Shallow-Parsing Stylebook for German

Frank Henrik Müller
fhm@sfs.uni-tuebingen.de

February 27, 2002

Abstract

The presented system provides a shallow syntactic annotation for unrestricted German text. It requires POS-annotated text and annotates the layers of chunks, topological fields and clauses. This stylebook gives an overview of the various categories annotated in those different layers. The methodology of the annotation process is mentioned in those cases where it has an impact on the annotation scheme. Example sentences are taken from real language data, but were simplified where necessary.

1 Introduction

The presented system works with POS-annotated text, which uses standard taggers (e.g. [Bra98]) to annotate text according to the STTS-tagset [STTS95]. After this, topological fields, clauses and chunks are annotated before language data is made available for global sentence analysis (See [MU01]). The decision to split the annotation task has been made on the basis of a decision to first annotate structure which can be handled with transducers using syntactic restrictions. More powerful formalisms may be used afterwards.

2 The Chunk Layer

2.1 The Notion of *Chunks*

Chunks are defined as non-recursive continuous kernels of phrases. This means that chunks may contain chunks of other categories but that they may

```
[APPR bei]
[ART der]
[PC
   [APPR durch]
   [NC
      [AJAC
         [ADJA jahrelange]
      ]
      [NN Fehlentscheidungen]
   ]
]
[NC
   [AJAC
      [ADJA hochverschuldeten]
   ]
   [NN Bahn]
]
```

Figure 1: Chunk structure of a complex prepositional phrase

not contain chunks of the same category. The prepositional chunk (PC) *durch jahrelange Fehlentscheidungen* in the example in **Figure 1** is part of a complex adjective phrase with the adjective *hochverschuldeten* as its head. There are two separate chunks, however, as an adjective chunk cannot contain another adjective chunk (like the one contained in the PC).[1] As a noun chunk (NC) cannot contain another NC for the same reason, the PC is not part of the following NC, either. This means that the article *der* cannot be in one chunk with *hochverschuldeten Bahn* as this is a discontinuous structure now, because the PC is a constituent between the article and the rest of the NC which cannot be integrated into the NC. Thus, constituents like articles or prepositions which cannot be attached to their respective chunks are left 'stranded'.

There are some other constituents which are never attached to any chunk. Those are subordinators, some adverbial interrogatives and interjections. Coordinators may also be left unattached, when they do not coordinate chunks.

[1]Please note: There would also be two chunks, if there was no adjective chunk in the PC, the definition of recursion being that a constituent **may** contain a constituent of the same type.

2.2 Outline of Chunk Types

There are five major types of chunks. Verb chunks (VC_)[2], noun chunks (NC), adjective chunks (AJ_C), adverb chunks (AVC) and prepositional chunks (PC). They are discussed in the order of their recognition in the annotation process. Prepositional Chunks always contain a noun chunk and noun chunks may contain adverb chunks and adjective chunks. Verb chunks are contained in no chunk and cannot contain any other chunk.

Verb chunks play a special role in the chunk structure, as they are both chunks and part of the clausal frame. This fact and the fact that verb chunks (and their combination) already contain a lot of information about the structure and type of the sentence they occur in has lead to a very fine grained distinction among the different verb chunks, which distinguishes them from the other chunks.

2.3 The Chunk Types

2.3.1 Truncated Chunks

Truncated Chunks (_CTRUNC) are chunks with fragmentary words. Those words receive the tag TRUNC in the STTS. They cannot usually be assigned to the appropriate word category, as they lack important morphological information. Using contextual information, it is, however, very often possible to detect the proper word category. Fragmentary words are, thus, chunked into a chunk corresponding to the respective word category, which is then treated like any ordinary chunk of its category (See **Figure 2**). Truncated chunks are, thus, not a chunk category on its own, but rather belong – according to their head words – as a subcategory to the respective chunk category. Truncated words, are, thus, treated just like their non-fragmentary counterparts. In cases where disambiguation of the word category is impossible, truncated words are not chunked at all.

2.3.2 Verb Chunks

Verb chunks are categorized on the basis of their syntactic distribution and their inner structure. As regards syntactic distribution, there are four main types of verb chunks: VCL-, VCR-, VCF- and VCE-. VC$\boxed{\text{L}}$-Chunks are chunks which are the **left** part of the clausal frame and VC$\boxed{\text{R}}$-Chunks are chunks which are the

[2]The '_' stands for one letter, if it occurs within a chunk name, and for one or more letters, if it occurs at the beginning or end of a chunk name.

```
[PC
     [APPR in]
     [NCC
          [NC
               [ART einer]
               [NCTRUNC
                    [TRUNC Rundfunk-]
               ]
          ]
          [KON und]
          [NC
               [NN Fernsehansprache]
          ]
     ]
]
```

Figure 2: Truncated noun chunk (NCTRUNC)

right part of the clausal frame. VC$\boxed{\text{F}}$-Chunks are chunks which in the basic word order would be the right part of the clausal frame but which are **front**ed and, thus, topicalized. VC$\boxed{\text{E}}$-Chunks occur in **Ersatz**infinitiv constructions. They contain the finite verb, which – without the Ersatzinfinitiv occuring – would be part of the right part of the clausal frame but which is moved to the left of it in such a construction. Currently, VC$\boxed{\text{E}}$-Chunks are only recognized if no constituents intervene between the VC$\boxed{\text{E}}$-Chunk and the VC$\boxed{\text{R}}$-Chunk.

