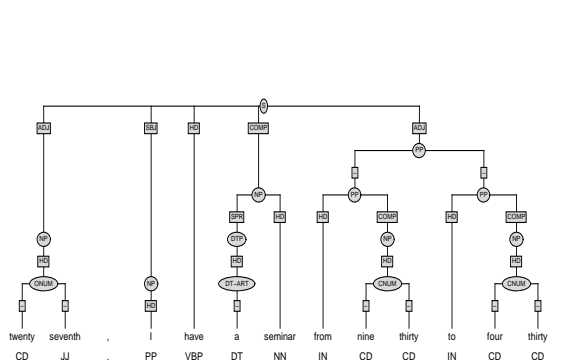
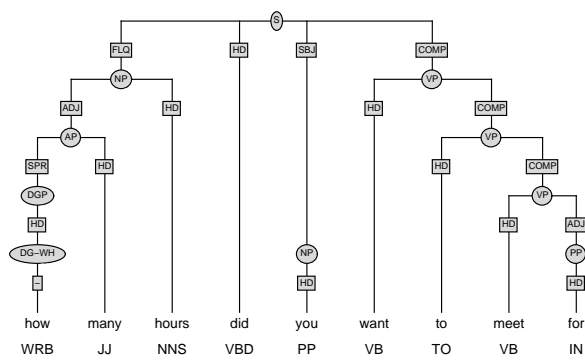
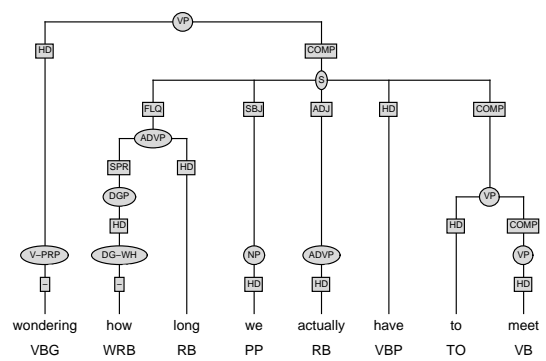
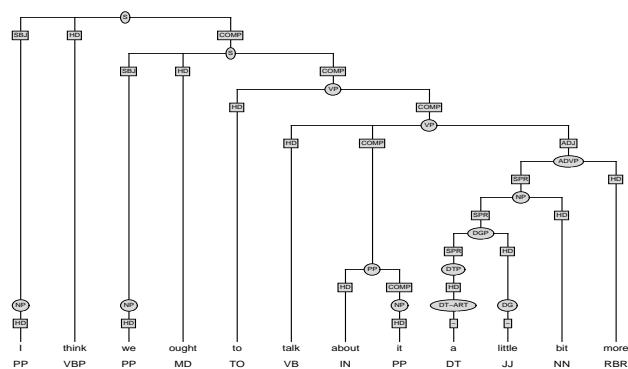
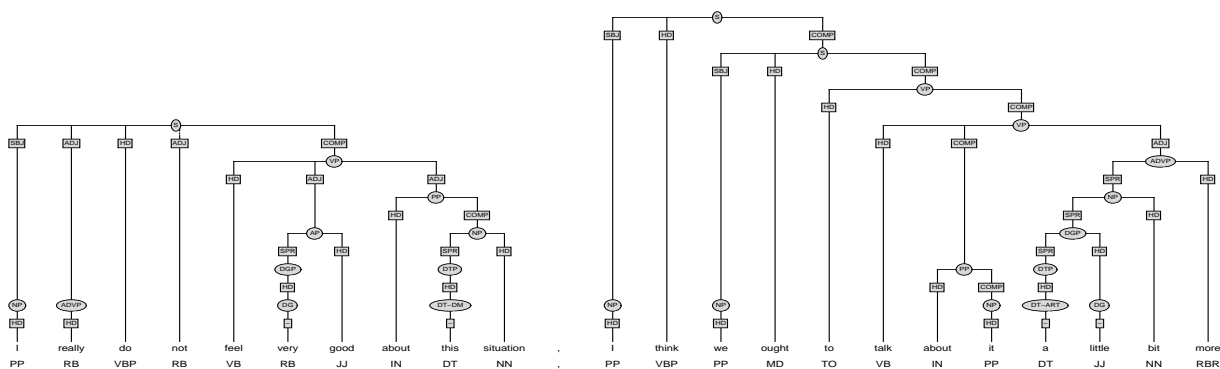
4.1.3 Degree Phrases

Adverbs like *totally*, *very*, *really*, *completely*, *awfully*, *extremely*, *terribly*, *rather*, *how* etc., noun phrases like *a little bit*, *a lot* etc., adjectival phrases like *how many*, quantifiers like *no*, *any*, *some* etc., when they are followed by adjectives or adverbs, determiners like *that*, *this* etc., in their “adverbial” use, expression like *as soon as*, *as long as*, *as well* etc. serve as degree words in the English treebank.

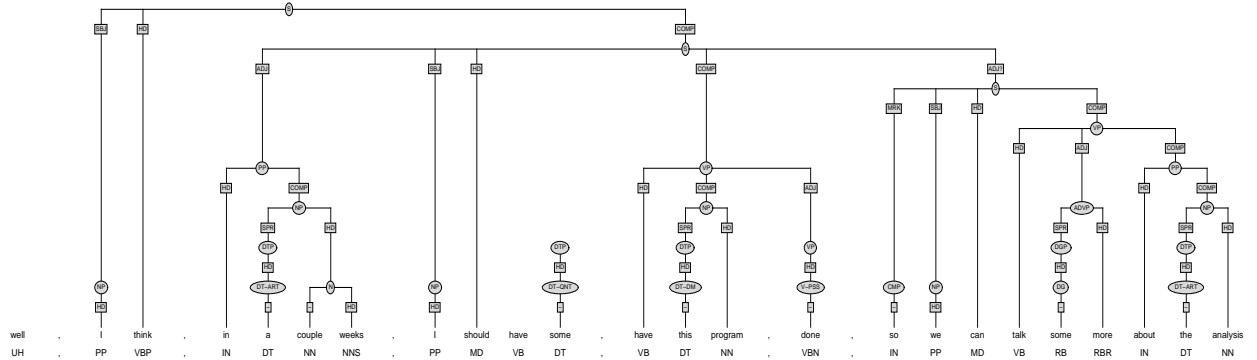
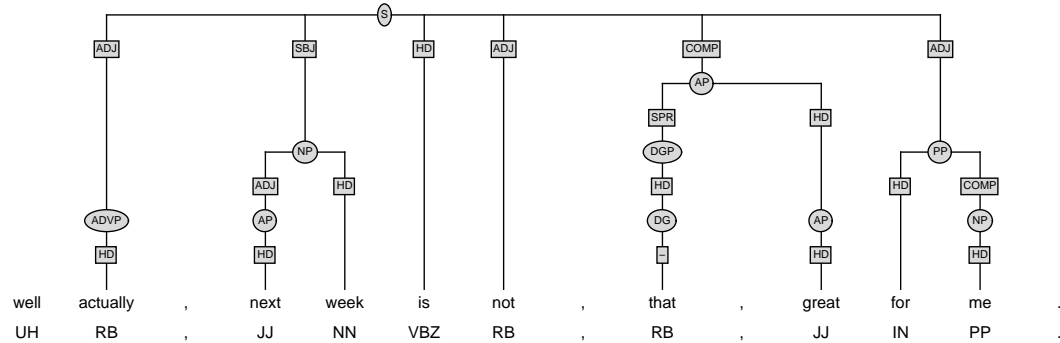
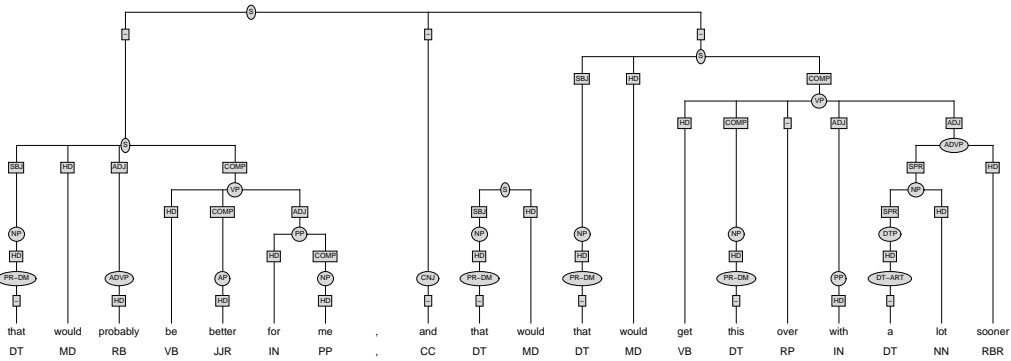
The strategy for the syntactic treatment of degree words, and ultimately D(e)G(ree)P(hrase)s in the English treebank parallels that of D(e)T(erminer)P(hrase)s as we have described it immediately above. That is, we treat degree words in a uniform way at the part-of-speech level, labelling them mainly as adverbs (RB), we differentiate among them at a node label immediately above the terminal categories (where one can find “simple” degree words, labelled as DG, interrogative degree words, labelled as DG-WH, etc.), and thereafter we project them to a D(e)G(ree)P(hrase), which like a D(e)T(erminer)P(hrase) serves as the SP(ecifier) of the corresponding phrase, since determiner phrases and degree phrases are both subtypes of specifiers.



