Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)

Frank Henrik Müller

fhm@sfs.uni-tuebingen.de

January 15, 2004

Abstract

The stylebook at hand is a guide to the shallow annotation structure of the **Tü**bingen **P**artially **P**arsed Corpus of Written German (TüPP-D/Z). The annotation is performed automatically by the partial parser KaRoPars (**Ka**skadierter **Ro**buster **Pars**er, i.e. Cascaded Robust Parser). This stylebook focuses on the description of the layers of chunks, topological fields and clauses, which are part of the KaRoPars output. The methodology of the annotation process is only mentioned in those cases in which it has an impact on the annotation structure. Example sentences are taken from real language data, but are simplified where necessary. This stylebook is an updated version of Müller (2002). The actual encoding of the linguistic phenomena in the corpus is described in the *Markup Manual for TüPP-D/Z* (see Ule (2004)).

Contents

1 Introduction

The KaRoPars system (see Ule and Müller (2004)), which was used for the automatic annotation of the TüPP-D/Z, first assigns POS labels to words according to the scheme of the STTS-tagset (see Schiller et al. (1995)). On top of the POS-annotated text, KaRoPars annotates topological fields, clauses and chunks. After this shallow annotation, the text is available for global syntactic analysis (see Müller and Ule (2002) and Müller and Ule (2003)). The decision to split the annotation task has been made on the basis of a decision to first annotate structures which can be handled with transducers using just POS tag information and syntactic restrictions. This structure is what we define as the shallow annotation structure. More powerful formalisms also using lexical subcategorization information may be used afterwards for the annotation of deeper linguistic structures. Please consult the TüPP-D/Z homepage for up-to-date information:

`http://www.sfs.uni-tuebingen.de/tupp`

2 The Chunk Layer

2.1 The Notion of *Chunks*

Typically, chunks are defined as non-recursive continuous kernels of phrases. This means that chunks may contain chunks of other categories but that they may not contain chunks of the same category. This concept works well with English. There are, however, problems in German, because, in German, there are center-embedding phenomena like the one in figure 1. In center-embedded structures, a constituent is embedded into a constituent of the same type with elements of the embedding constituent surrounding the embedded constituent. In figure 1, the prepositional chunk (PC) ‘bei der Bahn’ contains the attributive adjective chunk (AJAC) ‘durch jahrelange Fehlentscheidungen hochverschuldeten’. Since, in the AJAC, the adjective ‘hochverschuldeten’ is modified by the PC ‘durch jahrelange Fehlentscheidungen’, the AJAC and, thus, also the dominating PC contain a PC, i.e. a constituent of the same type.

If we did not allow this type of recursion in chunks, center-embedded structures would lead to problems for two reasons: In the first place, chunks are supposed to represent meaningful linguistic units and, secondly, the chunk structure is supposed to be a structure on which further and deeper annotation can be built. If, however, center-embedding structures like the one in figure 1 are not annotated,

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

Figure 1: Center-embedded structure as annotated

then structures like the one in figure 2 are generated by the shallow parsing component. It is obvious that the structure in figure 2 does not represent meaningful linguistic units since the determiner ‘der’ belonging to the head noun ‘Bahn’ is separated from it. The same is true of the preposition ‘bei’ which is also separated from the related noun ‘Bahn’. In addition, what is equally important, is that from this structure it is not possible to straightforwardly generate the correct structure, which is the one in figure 1. This is different from, e.g., a noun chunk (NC) which is followed by a prepositional chunk. If, in a further annotation step, it turns out that the PC is a modifier, then the NC followed by the PC can be grouped together as an NP. Such as simple grouping is not possible if one wants to generate the structure in figure 1 from the one in figure 2. Thus, center-embedded structures are annotated by the shallow parsing component.