The following letters in the name of the verb chunk correspond to the second and third letters in the verb tag, which denote verb type (V=lexical, A=auxiliary, M=modal) and finiteness (F=finite, I=infinitive, P=perfect participle)[3]. The sequence of the verbs in the chunk type name corresponds to the syntactic dependence, which is the opposite of the sequence of the occurence of the verbs in the sentence. Thus, in the sequence *Dialoge, die nicht ohne Weiteres zu verstehen sind* the verbal complex *zu verstehen sind* is assigned the chunk type VC$\boxed{\text{R}}$ for right part of clausal frame, $\boxed{\text{AF}}$ for auxiliary finite, $\boxed{\text{VZ}}$ for lexical verb/infinitive with 'zu' (See **Figure 3**). As they are parts of verbs, verbal particles are also included in the class of the verb chunks with the chunk type name VCRPT.

[3]An exception is being made in the treatment of the infinitive with 'zu', where 'I' is replaced with 'Z' in the chunk type name (See **Figure 3**) and with the imperative where a 'B' is assigned.

```
[NC
     [NN Dialoge]
]
[$, ,]
[NC
     [PRELS die]
]
[AVC
     [PTKNEG nicht]
]
[PC
     [APPR ohne]
     [NC
          [NN Weiteres]
     ]
]
[VCRAFVZ
     [PTKZU zu]
     [VVINF verstehen]
     [VAFIN sind]
]
```

Figure 3: Verb chunk in relative clause

The denotation of the verb chunk names has been chosen in such a way as to make the annotation system transparent and, thus, user-friendly. The idea was to create a system which provides the user with chunk type names which are self-explanatory. The system is also easily extensible using a method like this. As a side effect, the hierarchical structure of the chunk type names makes the corpus more easily accessible to query tools, as e.g. verb chunks governed by a finite auxiliary can be searched for with an expression like 'VC(L|R|F)AF.*'. Although decisions like this seem to be purely technical, they are nevertheless important because an inadequate annotation scheme may lead to mistakes in the evaluation of the corpus data or even to mistakes in subsequent annotation.

2.3.3 Noun Chunks

Noun chunks are the most common chunks. They consist at least of a noun

```
[PC
     [APPRART im]
     [NC
          [NN Interesse]
     ]
]
[NC
     [PIAT aller]
     [NN Mitgliedstaaten]
]
```

Figure 4: Noun chunks

```
[PC
     [APPR um]
     [NC
          [CARD sechs]
     ]
]
```

Figure 5: Cardinal as the head of a noun chunk

```
[PC
     [APPR aus]
     [NC
          [AJAC
               [AVC
                    [ADV bloß]
               ]
               [ADJA wirtschaftlichen]
          ]
          [NN Motiven]
     ]
]
```

Figure 6: Modifying adverb chunk in attributive adjective chunk

```
[NCC
     [NC
          [ART die]
          [AJAC
               [ADJA großen]
          ]
          [NN Gnus]
     ]
     [KON und]
     [NC
          [NN Zebras]
     ]
]
```

Figure 7: Two coordinated NCs

(or a cardinal number) as a head word (with the exception mentioned in section 2.3.4 and below) and of optional determiners, adverb chunks or attributive adjective chunks (See **Figures 4, 11 and 10** and the examples mentioned in section 2.3.4). Noun chunks can only be contained in prepositional chunks. As the definition of chunks does not allow recursive structures, postmodifying prepositional or nominal constituents may not be part of a noun chunk (See **Figure 4**). As mentioned in section 2.1, discontinuous chunks are not allowed and, thus, determiners might in some cases be separated from the noun chunk (See **Figure 1**). In case of preposition-article contraction (APPRART), the contraction is annotated in its function as a preposition (See **Figure 4**).