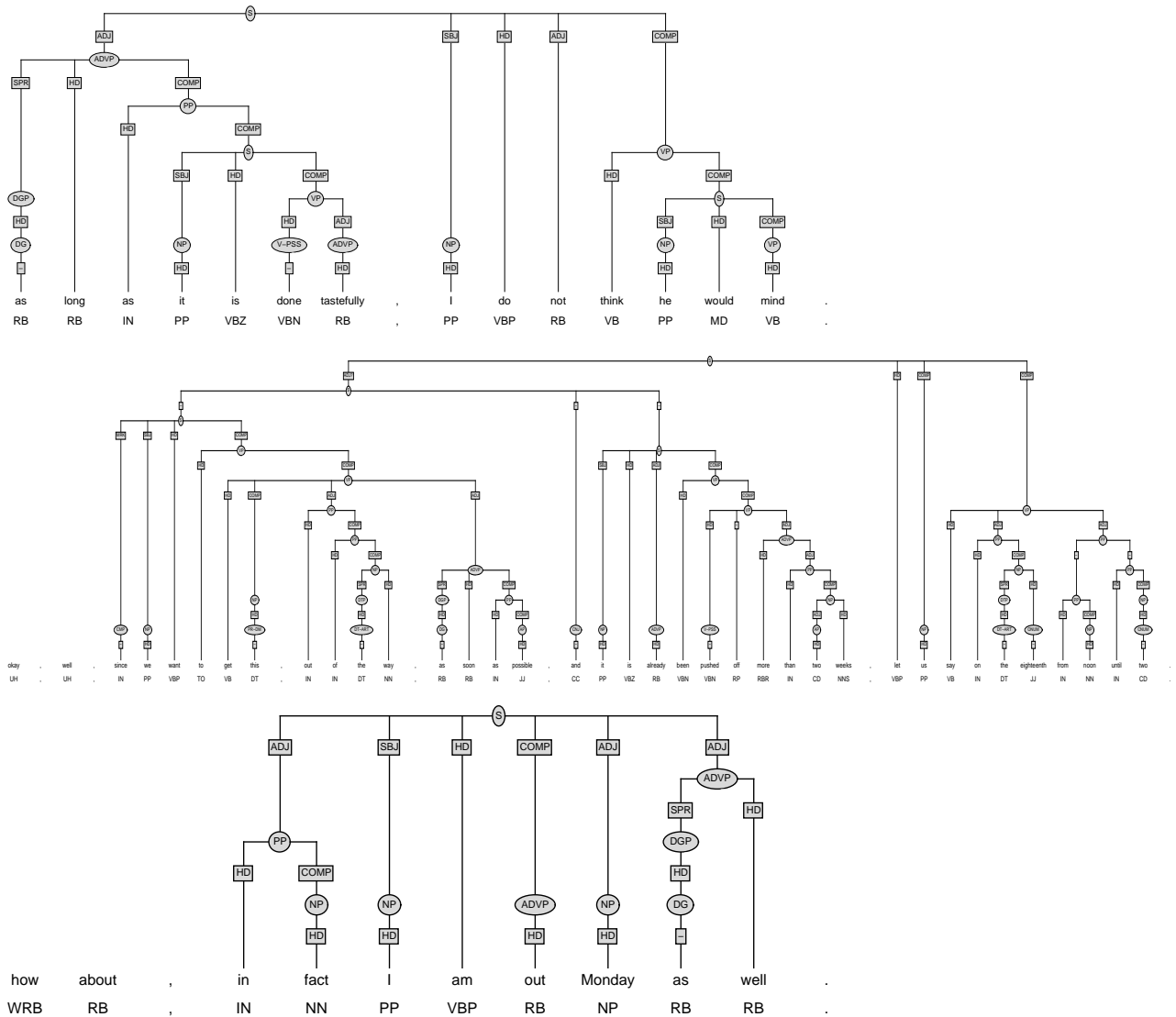
Stylebook for the English Treebank



Verbmobil Report 241



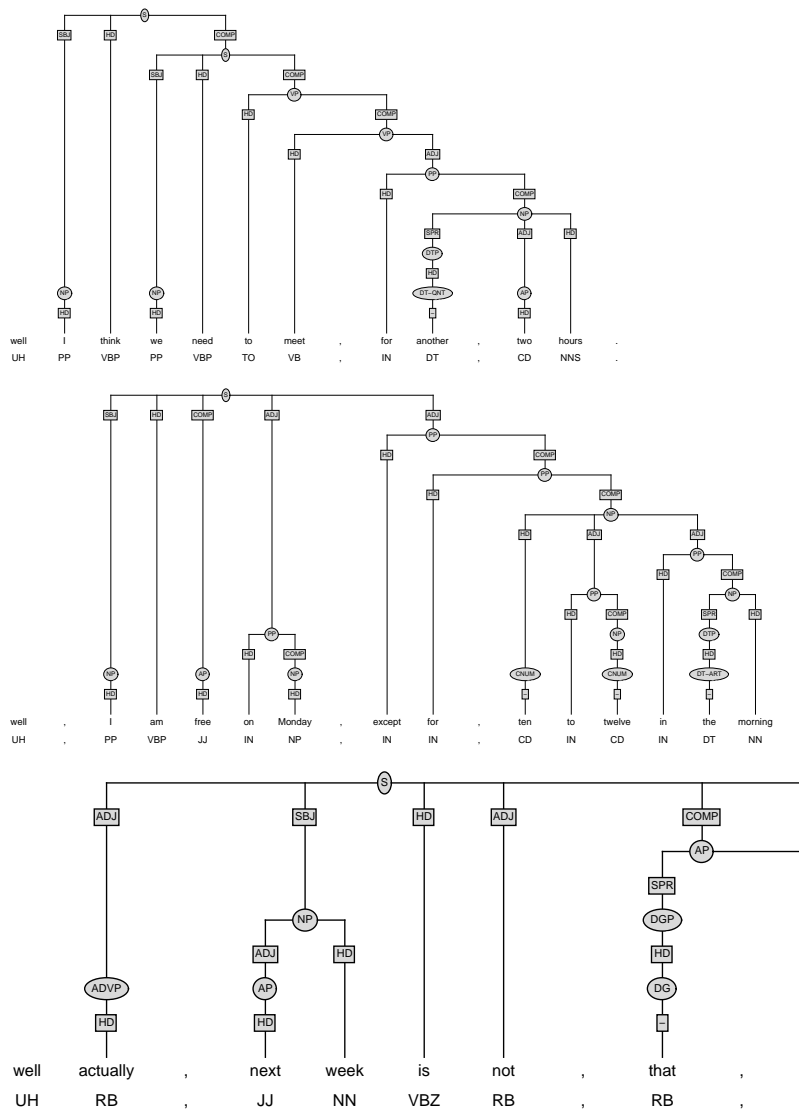
Stylebook for the English Treebank



4.1.4 Prepositional Phrases

Unlike the German treebank in Verbmobil, where for purely pragmatic project requirements the complement in prepositional phrases is annotated as the head of the phrase, the English treebank follows the common in all linguistic theories assumption that the head of a prepositional phrase is the preposition itself, which is also “responsible” for selecting its COMP(lement) NP, as in the following:

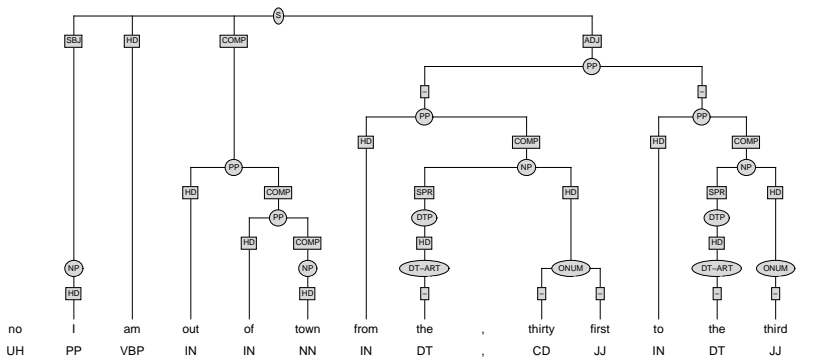
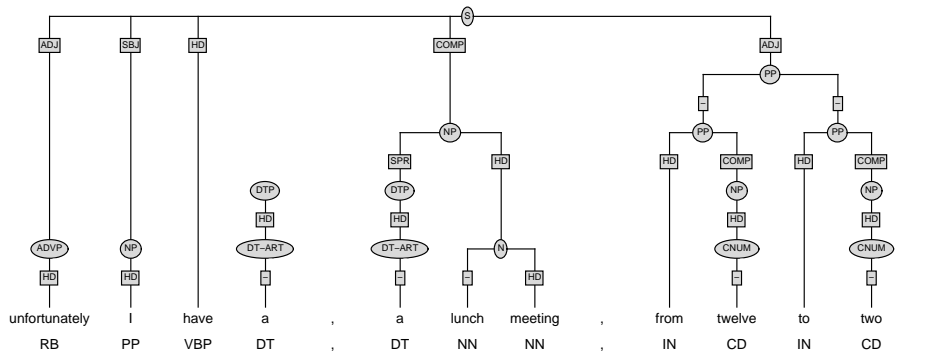
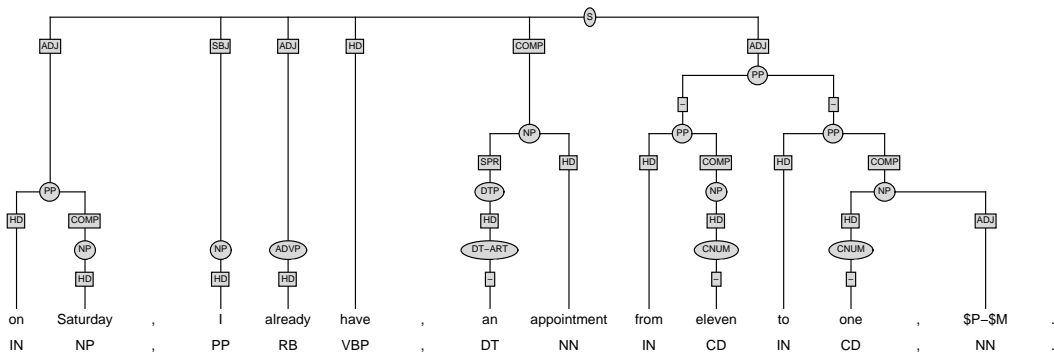
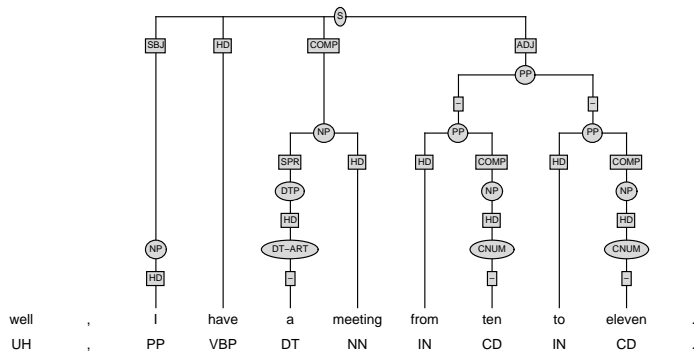
Verbmobil Report 241



There are, though, two specific instances of complex (i.e., consisting of more than two PPs) prepositional phrases in Verbmobil, where the annotation scheme intentionally combines a data- as well as theory-driven approach in order to account for them.

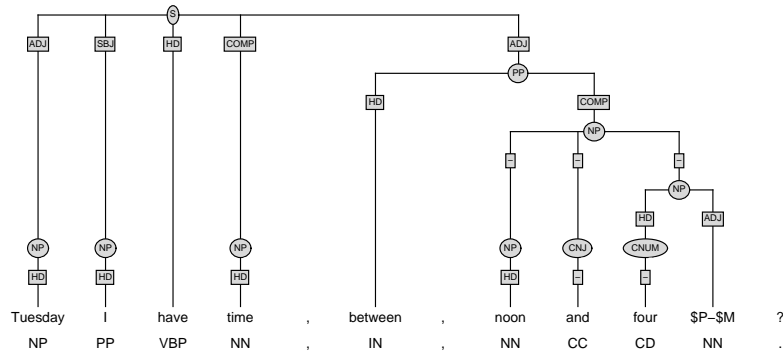
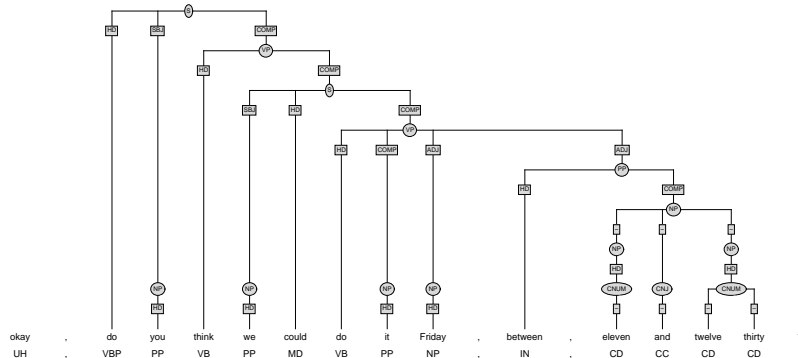
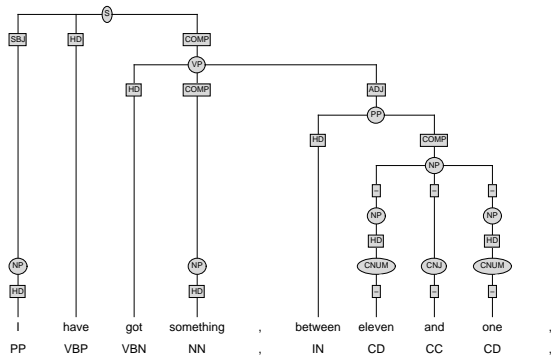
The first case involves complex temporal prepositional phrases expressing intervals of the form *from/to*, like it is shown in the following examples:

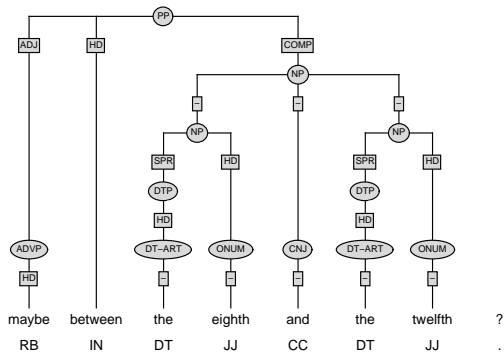
Stylebook for the English Treebank



Verbmobil Report 241

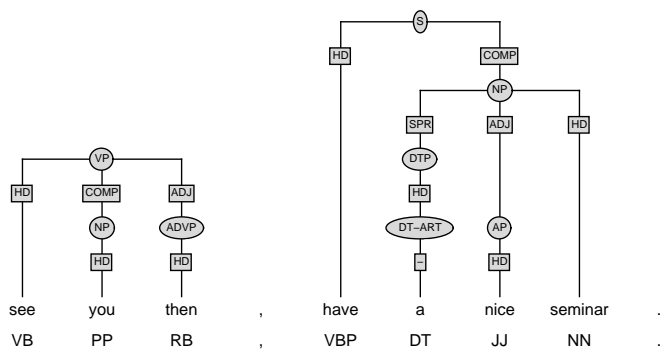
The second one involves *between*. It seems as far as the data in the English treebank has shown us so far that the preposition in question requires a complement containing a coordination structure:



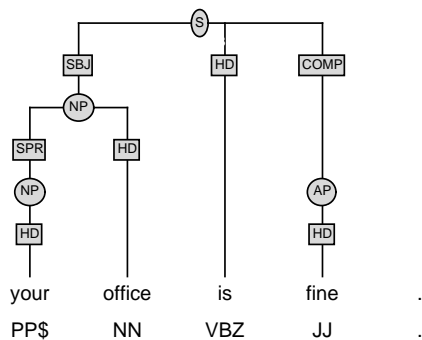


4.1.5 Adjectival Phrases

In the English treebank, adjectival phrases are treated either as ADJ(uncts), in which case they are attached, either prenominally or postnominally, to the head of the phrase they modify on the same level yielding this way a flat structure:

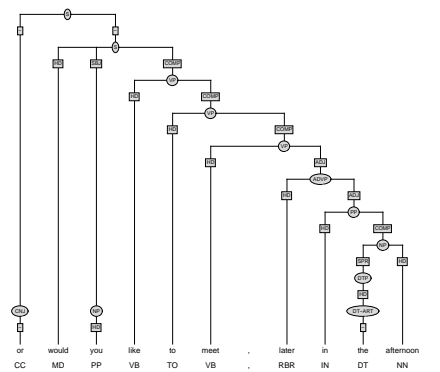
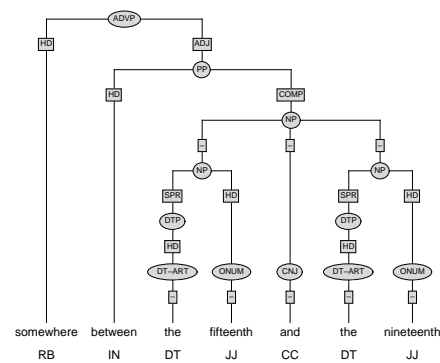
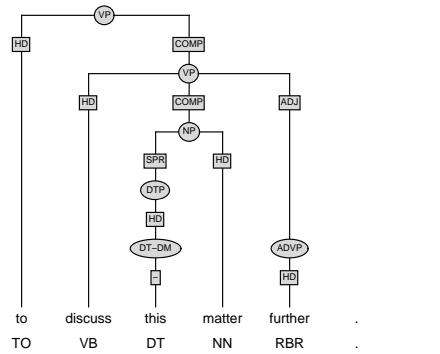


or as COMP(lements) of the verb *be* in the copula construction:



4.1.6 Adverbial Phrases

A few examples of ADV(erbial)P(hrases):

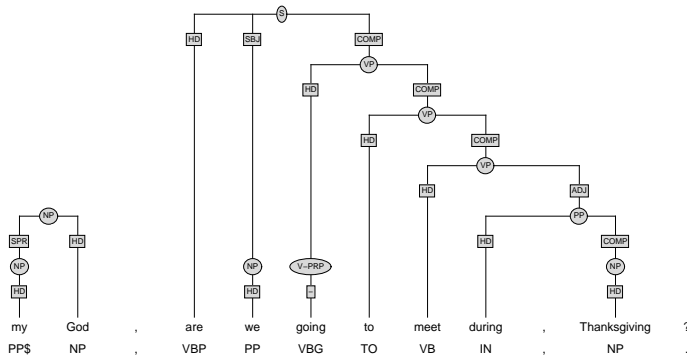
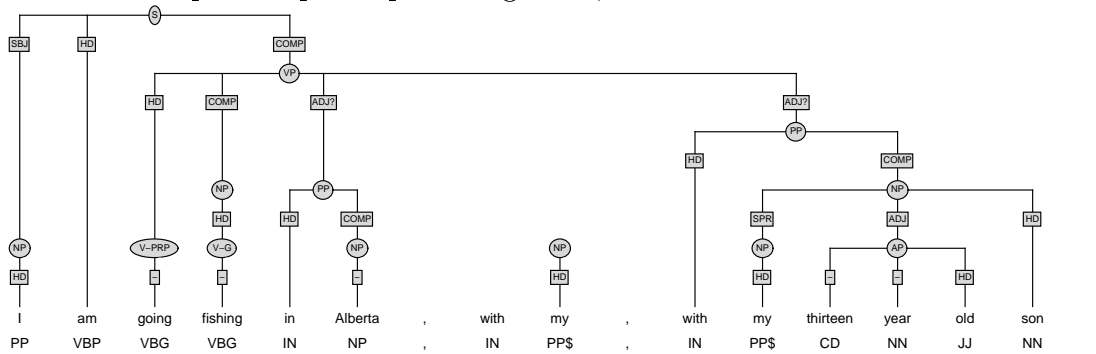


4.1.7 Verb Phrases

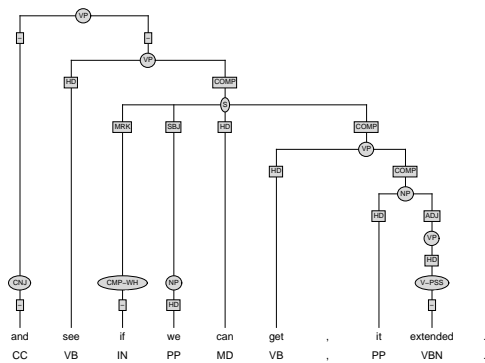
Among the verb phrases:

1. VBP labels a finite, non-3rd-person-singular verb in present tense,
2. VBZ a finite, 3rd-person-singular verb in present tense,

3. VBD a finite, past tense verb,
4. MD a modal verb
5. VBG a present participle or a gerund,



6. VBN a past participle,



7. VB the infinitival form of a verb
8. TO a verb phrase (mostly an infinitival one) introduced by TO

4.1.7.1 Particle Verbs

Following the *Part-of-Speech Tagging Guidelines for the Penn treebank Project* (Santorini 1990:pg.10-11), we differentiate between prepositions (IN) and particles (RP) according to the following criteria:

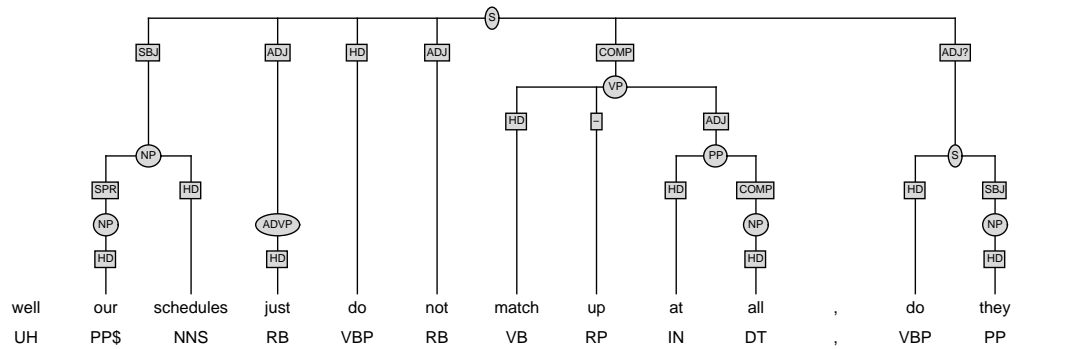
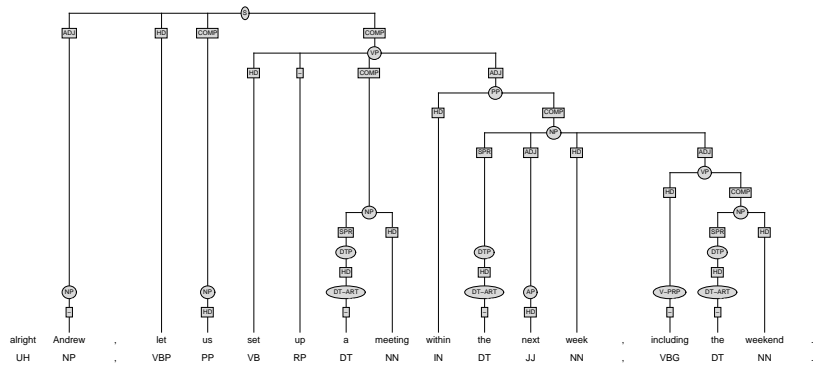
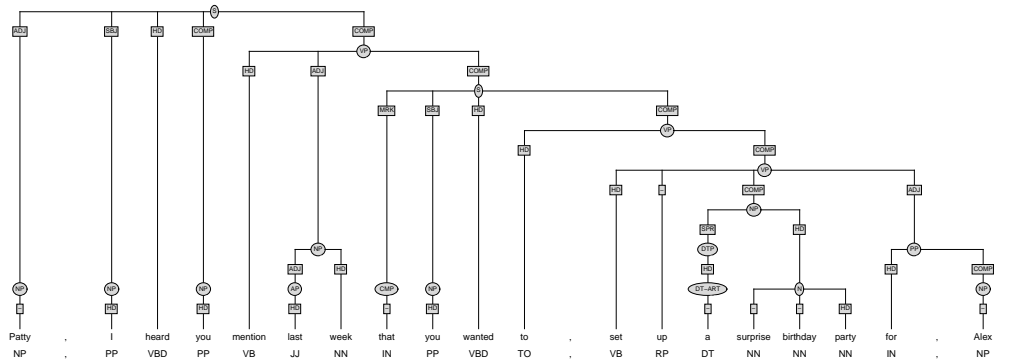
A word is a particle (RP) rather than a preposition (IN):

1. if it can either precede or follow a noun phrase object: She told off/RP her friends. / She told her friends off/RP;
2. if when one replaces a noun phrase object by a pronoun, the pronoun must precede the word: She told them off/RP. / * She told off/RP them.
If the results of this test conflict with the results of the first test, go by the results of the second: ??? to run a bill up/RP / to run up/RP a bill; to run it up/RP / * to run up/RP it;
3. if it can be part of a noun that is derived from a particle-verb collocation: to break down/RP; breakdown / to break through/RP; breakthrough / to be left over/RP; leftovers / to push over/RP; pushover / to put down/RP; putdown. The results of this test are one-directional only; if there is no related noun, the word can still be a particle: to pass out/RP; * passout / to pull off/RP; * pull off;
4. if it bears stress in clause final position (this criterion only applies to monosyllabic words): Why don't you come by/RP? vs. Real bargains are hard to come by/IN.
5. While particles usually occur in verb constructions, they can also co-occur with parts of speech derived from verbs: the cutting/NN off/RP of the top / the setting/NN up/RP of the problem.

On the other hand, a word is a preposition (IN) rather than a particle (RP):

1. if it must precede a noun phrase object: I live off/IN campus. / * I live campus off/IN.
2. if when one replaces a noun phrase object by a pronoun, the pronoun cannot precede the word: She has been into/IN it for a year. / * She has been it into/IN for a year.
3. if it cannot bear stress in clause-final position (this criterion only applies to monosyllabic words): Real bargains are hard to come by/IN vs. Why don't you come by/RP?

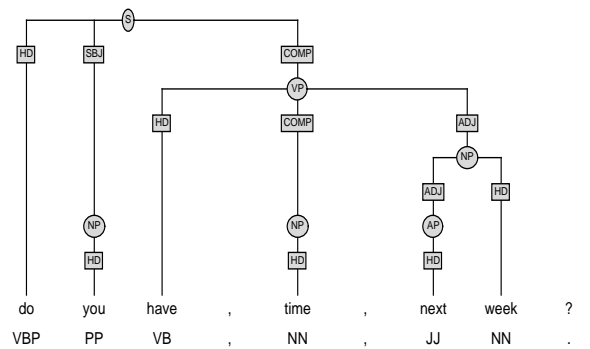
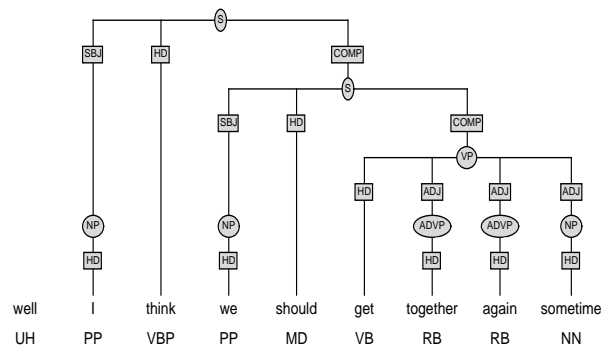
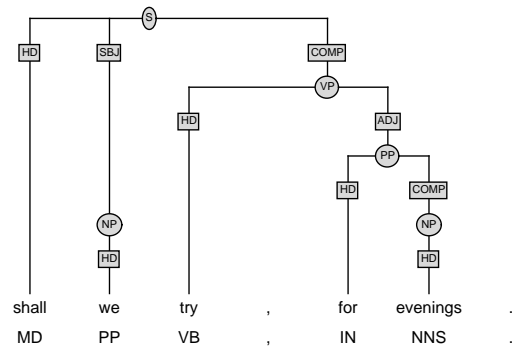
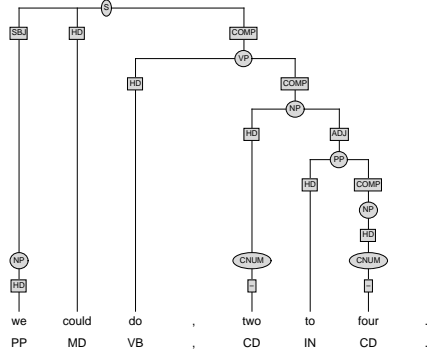
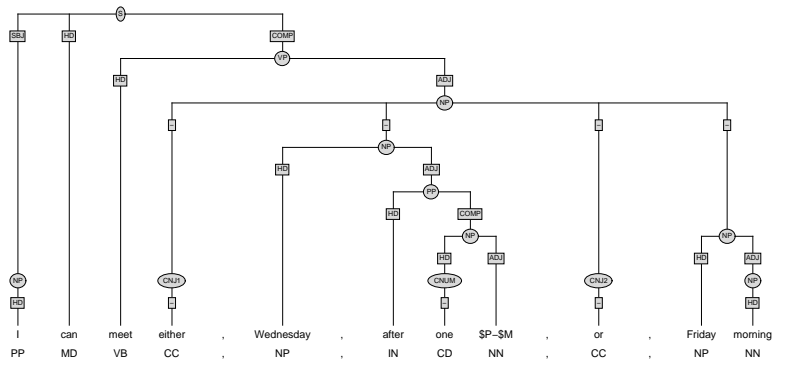
Stylebook for the English Treebank