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

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

[NC
    [AJAC
        .ADJA hochverschuldeten ]
    .NN Bahn ] ]
```

Figure 2: Center-embedded structure if recursion is excluded

2.2 Outline of Chunk Types

There are five major types of chunks. Verb chunks (VC_)[1], noun chunks (NC), adjective chunks (AJ_C), adverb chunks (AVC) and prepositional chunks (PC). Verb chunks play a special role in the chunk structure, as they are both chunks and part of the sentence bracket. 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

[1]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 3: Truncated noun chunk (NCTRUNC)

its category (see figure 3). Truncated chunks are, thus, not a chunk category on its own, but rather belong – according to their head words – as a sub-category to the respective chunk category. Truncated words, are, thus, treated just like their non-fragmentary counterparts. In cases in which 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 sentence bracket and VC$\boxed{\text{R}}$ chunks are chunks which are the **right** part of the sentence bracket. VC$\boxed{\text{F}}$ chunks are chunks which in the basic word order would be the right part of the sentence bracket but which are **front**ed and, thus, topicalized. VC$\boxed{\text{E}}$ chunks occur in **Ersatz**infinitiv constructions in subordinate clauses with verb-last position. They contain the finite verb, which – without the Ersatzinfinitiv occuring – would be part of the right part of the sentence bracket but which is moved to the left of it in such a construction (see figure 5). 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)[2]. The se-

[2] An exception is being made in the treatment of the infinitive with 'zu', where 'I' is replaced

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

Figure 4: Verb chunk in relative clause

```
[NC
     .PRELS die ]
[NC
     .PPER er ]
[VCEAF
     .VAFIN hätte ]
[VCRMIVI
     .VVINF sehen
     .VMINF sollen ]
```

Figure 5: VCE_ in Ersatzinfinitivkonstruktion in relative clause

quence 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 sentence bracket, $\boxed{\text{AF}}$ for auxiliary finite, $\boxed{\text{VZ}}$ for lexical verb/infinitive with 'zu' (see figure 4). As they are parts of verbs, verbal particles are also included in the class of the verb chunks with the chunk type name VCRPT.

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.*'.

2.3.3 Noun Chunks

Noun chunks are the most common chunks. They consist at least of a noun (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 6, 12 and 13 and the examples mentioned in section 2.3.4). As the definition of chunks does not allow recursive structures except in cases of center-embedding, post-modifying prepositional or nominal constituents may not be part of a noun chunk (see figure 6). In case of preposition-article contraction (APPRART), the contraction is annotated in its function as a preposition (see figure 6).

As pronouns are normally not modified, they are the only element of a noun chunk when they occur (see figure 18). Adverb chunks are only included in a noun chunk when they occur after clear syntactic 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 (see figures 8 and 13). 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 8). If there is no clear syntactic indication that the adverb belongs to the NC, it is not included.

with 'Z' in the chunk type name (see figure 4) and with the imperative where a 'B' is assigned.

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

Figure 6: Noun chunks

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

Figure 7: Cardinal as the head of a noun chunk

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

Figure 8: Modifying adverb chunk in attributive adjective chunk

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

Figure 9: Two coordinated NCs

```
[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 10: Sentence containing elliptical NC

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

Figure 11: Elliptical NC coordinated with NC

If two noun chunks are coordinated by a coordinator, they are chunked as a coordinated noun chunk NCC. If 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 the chunk level whether the adjective chunk modifies both nouns or the first noun only (see figure 9). 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 11).

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 (attributive) and ADJD (predicative/adverbial)).

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, the AJAC chunk is always part of a noun chunk. AJAC chunks may contain

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

Figure 12: AJAC chunk with modifying ADJD attached outside chunk

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

Figure 13: Cardinal number as head of an attributive adjective chunk

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

Figure 14: AJAC chunks coordinated by comma

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

Figure 15: AJAC chunks coordinated by *und*

modifying PCs as explained in section 2.1. A modifying adverb (ADV) or ADJD may be included in the AJAC if there is clear indication of its containment in the noun chunk (see section 2.3.3). If, however, there is no such clear indication, the ADV or ADJD is left outside the NC and, consequently, the AJAC (see figure 12).

AJAC chunks may be coordinated in two ways: with commas or with a coordinator (see figure 14 and figure 15). In this case 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 16). 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 (see figure 10) or follow (see figure 11) the elliptical noun chunk.

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

Figure 16: Two AJAC chunks

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

Figure 17: AJVC as a predicative adjective chunk

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

Figure 18: One AJVC as an adverbial and the other as a predicative adjective

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 17) or as an adverb (see figure 18). Obviously, in the tagset, no distinction was made because it is impossible to draw it without making a full parse. As this is still true as regards 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. The label AJVC has, thus, to be read as 'either predicative adjective **or** adverb chunk'.