As pronouns are normally not modified, they are the only element of a noun chunk when they occur (See **Figure 16**). Adverb chunks are only included in a noun chunk when they occur after clear indicators for the beginning of a noun chunk and before the head word, because in the other cases their attachment is ambiguous. In case they modify an adjective or a cardinal number, they are attached to the AJAC chunk. Clear indicators of the beginning of a noun chunk are determiners or attributive adjectives, but also prepositions (which are themselves not part of a noun chunk) (See **Figure 6**). When two noun chunks are coordinated by a coordinator, they are chunked as a coordinated noun chunk NCC. When the coordinated noun chunks are modified by an attributive adjective chunk, the adjective chunk becomes part of the first noun chunk as it is impossible to decide on

```
[NC
     [ART Der]
     [AJAC
          [ADJA älteste]
     ]
     [NN Künstler]
]
[VCLAF
     [VAFIN ist]
]
[NC
     [NN Jahrgang]
]
[NC
     [CARD 1929]
]
[$, ,]
[NCell
     [ART der]
     [AJAC
          [ADJA jüngste]
     ]
]
[NC
     [CARD 1963]
]
```

Figure 8: Sentence containing elliptical NC

```
[NCC
     [NCell
          [ART das]
          [AJAC
               [ADJA neunzehnte]
          ]
     ]
     [KON und]
     [NC
          [ART das]
          [AJAC
               [ADJA zwanzigste]
          ]
          [NN Jahrhundert]
     ]
]
```

Figure 9: Elliptical NC coordinated with NC

the chunk level whether the adjective chunk modifies both nouns or the first noun only (See **Figure 7**). Very often world knowledge or the understanding of the wider textual context would be required to solve the ambiguity. The same applies to articles which might refer to both noun chunks or the first noun chunk only.

Elliptical noun chunks (NCell) (i.e. noun chunks without a noun as their head word) must consist of at least an attributive adjective chunk. There are various kinds of elliptical noun chunks (See section 2.3.4). In cases where an elliptical noun chunk is coordinated with the noun chunk containing its head word, those noun chunks are also chunked as a coordinated noun chunk (See **Figure 9**).

2.3.4 Attributive Adjective Chunks

There are two main types of adjective chunks: attributive adjective chunks and predicative/adverbial adjective chunks. The distinction between them is made on the basis of the POS-tags of the head word of the chunk (i.e. the adjective tags ADJA and ADJD).

AJAC-chunks are chunks with an attributive adjective (or cardinal number) as their head. The definition of an attributive adjective being that it modifies a noun,

```
[NC
     [AJAC
          [AVC
               [ADJD furchtbar]
          ]
          [ADJA militaristische]
     ]
     [NN Trick-Aufnahmen]
]
```

Figure 10: AJAC chunks with ADJA modified by ADJD

```
[NC
     [ART die]
     [AJAC
          [AVC
               [ADJD ungefähr]
          ]
          [CARD vierzig]
     ]
     [AJAC
          [ADJA jungen]
     ]
     [NN Männer]
]
```

Figure 11: Cardinal number as head of an attributive adjective chunk

```
[NC
      [ART die]
      [AJACC
            [AJAC
                  [ADJA große]
            ]
            [$, ,]
            [AJAC
                  [ADJA weite]
            ]
      ]
      [NN Welt]
]
```

Figure 12: AJAC chunks coordinated by comma

the AJAC-chunk is always part of a noun chunk. As mentioned in section 2.1, complements of the adjective are not part of the AJAC-chunk. However, an AVC-chunk containing an ADJD can be part of an AJAC-chunk as a modifier (See **Figure 10**).

AJAC-chunks may be coordinated in two ways: with commas or with a coordinator (See **Figure 12** and **Figure 13**). In this case – analogous to the treatment of coordinated AJVC-chunks – they are projected to an AJAC$\boxed{\text{C}}$ chunk. Two immediately successive attributive adjective chunks are not projected to a coordinated chunk, as they are not in a relation of coordination (See **Figure 14**). In cases in which there is no head noun after the attributive adjective chunk, this chunk – together with other elements like determiners – forms the noun chunk. The head noun of this elliptical noun chunk (NCell) may be inherent or it may precede (**Figure 8**) or follow (**Figure 9**) the elliptical noun chunk.

2.3.5 Predicative Adjective Chunks/Adverbial Adjective Chunks

The tag ADJD, on the basis of which the chunk type AJVC is recognized, is assigned on the grounds of morpho-syntactic features, but the word to which the tag ADJD is assigned may function either as an adjective (See **Figure 15**) or as an adverb (See **Figure 16**). Obviously, in the tagset, no distinction was made, as it is impossible to draw it without making a full parse. As this is still true as regards

```
[NC
     [AJACC
          [AJAC
               [ADJA ausgekochte]
          ]
          [KON und]
          [AJAC
               [ADJA geschäftstüchtige]
          ]
     ]
     [NN Musiker]
]
```

Figure 13: AJAC chunks coordinated by *und*

```
[NC
     [AJAC
          [ADJA traditionelle]
     ]
     [AJAC
          [ADJA klassische]
     ]
     [NN Musik]
]
```

Figure 14: Two AJAC chunks

```
[AVC
     [ADV Auch]
]
[NC
     [ART die]
     [NN Kelten]
]
[VCLAF
     [VAFIN waren]
]
[AJVC
     [ADJD eitel]
]
```

Figure 15: AJVC as a predicative adjective chunk

the chunking level, the decision is not made on this level, either; especially as it does not lead to any negative effects on the chunking level. The chunk is, thus, left partially disambiguated, the disambiguation being left open for further annotation processes.