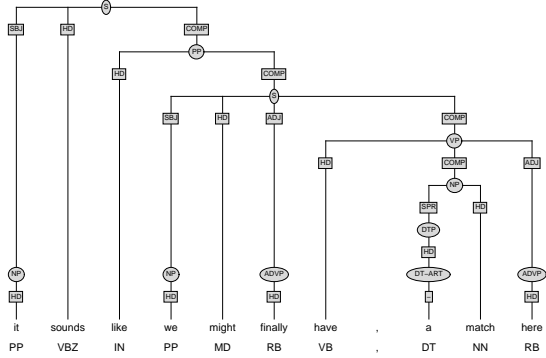
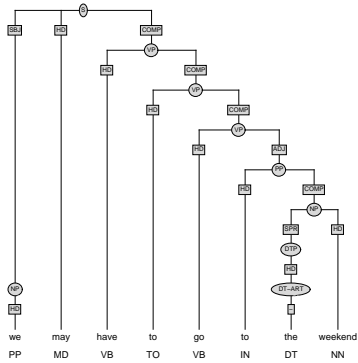
4.1.7.2 Modal Verbs

A few examples with modal verbs:

Verbmobil Report 241



Stylebook for the English Treebank

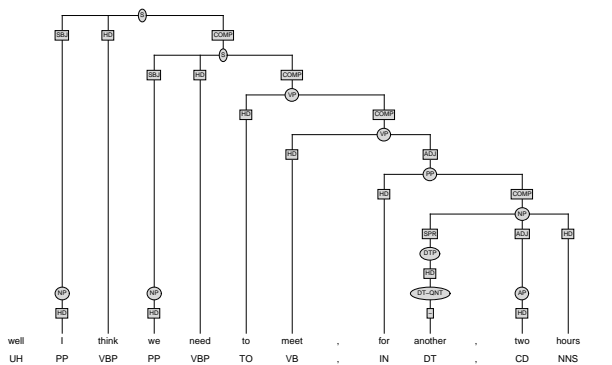


Chapter 5

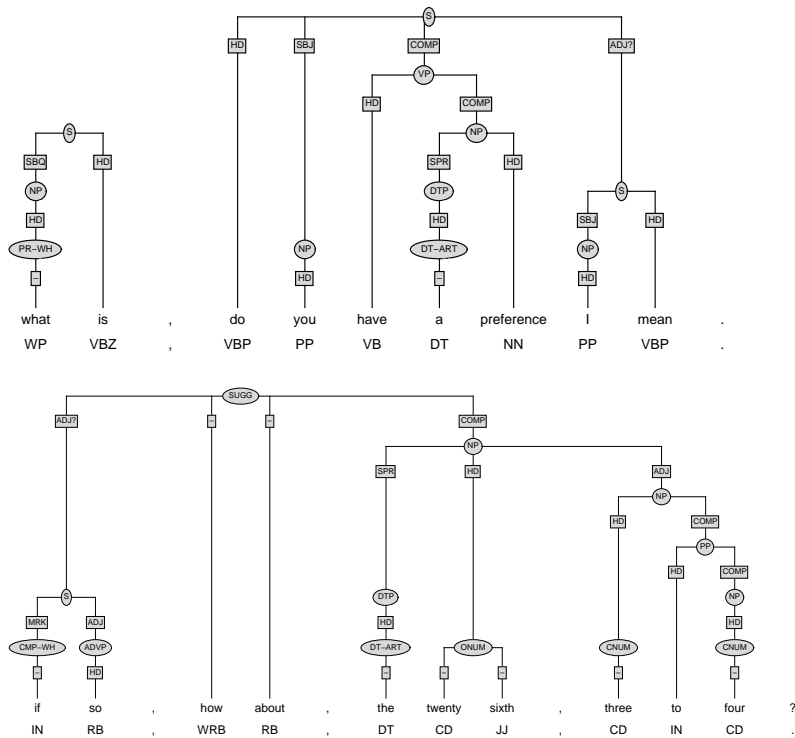
At the Sentential Level

5.1 Combining Phrases into Sentences

As has been already mentioned above, a grammatical S(entence) in the English treebank should consist mainly of a S(u)BJ(ect),¹ and a finite or a modal verb. COMP(lement)s' attachment is dictated by the verb's subcategorization frame, which is lexical-semantics driven (cf., document e-Verblast). Adjuncts' attachment is also semantically-driven, and in cases of ambiguities, adjuncts are attached to the highest possible level, given at the same time the grammatical function of ADJ?, which is a shorthand for adjunct ambiguity statement:



¹S(u)BJ(ect) for affirmative clauses, S(u)B(ject)Q for interrogative clauses, and S(u)B(ject)R for relative clauses.



5.1.1 Grammatical Functions

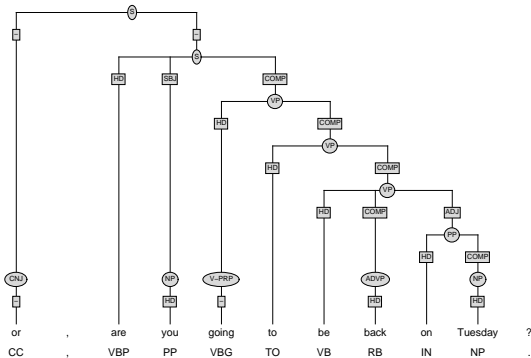
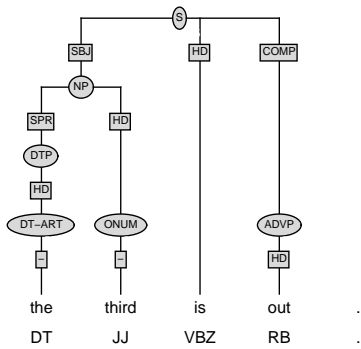
Edge labels express grammatical functions of constituents in the treebanks.

Within phrases, the main grammatical function, as far as the English treebank is concerned, is the head (HD).

The rest of the grammatical functions used in the annotation scheme of the English treebank are the following:

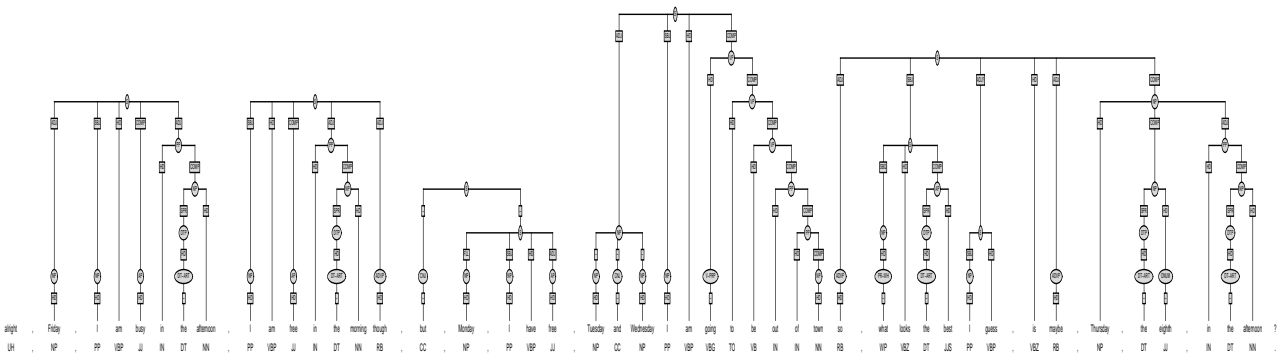
1. COMP for complements; all phrasal categories can serve as complements in the English treebank, as long as the sentence matrix verb's subcategorization requirements are satisfied:

Verbmobil Report 241



2. SPR for specifiers

3. SBJ for the subjects of affirmative clauses:

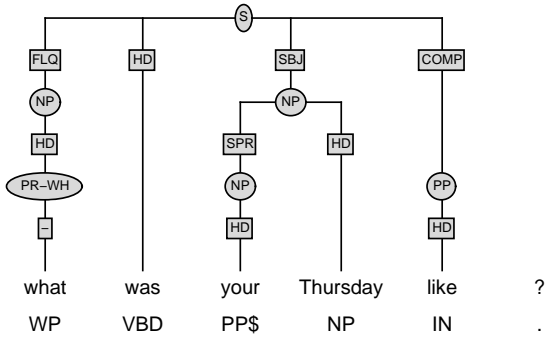
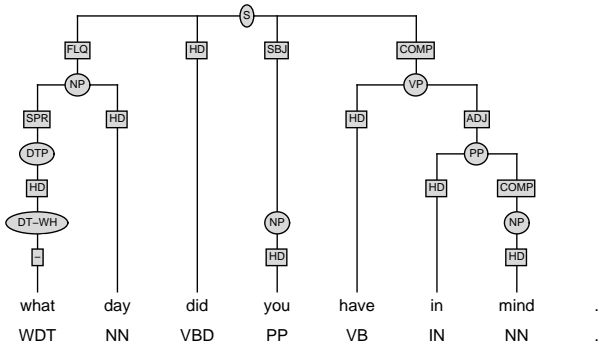
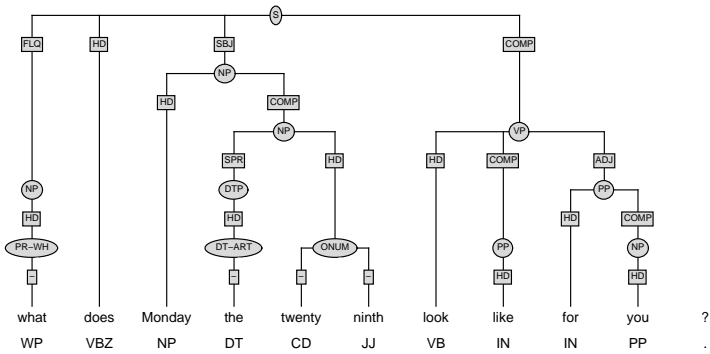
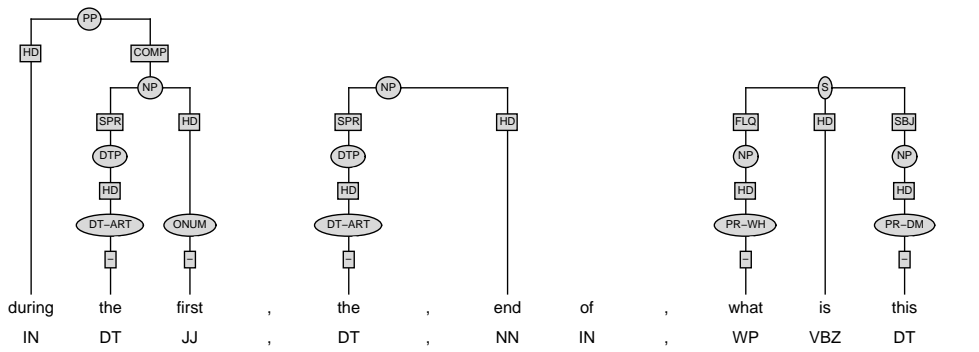


4. SBQ for the subjects of interrogative clauses

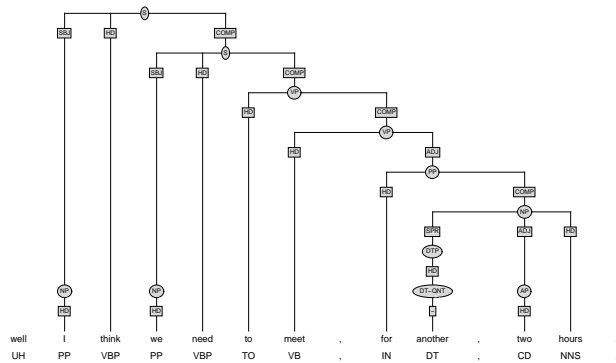
5. SBR for the subjects of relative clauses

6. ADJ for adjuncts; again all phrasal categories can serve as adjuncts in the English treebank, as long as their grammatical function is semantically motivated