Since it cannot be detected by syntactic restrictions whether an ADV or an ADJD is a modifier of an ADJD, modifiers of an ADJD are not chunked together with the AJVC (see figure 18). In the case of coordination of AJVC chunks, they are projected to an AJVC$\boxed{\text{C}}$ chunk (see figure 19). Coordination may occur with or without coordinator.

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

Figure 19: Coordinated AJVC chunks

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

Figure 20: Coordinated AVC chunks

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 (AVC) is then not part of the modified chunk. In most cases, adverb chunks consist of a single adverb only. The particle *nicht*, which is tagged PTKNEG, is counted among the adverbs. Adverb chunks cannot contain any other constituents than adverbs. Coordinated adverb chunks are grouped into a chunk labelled AVC$\boxed{\text{C}}$ analogous to the adjective chunks and the noun chunks (see figure 20).

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 (post-position) (see figure 21) or includes it (circumposition) (see figure 22). 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 23). In some cases, a prepositional chunk may contain what might be called a complex

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

Figure 21: PC with post-position

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

Figure 22: PC with circumposition

preposition, e.g. the token *bis* followed by a preposition (see figure 24).

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

Figure 23: Prepositional chunk with adverb head

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

Figure 24: 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	at least two coordinated NCs
NCell	elliptical noun chunk (i.e. without head noun)
PC	prepositional chunk
VCTRUNC	verb chunk with truncated verb
VCL_	verb chunk as left part of sentence bracket
VCR_	verb chunk as right part of sentence bracket
VCF_	verb chunk in topicalized (fronted) position
VCE_	verb chunk containing finite verb in *Ersatzinfinitiv* structure

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 (and the subordinator in subclauses). There are three different types of clauses (see table 2): verb-last clauses (VL), verb-first clauses (V1) and verb-second clauses (V2). VL clauses comprise all introduced subclauses, V1 clauses mainly comprise imperatives and yes/no questions and V2 clauses mostly comprise affirmative clauses. The topological fields CF/VCL and VCR constitute the sentence bracket, relative to which the other fields can be described.[3] The section preceding the left part of the sentence bracket is called the *Vorfeld* (VF; initial field; only in V2 clauses), the section included in the sentence bracket is called the *Mittelfeld* (MF; middle field) and the section following the right part of the sentence bracket is called the *Nachfeld* (NF; final field). There may also be a Koordinationsfeld (KOORDF; coordinator field), which contains a coordinating particle, or an alternative field to the KOORDF, the field PARORDF, which only occurs in V2 clauses. Additionally, there may be a field for resumptive constructions (LV; Linksversetzung). While the ordering of phrases is relatively free in German, the ordering of topological fields is subject to syntactic restrictions which adhere to the unvarying pattern outlined in table 2.

Table 2: The topological field model

clause type	topological fields						
VL:	KOORDF	LV		**CF**	MF	**VCR**	NF
V1:	KOORDF	LV		**VCL**	MF	VCR	NF
V2:	KOORDF/ PARORDF	LV	**VF**	**VCL**	MF	VCR	NF

The model of topological fields describes the distribution of constituents relative to the sentence bracket. It is therefore primarily a distributional model. It does not give any account of the verb-grammatical function structure and it does not reveal the relations between the constituents within the topological fields, either. In fact, the very structure of constituents **within** topological fields is left open

[3]Obligatory fields are in bold type.

in this theory. The model still has some clear advantages from both a theoretical and an annotational perspective: As regards the theoretical perspective it is, for example, important to point out that a lot of constituent-order phenomena can be described relative to topological fields. As regards the annotation of further grammatical information, the constituent order in the different topological fields may be utilized for annotation. This is possible because, although there are, in German, 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, clause boundaries and potential points of attachment are known. The annotation of topological fields considerably reduces the scope of ambiguity because the verb is always part of the sentence bracket and the respective grammatical functions are always realized in the corresponding fields. Without the annotation of topological fields the scope of the grammatical functions of the verbs is much wider, especially in complex sentences, in which, additionally, it is not clear which potential functions belong to which verb. After the annotation of topological fields, the syntactic restrictions and preferences which are valid in them might be utilized for further annotation (together with other linguistic information).