AJVC-Chunks may contain more than one ADJD, as one ADJD can be modified by another one. In that case the modifying ADJD is first projected to an AVC (See **Figure 16**). In the case of coordination of AJVC-chunks, they are projected to an AJVC$\boxed{\text{C}}$-Chunk (See **Figure 17**). Coordination may occur with or without coordinator.

2.3.6 Adverb Chunks

In the cases where the attachment of adverbs is ambiguous, the site of their attachment is not specified. The adverb chunk is then not part of the modified chunk. In most cases, adverb chunks consist of a single adverb only. Adverb chunks cannot contain any other constituents than adverbs (including *nicht*, which is tagged PTKNEG). Coordinated adverb chunks are grouped into a chunk labelled AVCC analogous to the adjective chunks and the noun chunks (See **Figure 18**).

```
[AVC
     [ADV Da]
]
[VCLVF
     [VVFIN kommt]
]
[NC
     [PPER Dir]
]
[NC
     [ART das]
     [AJAC
          [ADJA normale]
     ]
     [NN Leben]
]
[AJVC
     [AVC
          [ADJD richtig]
     ]
     [ADJD langweilig]
]
[VCRPT
     [PTKVZ vor]
]
```

Figure 16: AJVC as an adverbial chunk

```
[AJVCC
     [AJVC
          [ADJD schnell]
     ]
     [KON und]
     [AJVC
          [ADJD frisch]
     ]
]
```

Figure 17: Coordinated AJVC-chunks

```
[AVCC
     [AVC
          [ADV nachts]
     ]
     [KON oder]
     [AVC
          [ADV sonntags]
     ]
]
```

Figure 18: Coordinated adverb chunks

```
[PC
     [NC
          [AJAC
               [CARD drei]
          ]
          [NN Wochen]
     ]
     [APPO lang]
]
```

Figure 19: PC with postposition

```
[PC
     [APPR von]
     [NC
          [NN Anfang]
     ]
     [APZR an]
]
```

Figure 20: PC with circumposition

2.3.7 Prepositional Chunks

Prepositional chunks typically consist of a preposition and a noun chunk. In most cases, the preposition precedes the noun chunk. In some cases it follows the noun chunk (postposition) (See **Figure 19**) or includes it (circumposition) (See **Figure 20**). Contractions of pronouns and prepositions (e.g. *darauf*, *deswegen* or *hiermit*), which are tagged PROAV, may be the only constituents of a PC. Sometimes, the head of a prepositional chunk is a token tagged as an adverb (See **Figure 21**). In some cases, a prepositional chunk may contain what might be called a complex preposition, the token *bis* followed by a preposition (See **Figure 22**).

```
[PC
     [APPR seit]
     [ADV gestern]
]
```

Figure 21: Prepositional chunk with adverb head

```
[PC
     [APPR bis]
     [APPRART zum]
     [NC
          [AJAC
               [ADJA letzten]
          ]
          [NN Augenblick]
     ]
]
```

Figure 22: PC with ’complex’ preposition

Table 1: Overview of the Chunk Labels

Chunk Label	Definition
AJAC	attributive adjective chunk
AJACTRUNC	AJAC with truncated adjective
AJACC	at least two coordinated AJACs
AJVC	predicative adjective chunk/adverb chunk
AJVCTRUNC	AJVC with truncated adjective/adverb
AJVCC	at least two coordinated AJVCs
AVC	adverb chunk
AVCC	coordinated AVC
NC	noun chunk
NCTRUNC	NC with truncated noun
NCC	two NCs coordinated by coordinator
NCell	elliptical noun chunk (i.e. without head noun)
PC	prepositional chunk
VC_	verb chunk
VCTRUNC	verb chunk with truncated verb
VCL_	verb chunk as left part of clausal frame
VCR_	verb chunk as right part of clausal frame
VCF_	verb chunk in topicalized (fronted) position

3 The Topological Field Layer

3.1 The Notion of *Topological Fields*

Topological fields describe sections in the German sentence with regard to the distributional properties of the verb. In German, the verb complex is divided into two parts in the affirmative sentence with the finite verb coming first and the rest of the verb complex following later on to the end of the sentence. This construction is called the Satzklammer (clausal frame), the finite verb being the left part of the clausal frame (Linke Klammer, LK) and the non-finite verb complex being the right one (Rechte Klammer, RK). Thus, in the affirmative sentence, the sections preceding the finite verb may be called the Vorfeld (front field, VF), the section included in the clausal frame the Mittelfeld (middle field, MF) and the section following the non-finite verb complex the Nachfeld (post field, NF). There may also

be a KOORD-Feld (coordinator field, KOORDF). Sentences in which the verb is fronted (e.g. imperatives and yes/no-questions) are like affirmative sentences except that they lack a VF. Introduced subordinated sentences are different in that the finite verb goes together with the verb complex (as the last element in it). They also lack a VF, and the LK is occupied by the complementizer field (CF) in them.