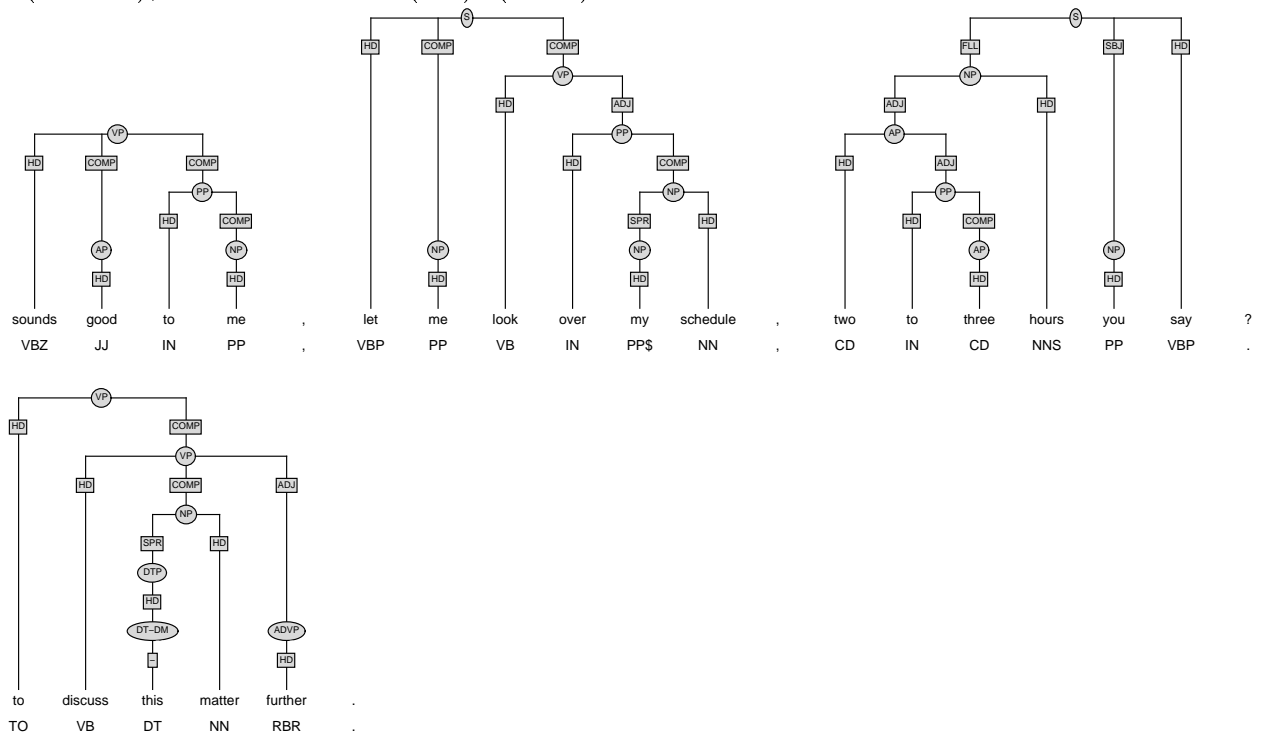
Verbmobil Report 241



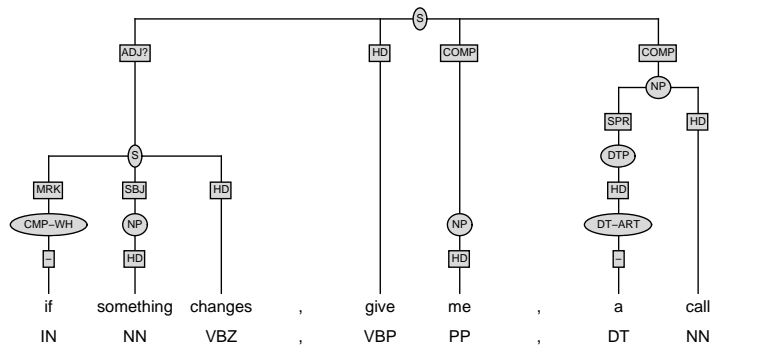
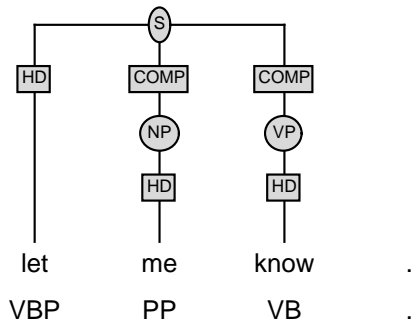
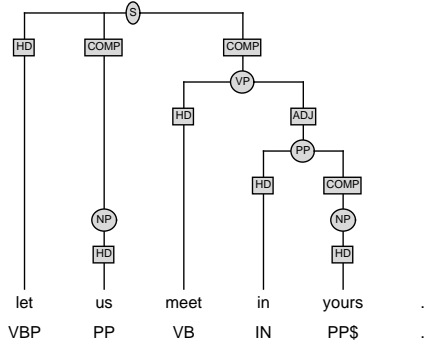
Verbmobil Report 241



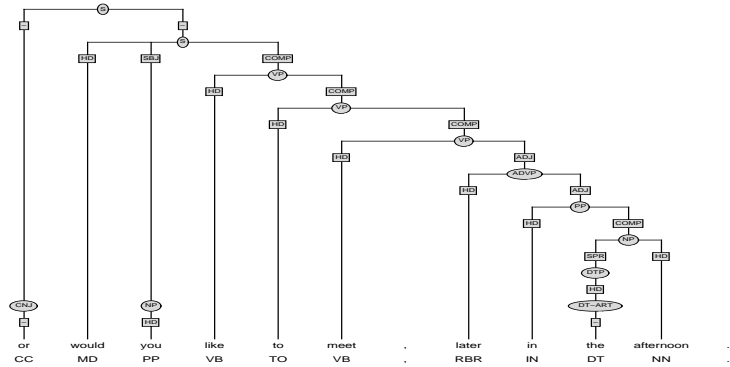
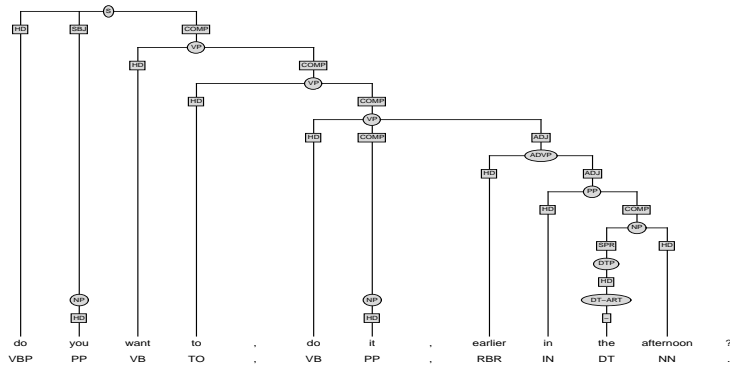
What is characteristic as far as the English treebank is concerned is that for a whole utterance to be eligible to be annotated as a S(entence), this should contain not only a finite verb, but also a noun phrase eligible to serve as the S(u)BJ(ect) of this verb. In the case that the S(u)BJ(ect) is missing, due, for instance, to fast speech phenomena, the utterance is no longer annotated syntactically as a S(entence), but rather as a V(erb)P(hrase):



The only exception here are sentences containing **imperatives**, as in the following:

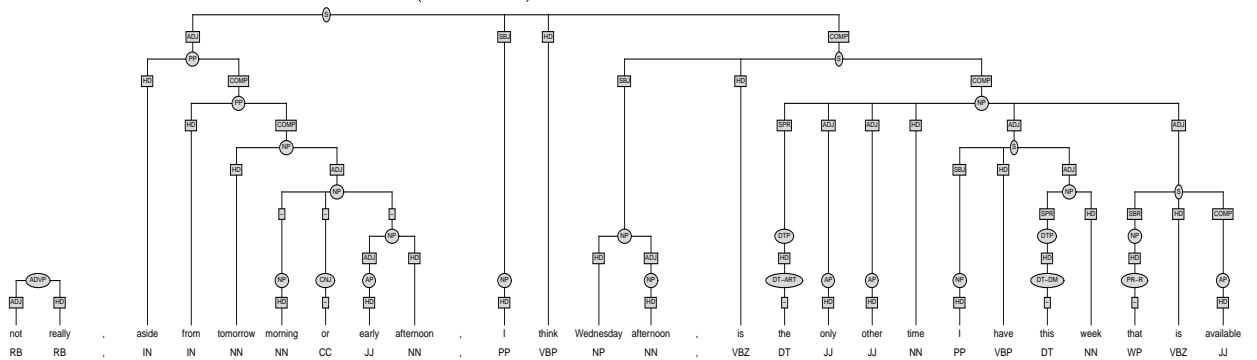


One further element characterizing the syntactic annotation scheme developed for the English treebank is that at the uppermost level of a grammatical sentence, the subject combines immediately with the sentence's verbal head, and not with a verb phrase (VP, as in traditional HPSG), in order to form a fully saturated S(entence):

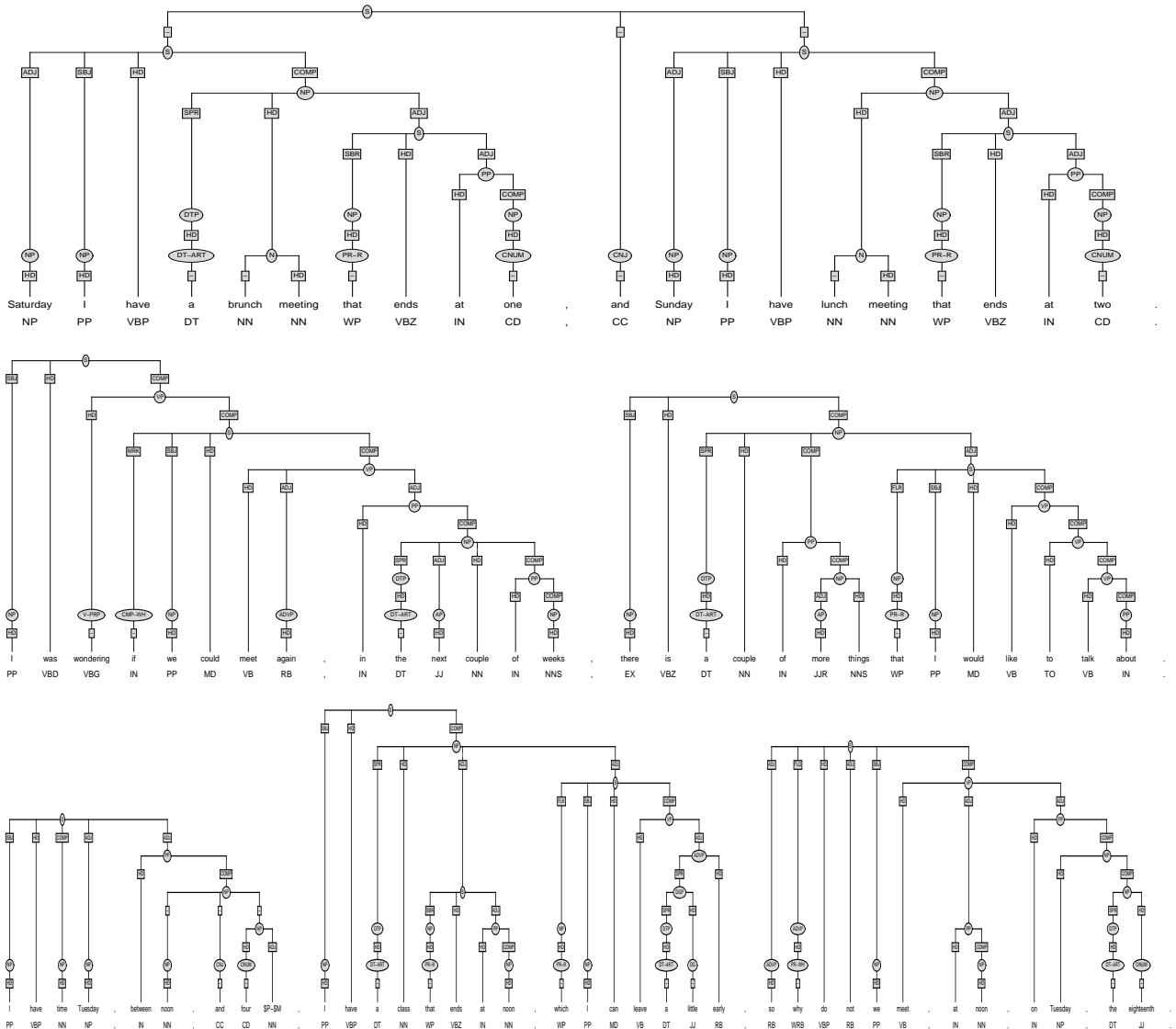


5.2 Relative Clauses

Here are a few examples of relative clauses, which most of the times in the English modify either NPs or whole S(entence)s:



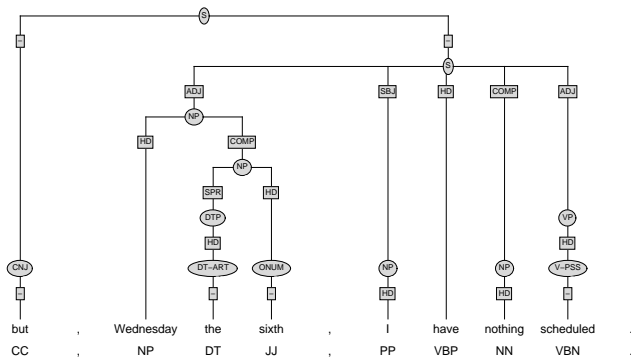
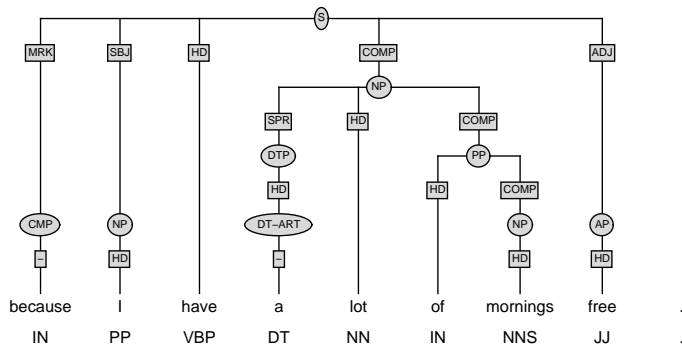
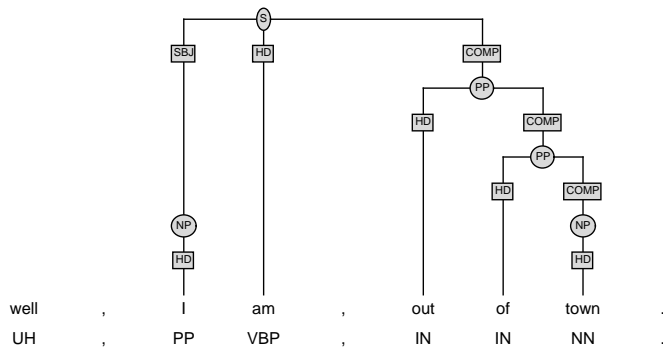
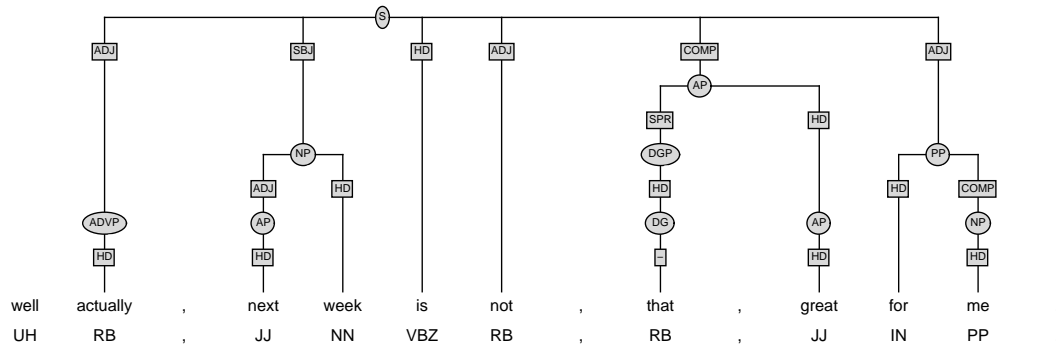
Stylebook for the English Treebank