The advantages of annotating topological fields before annotating verb-grammatical function structure can be illustrated by figure 25, which shows a sentence containing five verbs. As the grammatical functions may be realized on both sides of the verb, it is by no means clear, where the respective grammatical functions of the verb are. After the annotation of the topological fields, however, the scope is reduced: First of all, the annotation structure in figure 25 shows that 'wurde ...gezeigt' is one verb complex which is a passive. The topological field and clause structure reveal that the only phrasal grammatical functions of the verb 'zeigen' can be 'die Rede', 'des Forschers' or 'auf einem Bildschirm'. Since the VF typically just contains one element, it is clear that the subordinate clause (SUB) in the VF is part of the infinitive clause (INF). Since relative clauses cannot act as grammatical functions alone, the only remaining potential clausal grammatical function is the INF. Overall, the number of potential grammatical functions for 'zeigen' has been reduced to four constituents.

For the three subclauses, the scope of their verbs has been reduced to their MFs and to clauses following within the same field. The latter only applies to the INF. Figure 25 illustrates how the scope of potential grammatical functions 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 25 it is left open to which chunk the REL is attached. Still, the pre-structuring achieved with the annotation of the fields is a solid base for further annotation.

3.2 Outline of Field Types

A distinction can be made between two types of fields. On the one hand, there are those fields which are part of the sentence bracket. They can typically only contain tokens of a restricted number of Parts-of-Speech, namely complementizers and verbs. The corresponding fields are the complementizer field (CF) and the left part of the clausal frame (VCL_) and the verb complex as the right part of the clausal frame (VCR_). VCL_ and VCR_ are chunks and topological fields at one and the same time. On the other hand, there are those fields which can be described relative to the sentence bracket. 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 LV. A special case is the KOORDF and the PARORDF. These fields just contain one constituent (the coordinator). They are fields which may occur at the beginning of a clause.

3.3 The Types of Topological Fields

3.3.1 The Complementizer Field (CF)

The CF only occurs in subordinated clauses introduced by a complementizer. It is always the left part of the sentence bracket. 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

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

```
{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 25: Complex field and clause structure

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.

3.3.2 The Left Part of the Sentence Bracket (VCL_)

While the CF is the left part of the sentence bracket in introduced subclauses, the VCL_ is the left part of the sentence bracket 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.

(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.

3.3.3 The Right Part of the Sentence Bracket (VCR_)

VCR_ is defined as being the right part of the sentence bracket. While VCL_ only occurs in main clauses and in non-introduced subclauses (i.e. verb-first and verb-second clauses), 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 sentence bracket (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).

(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].

3.3.4 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 sentence bracket (i.e. verbal elements and subordinators). An exception is the fronted and thus topicalized right part of the sentence bracket 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) [$_{VF}$ [$_{PC}$ Unter dieser Bedingung] [$_{AVC}$ freilich]] [$_{VCLVF}$ wäre] man mit der Vereidigung auf öffentliche Plätze gegangen.

3.3.5 The Mittelfeld (MF)

The MF is defined as the topological field which is enclosed by the left part of the sentence bracket (i.e. VCL_ or CF) and the right part of the sentence bracket (i.e. VCR_). In cases in which no constituent appears between the left part of the sentence bracket and the right part of the sentence bracket, no MF is annotated (see sentence 17). If there is no right part of the sentence bracket, 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, automatic 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]]].

3.3.6 The Nachfeld (NF)

The Nachfeld (NF) is defined as the topological field after the right part of the sentence bracket. It may contain constituents of various categories. It may, however, not contain all kinds of grammatical functions. As this is of less importance 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 Vergleichspartikel (see sentence 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 sentence 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].

3.3.7 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 referring pronoun may be situated in another place (see sentence 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 as shown in sentence 26, in which it is a clause which 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.

3.3.8 The Coordination Fields (KOORDF and PARORDF)

Höhle Höhle (1986) 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 referring to more than one preceding sentence. However, there are also sentences like 31 and 32, in which there are doubtlessly two conjuncts. Sentence 32 shows the alternative field to the KOORDF, the PARORDF.

The difference between PARORDF and KOORDF is that, if two clauses occur, the conjunction in KOORDF coordinates the two clauses both syntactically and semantically while the conjunction in PARORDF only coordinates the clauses syntactically. Semantically, there exists a relation of subordination because the second clause in sentence 32 gives the reason for the proposition made in the first clause. This can be tested by reverting the order of the two conjuncts. With KOORDF which contains conjunctions like 'und' or 'oder' the meaning of the sentence does not change. With PARORDF, however, which contains conjunctions like 'denn' or 'weil' (this subsume just the 'weil' which is used in V2 clauses) the meaning changes.

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

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

3.3.9 Coordination of Topological Fields

A special case is the coordination of topological fields: In this case, one or more topological fields are a constituent in a coordination. For the TüPP-D/Z, only the coordination of MFs and VCR_ is annotated (with the label MFVCC) since this structure can be annotated quite reliably and would cause problems in further annotation if it was not annotated. This structure mainly occurs in VL clauses (cf. figure 26). The attachment is shallow: All the fields and the coordinator are directly attached to the field coordination and not grouped as a conjunct before.

4 Clauses

Clauses are annotated according to the scheme of V1, V2 and VL clauses. There is no further subdivision in the categories of V1 and V2 clauses since they are