The model of topological fields describes the distribution of constituents relative to the clausal frame. It is therefore primarily a distributional model. It does not give any account of the verb-argument structure and it does not reveal the relation of the constituents within the topological fields, either. In fact, the very structure of constituents within topological fields is left open. The model still has some clear advantages from both a theoretical and an annotation perspective: As regards the theoretical perspective it is for example important to point out that a lot of constituent order phenonema can be described relative to topological fields. As regards the annotation of further grammatical information, it remains to be seen whether the constituent order in the different topological fields may be utilized for annotation because, in German, there are very few syntactic restrictions in the constituent order. There are, however, a lot of syntactic preferences which may be utilized if connected with other information like morphological features and valency structure.

As regards automatic annotation, one of the main advantages which can be drawn from topological fields is that they are the skeleton of the sentence and that, thus, once topological fields are annotated, the outline of the sentence is known. The annotation of topological fields considerably reduces the scope of ambiguity because the verb is always part of the clausal frame and its arguments are always part of one of the corresponding fields. Without the annotation of topological fields the scope of arguments of the verbs is much wider, especially in complex sentences, in which, additionally, it is not clear which potential arguments belong to which verb. After the annotation of topological fields, the syntactic restrictions and preferences in them might be applied for further annotation (together with other linguistic information).

The advantages of annotating topological fields before annotating verb argument structure can be illustrated with **Figures 23** and **24**: **Figure 23** shows a sentence containing five verbs in three verb complexes. As the arguments may appear on both sides of the verb, it is by no means clear, where the respective arguments of the verb are. After the annotation of the topological fields, however, the scope is reduced: The arguments of *sagte* – as long as they are single phrases – may only occur in the preceding VF and the following MF. The same is true of the following clause, where the field structure shows that *könne . . . beendet werden*

is one verb complex and then it is clear that phrasal arguments may only occur in the VF, the MF or the NF. In the subclause in the NF it is again clear that the arguments of the verb must be in the MF.

Figure 23 illustrates how the scope of potential arguments is reduced by annotating field structure. It should be taken into account, however, that the annotation of field and clause structure is a shallow one like the one of the chunks. It is, thus, not always clear to which field or clause subclauses should be attached. In **Figure 23** it is left open whether the NF containing an adverbial clause is the NF of the preceding sublause or the NF of the main clause (i.e. the whole sentence). As regards the subclause *der Streit könne auf der Stelle beigelegt werden*, the annotation leaves it open that it is a subclause at all. This is done because without taking into account valency information it is not clear whether a sequence of Vorfeld, Linke Klammer and Mittelfeld is a separate main clause in an asyndetic construction (like in *Das Parlament debattiert, der Präsident handelt.*) or a subclause (like in **Figure 23**) Still, the pre-structuring achieved with the annotation of the fields is a solid base for further annotation. **Figure 24** shows that quite often the field structure is not ambiguous. The non-finite clause and the subclause in the VF of the sentence clearly are one constituent because there is typically just one constituent in the VF. The relative clause is clearly part of the MF because it is clearly within the clausal frame. This would of course also be true of any other subclause within the clausal frame.

3.2 Outline of Field Types

Some of the major characteristics of the different field types have already been mentioned at the beginning of section 3.1. A distinction can be made between two different types of fields. On the one hand, there are those fields which are part of the clausal frame. They can typically only contain tokens of a restricted number of Parts-of-Speech, namely complementizers and verbs. Those fields are the complementizer field (CF) and the finite verb as the LK (i.e. left part of the clausal frame) and the verb complex as the RK (i.e. right part of the clausal frame). As the fields containing chunks coincide with the verb chunks and as it is marked in the chunks whether they are LK or RK (See section 2.3.2), there is no further indication of those fields. On the other hand, there are those fields which can be described relative to the clausal frame. They can contain tokens of all other Parts-of-Speech. The constituent order in those fields is far more free than in those fields which are part of the clausal frame. Those fields are the VF, the MF, the NF and the Linksversetzung (LV). A special case is the KOORDF. This field just contains

```
{VF
    [NC
        .ART     Der
        .NN      Präsident ] }
[VCLVF
    .VVFIN   sagte ]
{MF
    [NC
        .ART     den
        [AJAC
            .ADJA    anwesenden ]
        .NN      Journalisten ] }
.$,      ,
{VF
    [NC
        .ART     der
        .NN      Streit ] }
[VCLMF
    .VMFIN   könne ]
{MF
    [PC
        .APPR    auf
        [NC
            .ART     der
            .NN      Stelle ] ] }
[VCRAIVP
    .VVPP    beendet
    .VAINF   werden ]
.$,      ,
{NF
    (SUB
        {CF
            .KOUS    wenn }
        {MF
            [NC
                .PRF     sich ]
            [NC
                .PIAT    beide
                .NN      Seiten ]
        [VCRVF
            .VVFIN   verständigten ] ) }
.$.      .
```