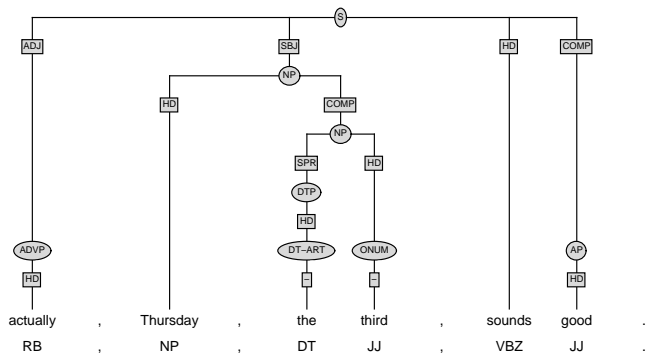
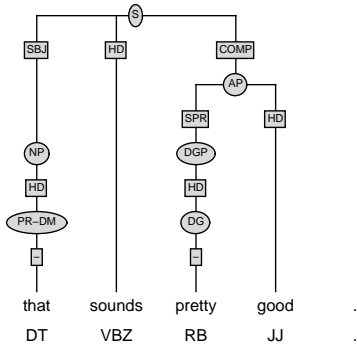
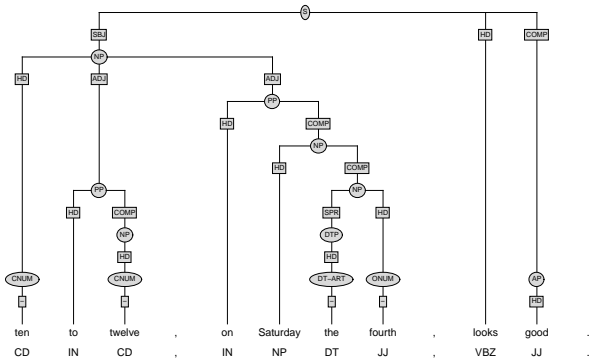
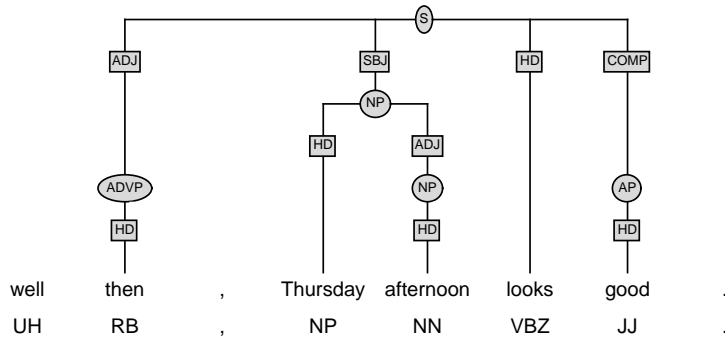
5.3 Copula Constructions

Verbs like *be*, *have*, *look*, *sound* support copula constructions:

Verbmobil Report 241

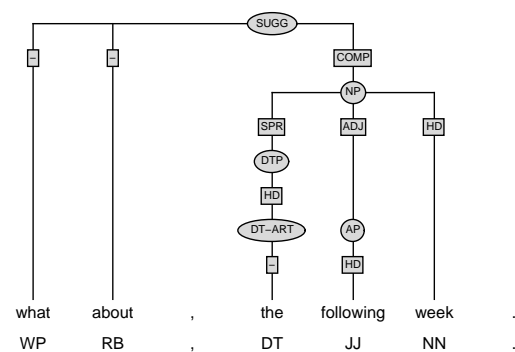
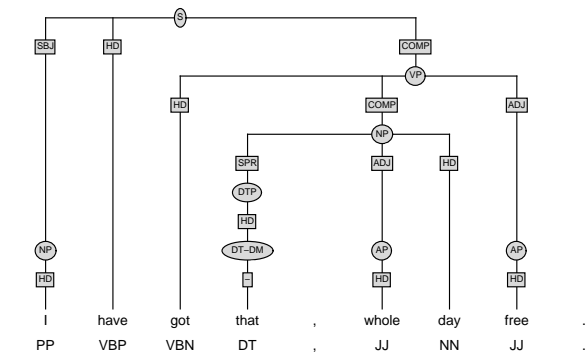
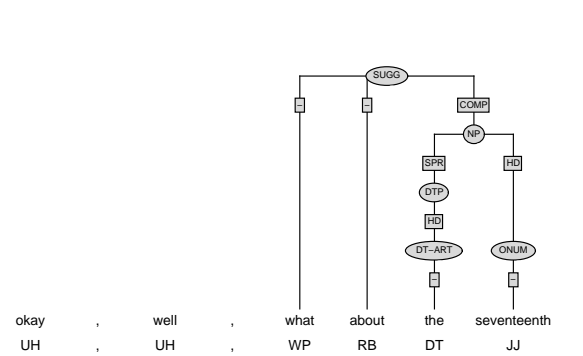
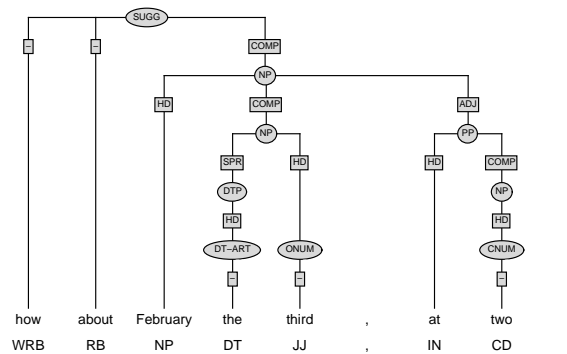


Stylebook for the English Treebank

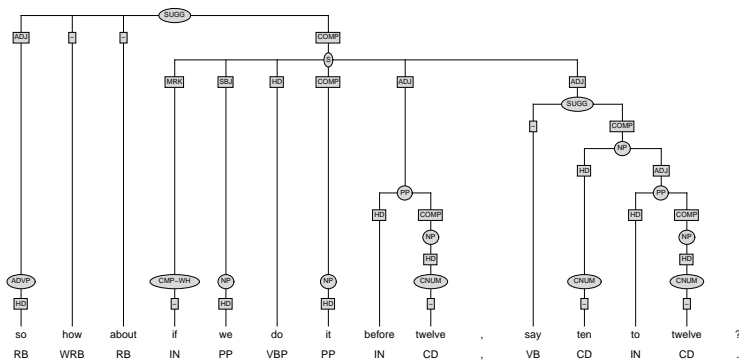
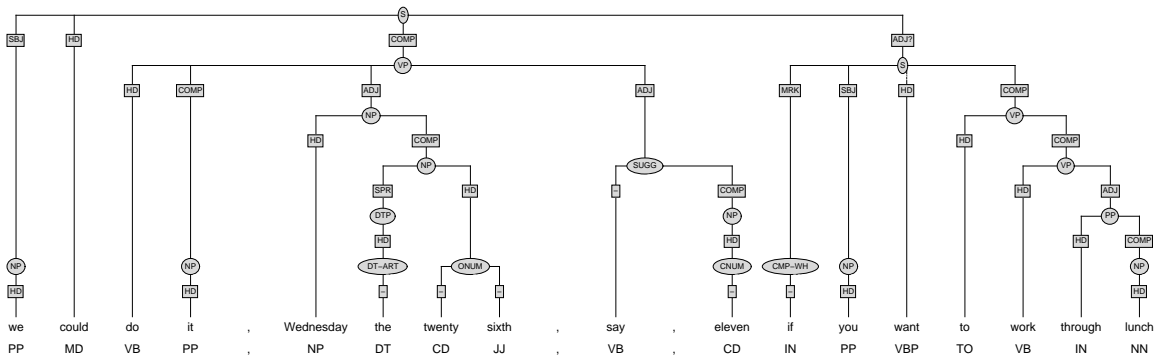
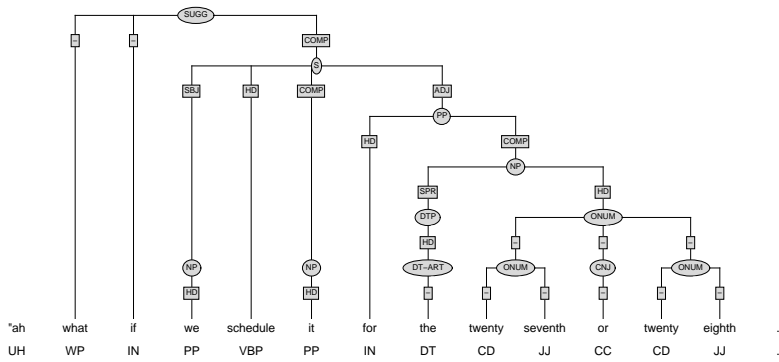
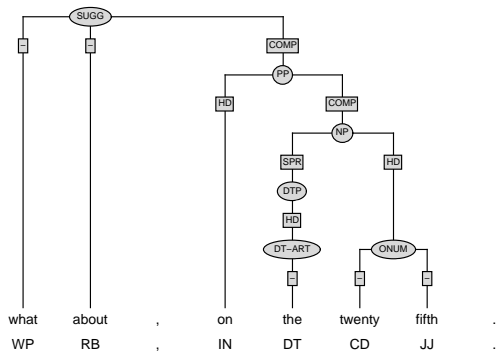


5.4 SUGG(estion)s

In the English treebank, utterances introduced with *how about*, *what about*, *what if*, and *say* are not given a syntactically complete infra-structure (i.e., they do contain only a COMP(lement), and no head), and they are treated as SUGG(estion)s (a Parentlabel, which should be viewed as a subset of S(entence)):