```
(V2
    {VF
        [NC
            .PWS     Was ] }
    [VCLMF
        .VMFIN   soll ]
    {MF
        [NC
            .PPER    ich ]
        [AVC
            .ADV     denn ]
        [PC
            .APPR    mit
            [NC
                .ART     einem
                [AJAC
                    .ADJA    freien ]
                [NC
                    .NN      Dienstag ] ] ] }
    .$,      ,
    {NF
        (REL
            {CF
                [PC
                    .APPR    an
                    [NC
                        .PRELS   dem ] ] }
            <MFVCC
                {MF
                    [NC
                        .PPOSAT  meine
                        .NN      Frau ] }
                [VCRVF
                    .VVFIN   arbeitet ]
                .KON     und
                {MF
                    [NC
                        .ART     die
                        .NN      Kinder ]
                    [PC
                        .APPR    in
                        [NC
                            .ART     der
                            .NN      Schule ] ] }
                [VCRAF
                    .VAFIN   sind ] > ) }
    .$.      . )
```


Figure 26: Coordination of topological fields

```
(V1
    [VCLMF
        .VMFIN   Kann ]
    {MF
        [NC
            .ART     die
            .NN      Talkshow ]
        [PC
            .APPRART im
            [NC
                .NN      Privatleben ] ]
        [NC
            [AJAC
                .ADJA    konstruktive ]
            .NN      Gespräche ] }
    [VCRVI
        .VVINF   initiieren ]
    .$.      ? )
```

Figure 27: V1 clause

are typically maximal clauses, i.e. they do not have any function within a super-ordinate clause (cf. figure 27 and 28). In cases in which they do have a function in super-ordinate clauses, this function has to be resolved on later annotation level because the shallow annotation level works with syntactic restrictions alone. With these, it is not possible to determine the function of a V1 or V2 clause. VL clauses, which are typically subclauses, are subdivided into three different categories: general subclauses (SUB, cf. figure 28), infinitive clauses (INF, cf. figure 28) and relative clauses (REL, cf. figure 29). 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). It includes clauses introduced by a complementizer and and those which are non-introduced. The category SUB subsumes all other introduced subclauses, which are mainly adverbial clauses.

The annotation of the clauses is shallow in the sense that the attachment of subordinate clauses is left unresolved just like the attachment of prepositional chunks is left unresolved in the chunk annotation. This means, for example, that the reference noun of a relative clause is not given. VL clauses are included in