Figure 23: Complex Sentence

```
{VF
    (INF
        {CF
            .KOUI    Um }
        [VCRVZ
            .PTKZU   zu
            .VVINF   demonstrieren ] )
    .$,      ,
    (SUB
        {CF
            .KOUS    wie }
        {MF
            [NC
                .PDS     das ] }
        [VCRVF
            .VVFIN   funktioniert ] )
    .$,      , }
[VCLAF
    .VAFIN   wurde ]
{MF
    [NC
        .ART     die
        .NN      Rede ]
    [NC
        .ART     des
        .NN      Forschers ]
    .$,      ,
    (REL
        {CF
            [NC
                .PRELS   der ] }
        {MF
            [NC
                .ART     das
                [AJAC
                    .ADJA    neue ]
                .NN      Programm ] }
        [VCRVF
            .VVFIN   vorstellte ] )
    .$,      ,
    [PC
        .APPR    auf
        [NC
            .ART     einem
            .NN      Bildschirm ] ] }
[VCRVP
    .VVPP    gezeigt ]
.$.      .
```


Figure 24: Complex Sentence

one constituent (the coordinator). It is a field which may occur at the beginning of a clause.

3.2.1 The Complementizer Field (CF)

The CF only occurs in subordinated clauses introduced by a complementizer. It is always the left part of the clausal frame. CFs usually contain just one token (i.e. the complementizer). These complementizers may be relative pronouns (PRELS, PRELAT), interrogative pronouns and adverbial interrogative pronouns in indirect questions (PWAT, PWS, PWAV) or subordinators (KOUI, KOUS). In the cases in which the tokens are attributive, the whole noun chunk belongs to the CF (See sentences 1 and 2)[4]. In some cases, a complementizer may consist of a complex token like *so dass/daß* or *als ob* (See sentence 3). The CF can also be occupied by an expression like the one in sentence 4, which also introduces a subordinate clause.

(1) Sie stammen aus Ländern, [$_{CF}$ deren/PRELAT Regierungen] keinerlei Respekt für die Menschenrechte haben.

(2) Niemand weiß, [$_{CF}$ welches/PWAT Datum] das Dokument trägt.

(3) Die meisten Besucher kennt man, [$_{CF}$ so daß] die Sicherheitsprozedur entfällt.

(4) [$_{CF}$ Je mehr Dinge] man zu erledigen hat, desto mehr Zeit hat man.

3.2.2 The Linke Klammer (VCL_)

While the CF is the left part of the clausal frame in introduced subclauses, the VCL_ is the left part of the clausal frame in main clauses (See sentences 5 and 6) and non-introduced subclauses (See sentences 7 and 8). The VCL_ always just contains one finite verb of the categories lexical verb, auxiliary verb or modal verb.

(5) Ein Almbauer aus Bayrischzell [$_{VCLVF}$ verhindert] den Skisport am Wendelstein.

[4]The example sentences in this section and the following sections just contain the linguistic markup relevant for the respective sections.

(6) Jetzt [$_{VCLMF}$ wollen] die Sozis einen Antrag [$_{VCRVF}$ einbringen].

(7) Kowaljow hatte angekündigt, er [$_{VCLAF}$ werde] den Vorwürfen [$_{VCRVF}$ nachgehen].

(8) [$_{VCLVF}$ Stimmt] der Präsident auch zu, fehlt immer noch das Ja des Parlaments.

3.2.3 The Rechte Klammer (VCR_)

VCR_ is defined as being the right part of the clausal frame. While the VCL_ only occurs in main clauses and in non-introduced subclauses (i.e. verb-first and verb-second clauses), the VCR_ must occur in all introduced subclauses (i.e. verb-last clauses) and may occur in all kinds of other clauses provided that they contain a complex predicate (i.e. a predicate consisting of two verbs or a verb and a verbal particle). VCR_ may contain one or more tokens. In introduced subclauses, VCR_ contains all the verbal elements and the CF constitutes the left part of the clausal frame (See sentences 9 and 10); in main clauses VCR_ contains all the verbal elements except for the finite verb (which is contained in the VCL_) (See sentences 11 and 12). The structure of the label of the VCR_ has been explained in section 2.3.2.