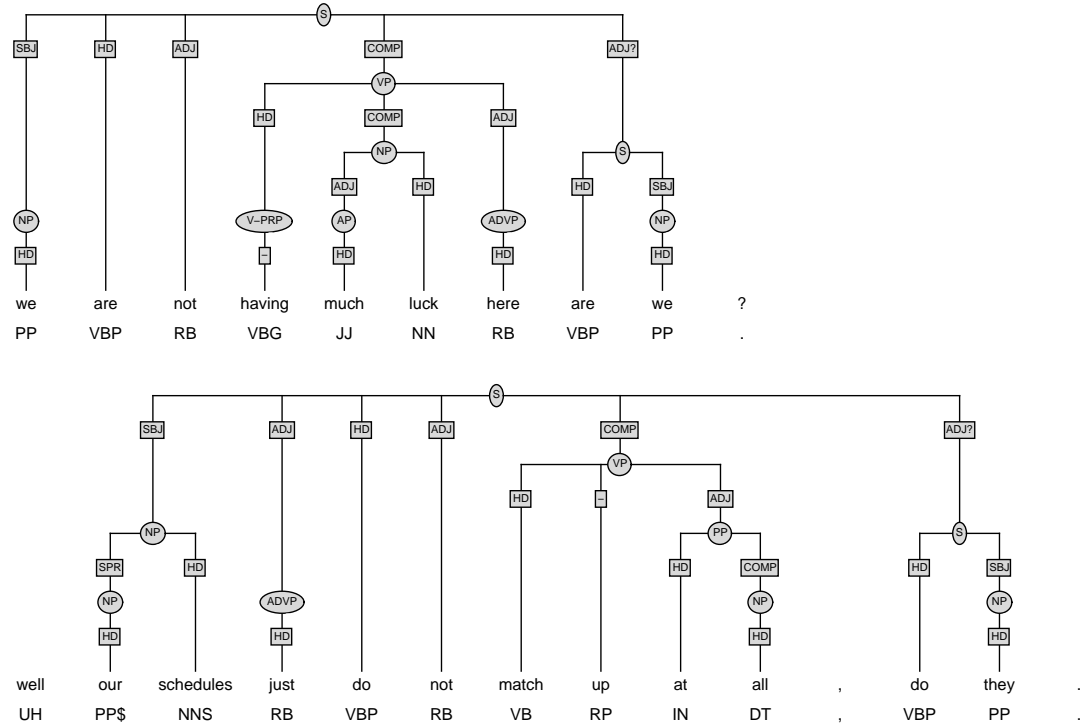


Stylebook for the English Treebank



5.5 The Periphery of the Sentence

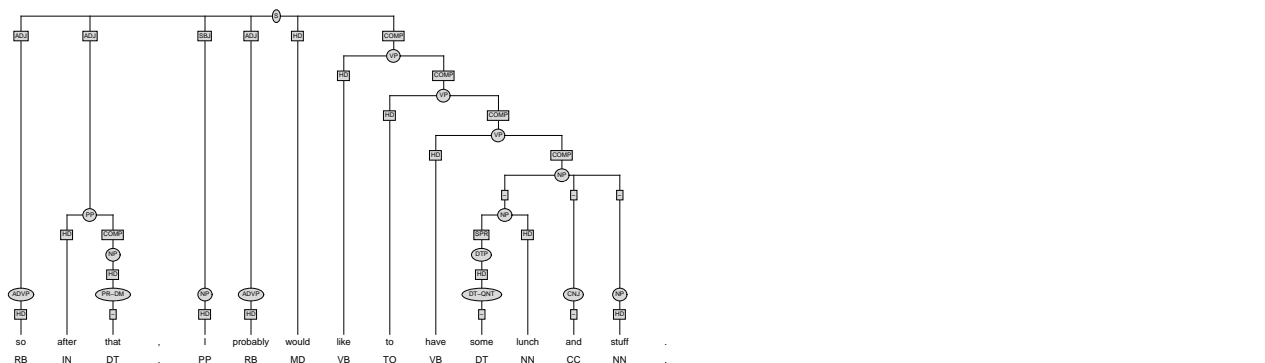
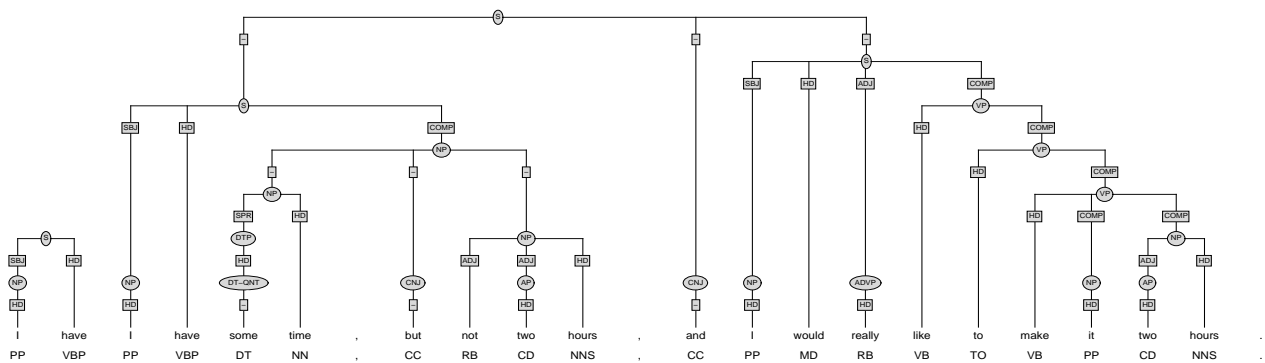
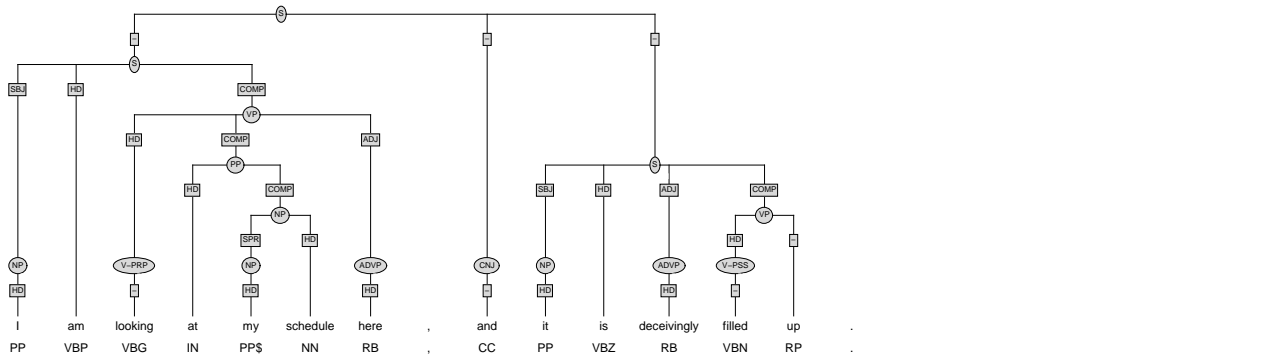
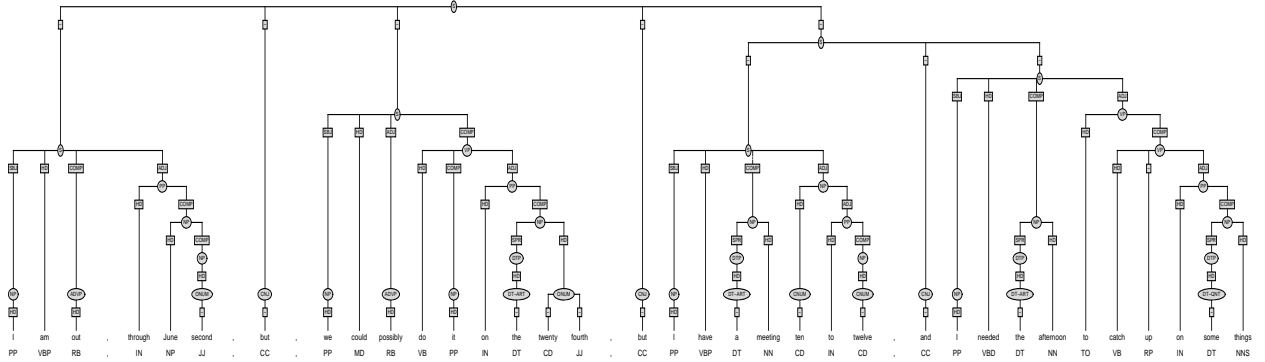
Sentences occurring in the periphery of another S(entence) (*root* sentence) with which they constitute a semantic unit, i.e., tail or tag questions, are treated in the English treebank as ADJ(uncts), modifying the root sentence (for a similar analysis in HPSG see also (Bender and Flickinger 1999)):



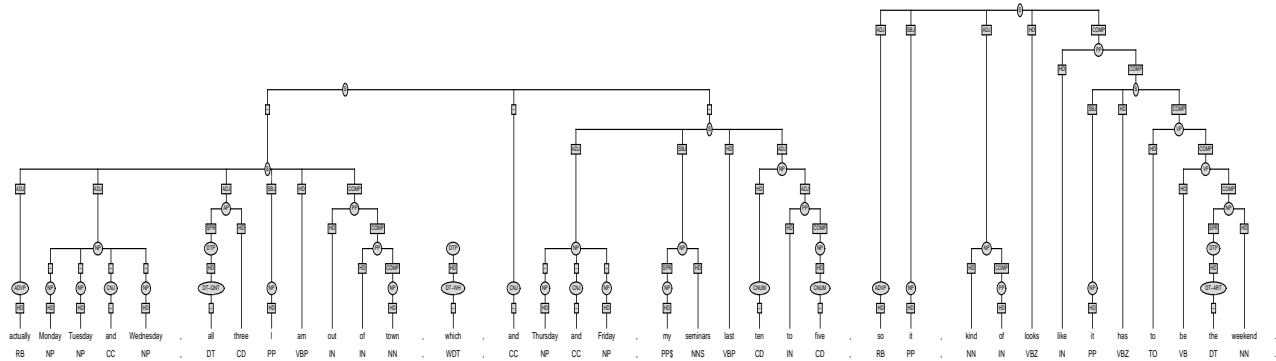
5.6 Coordination

Within coordinations in the English treebank, the conjuncts along with the conjunctive word are first projected to phrases, and then they are attached to the mother node (ternary branching with conjunction in the middle). The edge labels below the mother node of the coordination are empty. This scheme is the same for all syntactic categories:

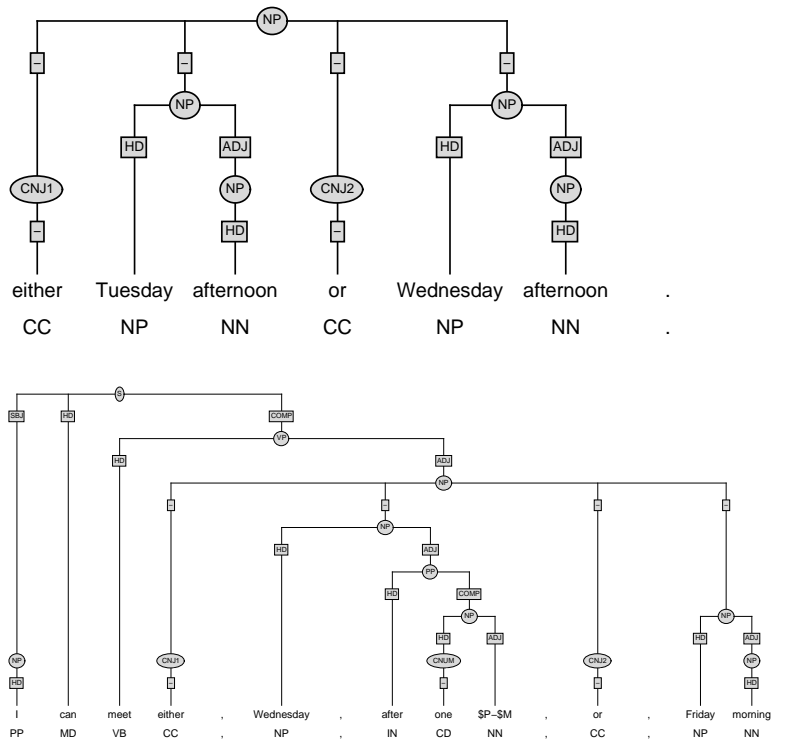
Stylebook for the English Treebank

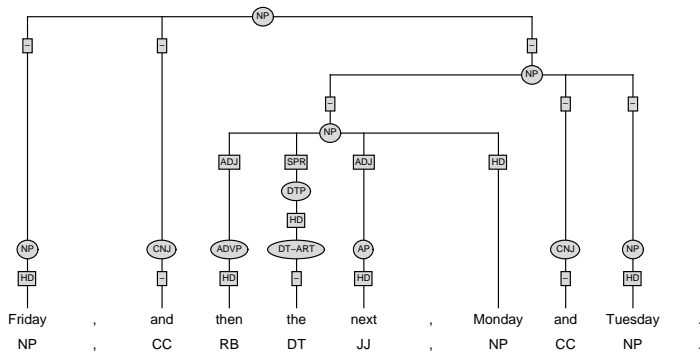


Coordinations with more than two conjuncts are treated as flat n -ary branching structures:



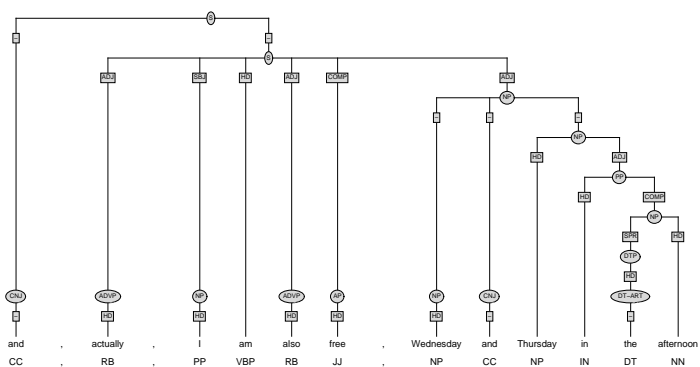
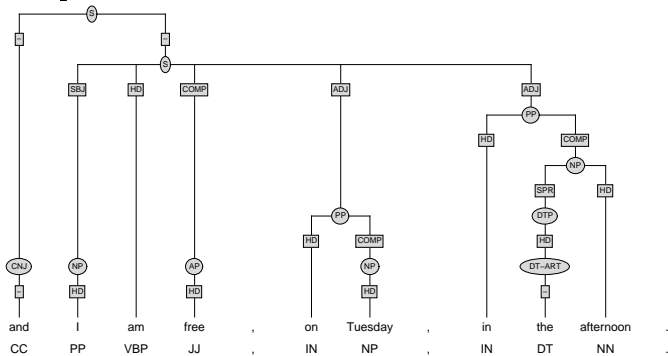
Either-or structures form a unit which is semantically closely related to coordinations and can be therefore treated in a parallel way:

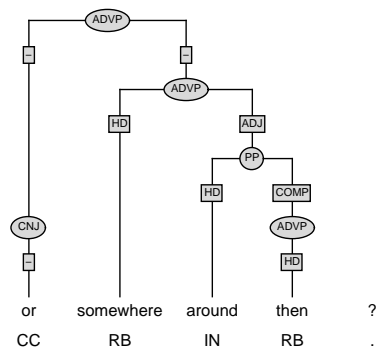
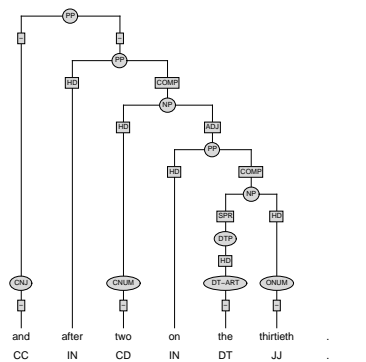
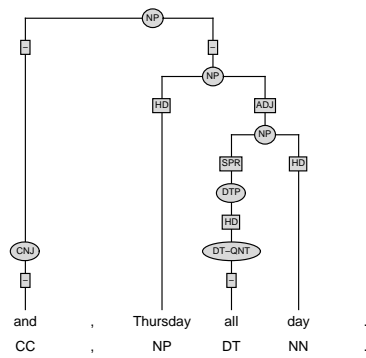




5.6.1 Isolated Conjuncts

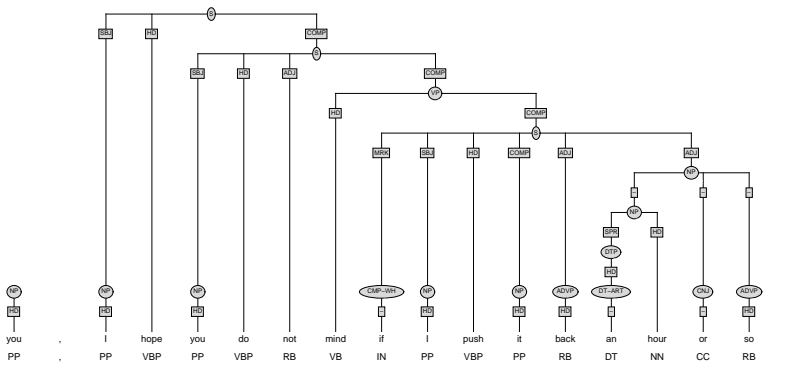
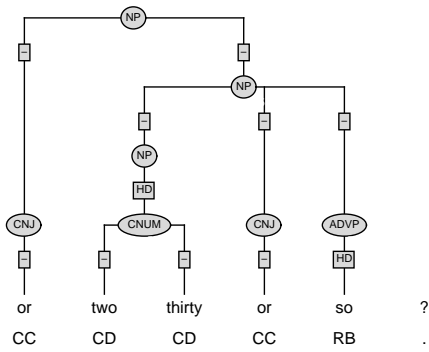
In VERBMOBIL, conjuncts often occur isolated and not as a *complete coordination* (i.e., a coordination with at least two conjuncts). In this case, the conjunct is attached to the conjunctive word one level higher, similar to the attachment in complete coordination:





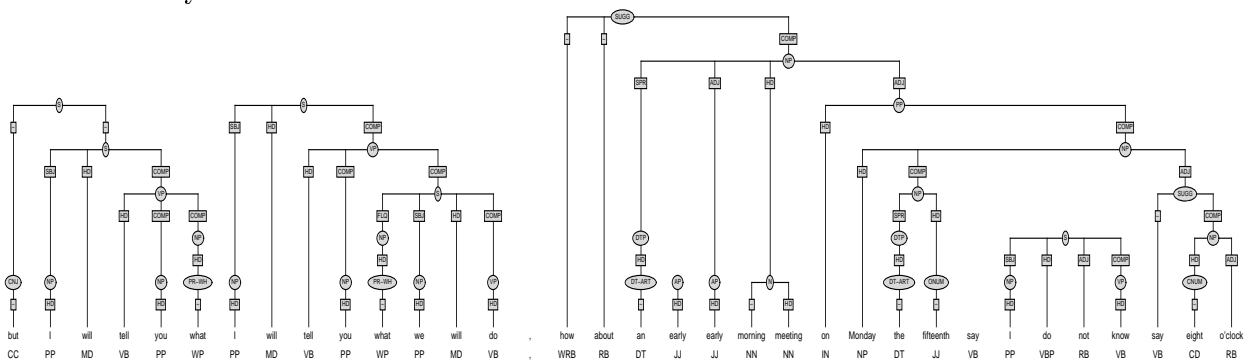
5.6.2 Unequal Conjuncts

In some cases of coordination the phrases are not of the same category. In this case, the strategy adopted for the English treebank is to choose the syntactic category of the **left-most** conjunct as the category of the whole coordination:



5.7 Parentheses

Parentheses occur in the English treebank as interjective utterances within sentences. They are not attached to the sentence inside which they appear. Most of the times they form small sentences of their own:

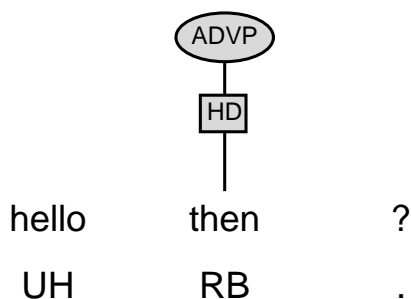
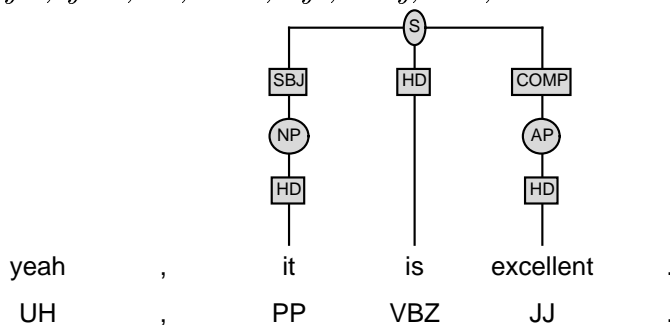


5.8 Discourse Markers

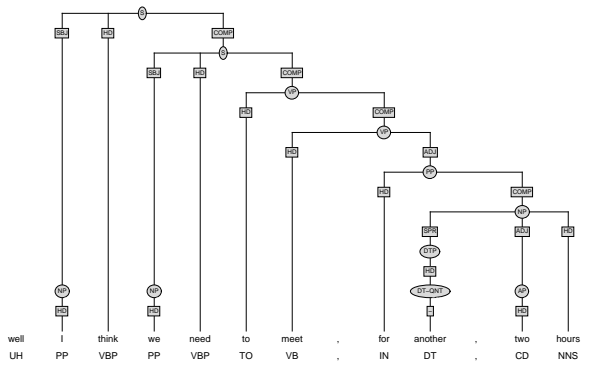
Discourse markers are sentence external words or phrases contributing to the course of the conversation. Expressions of greeting, apologising, thanking, sentence introducing utterances, short emotional utterances are considered to be discourse markers. All of them in the English treebank are pos-tagged as interjections (UH), and they are treated only on the terminal level.

Discourse markers are never attached to the sentence, but they can also occur as isolated expressions within the sentence boundaries.

Typical discourse markers in the English treebank are the following:
yes, yeah, no, hello, bye, okay, well, thanks



Stylebook for the English Treebank



Chapter 6

Conclusion

The purpose of this report has been to describe the design principles and the annotation scheme for the Tübingen treebank of English.

We have shown that the linguistic annotations of the Tübingen English treebank, which by the project completion counts 30 000 entries, pertain to the levels of morpho-syntax (part-of-speech tagging), syntactic phrase structure and function-argument structure.

In the Verbmobil context, the treebank was utilized as a data resource for a variety of processing modules in the Verbmobil system, including as a training data of stochastic parsers, as an on-line resource for the tree construction algorithm of the chunk parser (cf., (Hinrichs et al. 2000b)), and as a resource for the development of semantic construction rules and translation transfer rules.

In order, though, to ensure reusability of the data for purposes beyond the Verbmobil project, we have shown in this report that the treebank annotations follow accepted guidelines for corpus annotation (cf., (Leech 1993)).

Thus, this report is intended not only as a guide for the treebank annotators in Tübingen, but also as a reference for interested parties who will work directly with the annotated treebank data or who are interested in the construction or use of treebank data for English.

References

- Bender, E., and D. Flickinger. 1999. Peripheral constructions and core phenomena: Agreement in tag questions. In G. Webelhuth, J.-P. Koenig, and A. Kathol (Eds.), *Lexical and Constructional Aspects of Linguistic Explanation*, 199–214. Stanford: CSLI.
- Brants, T., and W. Skut. 1998. Automation of treebank annotation. In *Proceedings of the Conference on New Methods in Language Processing (NeMLaP3)*, Sydney, Australia, 49–57.
- Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy*.
- Burger, S. 1997. Transliteration spontansprachlicher Daten - Lexikon der Transliterationskonventionen - VERBMOBIL II. Technical Report 56, Verbmobil.
- Flickinger, D., A. Copestake, and I. A. Sag. 2000. HPSG Analysis of English. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation, Artificial Intelligence*, 255–265. Berlin, Heidelberg, New York: Springer-Verlag.
- Hinrichs, E. W., J. Bartels, Y. Kawata, V. Kordoni, and H. Telljohann. 2000a. The Tübingen Treebanks for Spoken German, English, and Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation, Artificial Intelligence*, 552–576. Berlin, Heidelberg, New York: Springer-Verlag.
- Hinrichs, E. W., S. Kübler, V. Kordoni, and F. H. Müller. 2000b. Robust Chunk Parsing for Spontaneous Speech. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation, Artificial Intelligence*, 163–182. Berlin, Heidelberg, New York: Springer-Verlag.
- Kawata, Y., and J. Bartels. 2000. Stylebook for the Japanese Treebank in VERBMOBIL. Technical Report 240, Verbmobil.
- Leech, G. 1993. Corpus annotations schemes. *Literary and Linguistic Computing* 8.4:275–281.
- Plaehn, O. 1998. ANNOTATE: Bedienungsanleitung. NEGRA Project. Technical report, Saarbrücken, Germany: Universität des Saarlandes, Computerlinguistik.
- Pollard, C., and I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Stylebook for the English Treebank

- Santorini, B. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science, University of Pennsylvania. Available as Technical Report MS-CIS-90-47.
- Skut, W., B. Krenn, T. Brants, and H. Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP), Washington, D.C.*
- Stegmann, R., H. Telljohann, and E. W. Hinrichs. 2000. Stylebook for the German Treebank in VERBMOBIL. Technical Report 239, Verbmobil.
- Wahlster, W. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin, Heidelberg, New York: Springer-Verlag.

Appendix: The Labels Used in the Treebank

Part-of-speech labels

CC: Coordinating conjunction
CD: Cardinal number
DT: Determiner
EX: Existential there
FW: Foreign word
IN: Preposition or subordinating conjunction
JJ: Adjective
JJR: Adjective, comparative
JJS: Adjective, superlative
LS: List item marker
MD: Modal
NN: Noun, singular or mass
NNS: Noun, plural
NP: Proper noun, singular
NPS: Proper noun, plural
PDT: Predeterminer
POS: Possessive ending
PP: Personal pronoun
PP: Possessive pronoun
RB: Adverb
RBR: Adverb, comparative
RBS: Adverb, superlative
RP: Particle
SYM: Symbol
TO: to
UH: Interjection
VB: Verb, base form
VBD: Verb, past tense
VBG: Verb, gerund or present participle
VBN: Verb, past participle
VBP: Verb, non-3rd person singular present
VBZ: Verb, 3rd person singular present
WDT: Wh-determiner
WP: Wh-pronoun
WP: Possessive wh-pronoun
WRB: Wh-adverb

- ,: Comma
- :: Sentence-final punctuation

Edge Labels

Edge labels indicate the grammatical functions of phrases.

- HD:** Head
- COMP:** Complement
- SPR:** Specifier
- SBJ:** Subject
- SBQ:** Subject,WH-
- SBR:** Subject,REL
- ADJ:** Adjunct
- ADJ?:** Adjunct?
- FLL:** Filler
- FLQ:** Filler,WH-
- FLR:** Filler,REL
- MRK:** Marker
- : for intentionally empty edge labels

Node Labels

Node labels in general indicate the syntactic category of phrases.

- AP:** Adjective Phrase
- APS:** Adj-headed sm.clause
- ADVP:** Adverb Phrase
- CMP:** Complementizer
- CMP-WH:** Complementizer,WH-
- CNJ:** Conjunction(single)
- CNJ1:** Conjunction(1 of 2)
- CNJ2:** Conjunction(2 of 2)
- DG:** Degree(non-wh)
- DG-WH:** Degree-WH(how...)
- DGP:** Degree Phrase
- DT-ART:** Det,Article(the,a)
- DT-DM:** Det,Demonstrative
- DT-QNT:** Det,Quantifier(every)

Verbmobil Report 241

DT-R: Det,Rel.clause
DT-WH: Det,Wh-(which,whose)
DTP: Det.Phrase
N: Noun,Common
CNUM: N,Cardinal Number
ONUM: N,Ordinal Number
NP: Noun Phrase
NPS: Noun-headed sm.clause
PR-DM: PR,Demonstrative
PR-WH: PR,WH-
PR-R: PR,Relative
PP: Prepositional Phrase
PPS: Prep-headed sm.clause
SUGG: Suggestion("How about Tuesday?")
S: Sentence(VP w/subject)
V-G: Verb,gerund
V-PRP: Verb,present participle
V-PSS: Verb,passive participle
VP: Verb Phrase(S if sub Vs sister)