```
(V2
    {VF
        (SUB
            {CF
                [NC
                    .PWS     Wer ] }
            {MF
                [AVC
                    .ADV     aber ]
                [PC
                    .APPR    von
                    [NC
                        .NN      Propaganda ] ] }
            [VCRAFVP
                .VVPP    verdorben
                .VAFIN   ist ] )
        .$,       , }
    [VCLAF
        .VAFIN   hat ]
    {MF
        [AVC
            .ADV     wohl ]
        [NC
            .NN      Grund ] }
    .$,       ,
    {NF
        (INF
            {MF
                [NC
                    .PRF     sich ] }
            [VCRVZ
                .PTKZU   zu
                .VVINF   fürchten ] ) }
    .$.       . )
```

Figure 28: V2 clause with VL clause SUB embedded into its VF and VL clause INF embedded into its NF

the super-ordinate clause if this is a V1 or V2 clause and they are included in the VL clause if they are center-embedded into it. Thus, the treatment of recursion in clauses is analogous to the treatment of recursion in chunks. Clauses cannot contain clauses of the same type unless they are center-embedded. The reason for this treatment is the same as with the chunks. If recursion in center-embedded clauses is not treated, structures are generated which are not adequate shallow annotation structures.

Acknowledgements

The annotation of TüPP-D/Z has taken great advantage of resources and tools originally set up in the DEREKO project (`http://www.sfs.uni-tuebingen.de/dereko`). The tools have been updated with support from the DFG project *Sonderforschungsbereich 441: Linguistische Datenstrukturen*.

TüPP-D/Z has been annotated using KaRoPars, which integrates a number of tools into a cascaded annotation system (Ule and Müller, 2004). We are grateful to the authors of these tools, which include

- `fsgmatch` – a general-purpose transducer operating on XML (Mikheev et al., 1999)
- `tnt` – a part-of-speech tagger (Brants, 2000)
- `xmlperl` – an XML processing/translating language (McKelvie, 1999)
- DMOR – Deutsche Morphologie (Schiller, 1995)

```
(V2
    {VF
        [NC
            .PPER    Es ] }
    [VCLVF
        .VVFIN   handelt ]
    {MF
        [NC
            .PRF     sich ]
        [AVC
            .ADV     eben ]
        [PC
            .APPR    um
            [NC
                .ART     einen
                [AJAC
                    .ADJA    doofen ]
                .NN      Trick ] ] }
    .$,      ,
    {NF
        (REL
            {CF
                [NC
                    .PRELS   der ] }
            {MF
                [AVC
                    .ADV     dennoch ] }
            [VCRVF
                .VVFIN   verfängt ] ) }
    .$.      . )
```

Figure 29: VL clause REL embedded into NF of V2 clause

References

Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of ANLP-2000, April*, Seattle, WA, 2000.

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

David McKelvie. *XMLPERL 1.0.4*. Language Technology Group, University of Edinburgh, Edinburgh, 1999. URL `http://www.cogsci.ed.ac.uk/~dmck/xmlperl`.

Andrei Mikheev, Claire Grover, and Marc Moens. XML tools and architecture for named entity recognition. *Markup Languages*, 1(3):89–113, 1999.

Frank Henrik Müller. Shallow-Parsing Stylebook for German. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, 2002. URL `http://www.sfs.uni-tuebingen.de/dereko/anno-doc.html`.

Frank Henrik Müller and Tylman Ule. Annotating topological fields and chunks – and revising POS tags at the same time. In Shu-Chuan Tseng, editor, *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, volume 2, pages 695–701, Taipei, Taiwan, 2002. Morgan Kaufmann.

Frank Henrik Müller and Tylman Ule. On the nature, annotation and use of shallow parsing structures. In Lea Cyrus, Hendrik Feddes, Frank Schumacher, and Petra Steiner, editors, *Sprache zwischen Theorie und Technologie. Festschrift für Wolf Paprotté zum 60. Geburtstag*, Sprachwissenschaft, pages 199–209. Deutscher Universitäts-Verlag, Wiesbaden, 2003.

Anne Schiller. DMOR Benutzerhandbuch. Technical report, IMS, Universität Stuttgart, 1995.

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. URL `http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz`.

Tylman Ule. *Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, January 15 2004. URL `http://www.sfs.uni-tuebingen.de/tupp/dz/markupmanual.ps`.

Tylman Ule and Frank Henrik Müller. KaRoPars: Ein System zur linguistischen Annotation großer Text-Korpora des Deutschen. In Alexander Mehler and Henning Lobin, editors, *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, Opladen, 2004. Westdeutscher Verlag.