(9) Ein Antrag, [$_{CF}$ dem/PRELS] weder die rot-grüne Koalition noch die PDS [$_{VCRMFVI}$ zustimmen mochten].

(10) Klar, [$_{CF}$ daß/KOUS] dieser Antrag keine Mehrheit [$_{VCRVF}$ fand].

(11) Für eine Feier auf öffentlichen Plätzen [$_{VCLAF}$ hätte] eine eindeutige Einladung [$_{VCRMIVI}$ vorliegen müssen].

(12) Etwaige Sicherheitsbedenken [$_{VCLVF}$ wies] er entschieden [$_{VCRPT}$ zurück].

3.2.4 The Vorfeld (VF)

The VF is defined as the topological field enclosed by the beginning of the sentence on the left-hand side and the VCL_ on the right-hand side. A VF may contain all kinds of constituents except the ones contained in the clausal frame (i.e. verbal elements and subordinators). An exception is the fronted and thus topicalized right part of the clausal frame which is labelled VCF_ and enclosed in

the VF (See sentence 13). As a VF may contain subclauses, a clausal frame as a part of such a subclause may be contained in the VF (See sentence 14). Typically, a VF just contains one constituent (which may, however, be very complex; See sentence 15). However, some adverbs (e.g. *freilich*; See sentence 16) may occur along other constituents.

(13) [$_{VF}$ [$_{VCFVI}$ Abnehmen]] [$_{VCLVF}$ kann] ihnen das keiner.

(14) [$_{VF}$ [$_{SUB}$ Daß ihr Vorhaben auf Widerstand stoßen würde]], [$_{VCLAF}$ war] den Transplanteuren in Hannover bewußt.

(15) [$_{VF}$ Die Lage der noch etwa 15.000 verbliebenen Einwohner Grosnys, die seit Wochen in den Kellern der belagerten Stadt ausharren], [$_{VCLAF}$ ist] katastrophal.

(16) [$_{VF}$ Nur unter dieser Bedingung freilich] [$_{VCLVF}$ wäre] man mit der Vereidigung auf öffentliche Plätze gegangen.

3.2.5 The Mittelfeld (MF)

The MF is defined as the topological field which is enclosed by the left part of the clausal frame (i.e. VCL_ or CF) and the right part of the clausal frame (i.e. VCR_). In cases in which no constituent appears between the left part of the clausal frame and the right part of the clausal frame, no MF is annotated (See sentence 17). If there is no right part of the clausal frame, the MF ends at the end of the sentence or at the beginning of a new main clause (See sentences 18 and 19), or at the beginning of the Nachfeld (NF) (See sentence 20). Sentence 20 also shows that an MF may begin after a comma in non-introduced non-finite clauses and that the MF of the matrix clause ends where this clause begins. In cases like this one, the annotation very much relies on punctuation.

(17) Mehrere weitere Menschen [$_{VCLAF}$ wurden] [$_{VCRVF}$ verletzt].

(18) Der Mann [$_{VCLVF}$ verletzte] [$_{MF}$ sich dabei zum Glück nur leicht] .

(19) Wir [$_{VCLVF}$ verkaufen] [$_{MF}$ ihnen keinen Reis], und dann kriegen wir keine Bananen.

(20) Lenin-Räuber [$_{VCLVF}$ versuchten] [$_{MF}$ vergeblich], [$_{NF}$ [$_{INF}$ [$_{MF}$ einen im Wald vergrabenen Lenin] [$_{VCRVZ}$ zu klauen]]].

3.2.6 The Nachfeld (NF)

The Nachfeld (NF) is defined as the topological field after the right part of the clausal frame. It may contain constituents of various categories. It may, however, not contain all kinds of syntactic functions. As this is of less imporatance in shallow annotation, it will not be discussed here. The most typical constituent of an NF is a subclause (See sentences 20 and 21); another typical but less frequent constituent is a phrase introduced by a Vergleichpartikel (See 22). Other constituents include prepositional phrases like the one in sentence 23 or even conjuncts like the one in 24. These cases are not typical cases of NF constituents but rather cases in which the author wanted to evoke some dramatic effect. This is even more the case with constituents which can be seen as a kind of addendum or afterthought (See 25).

(21) Plötzlich merkte ich, [$_{NF}$ was für ein ungeheurer Druck auf mir lastete].

(22) In Deutschland wurden 4,6 % mehr für Tabakwaren ausgegeben [$_{NF}$ als im Vorjahr].

(23) Die Dresdner Semperoper ist vollbesetzt [$_{NF}$ bis in den vierten Rang].

(24) Die Konfrontation solle in den Museen stattfinden – [$_{NF}$ oder auf der Straße].

(25) Er will beim Management umgerechnet 350 Mark rausholen, [$_{NF}$ das Doppelte des Monatslohns].

3.2.7 The Linksversetzung (LV)

The topological field LV is used to annotate resumptive constructions, in which a constituent is dislocated and moved to the left in front of the VF. This constituent is then resumed in its original place, which is the VF (See sentences 26 and 27). However, in the special case of the *nominativus pendens*, the refering pronoun may be situated in another place (See 28). Due to the limitations of a shallow annotation, these cases can, however, not be recognized. LVs are more likely to occur in spoken language. However, a construction like in sentence 26, in which a clause is fronted, is not infrequent in written language.

(26) [$_{LV}$ Wenn die Leute schon Skifahren müssen], [$_{VF}$ dann] sollen sie es tun, wenn genug Schnee da ist.

(27) [$_{LV}$ “Infos”], [$_{VF}$ das] sind vor allen Dingen lokale Nachrichten.

(28) [$_{LV}$ Ein frühes Tor], [$_{VF}$ jeder Trainer] würde sich wohl darüber freuen.

3.2.8 The Coordinator Field (KOORDF)

Höhle [H$\ddot{8}$6] states that the KOORDF is not a field in between sentences but a field introducing a sentence. He argues that sentences containing a KOORDF can be uttered without a preceding sentence which can be interpreted as its first conjunct. We share this view because it is supported by empirical data. Sentence 29, for example, has no first conjunct. Furthermore, there are examples like sentence 30, which very often must be interpreted as refering to more than one preceding sentence. However, there are also sentences like 31 and 32, in which there are doubtlessly two conjuncts.

(29) Welche Verleger sind mit welchen Konzepten in der Stadt ansässig? Was machen das Literaturkontor und die Literaturzeitschriften? [$_{KOORDF}$ Und] [$_{VF}$ auch die Bücher selbst] sollen gelobt oder verrissen werden.

(30) [$_{KOORDF}$ Doch] [$_{VF}$ daraus] wird nun nichts.

(31) Das sei ihm verziehen, [$_{KOORDF}$ denn] [$_{VF}$ um ihn] brennt die Luft beim Sendestart.

(32) Sie liegen auf der anderen Seite des Erdballs [$_{KOORDF}$ und] [$_{VF}$ ihr Streit] erscheint weit entfernt.

4 Clauses

There are three different kinds of clauses annotated by our system. Relative clauses (REL), non-finite clauses (INF) and general subclauses (SUB). A clause is defined as having one head verb (with the exception of coordinated head verbs). Main clauses are not annotated by the system because this would be beyond the scope of shallow annotation. This means that the relation between the subclauses is not made explicit. The implication of this has already been discussed in section 3.1. Generally, one can say that the attachment of the subclause is left open in shallow clause annotation just like the attachment of a prepositional chunk is left open in chunk annotation. This means for example that the reference noun of a relative clause is not given.

The category REL subsumes all relative clauses introduced by relative pronouns. Clauses introduced by adverbial relative pronouns (i.e. PWAV) are annotated as SUB because, from the perspective of a shallow annotation, it is not clear

in which cases a PWAV is in fact a relative pronoun. The category INF subsumes all non-finite clauses containing an infinitive (not the ones containing a participle) both introduced by a complementizer and non-introduced. The category SUB subsumes all other introduced subclauses, which are mainly adverbial clauses. Recursiveness is dealt with by applying the annotation cascade for subclauses twice. As the annotation of clauses is a shallow one, this does not affect subclauses in succession (because they are not attached at all), but just cases in which embedding occurs in the MF. Thus, in our system, a subclause cannot contain a subclause in its MF which again contains another subclause in its MF. These cases are rare, however, because centre-embedding is limited due to cognitive limitations.[5]

[5] A counterexample is a sentence like: *Wenn Du Deinen Freund, der, wenn er den Mund aufmacht, lügt, mitbringst, geh ich.*

References

[Bra98] Thorsten Brants. *TnT – A Statistical Part-of-Speech Tagger*. Universität des Saarlandes, Computational Linguistics, Saarbrücken, 1998.

[H̊86] Tilman Höhle. Der Begriff 'Mittelfeld', Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses*, pages 329–340, Göttingen, 1986.

[MU01] Frank Henrik Müller and Tylman Ule. Satzklammer annotieren und Tags korrigieren. Ein mehrstufiges Top-Down-Bottom-Up-System zur flachen, robusten Annotierung von Sätzen im Deutschen. In Henning Lobin, editor, *Proceedings der GLDV-Frühjahrstagung 2001*, pages 225–234, Gießen, 2001. Gesellschaft für Linguistische Datenverarbeitung.

[STTS95] Anne Schiller, Simone Teufel, Christine Thielen, and Christine Stöckert. *Guidelines für das Taggen deutscher Textcorpora mit STTS*. IMS Stuttgart und SfS Tübingen, Stuttgart und Tübingen, 1995.