Chunking German: An Unsolved Problem

Sandra Kübler
Indiana University
Bloomington, IN, USA
skuebler@indiana.edu

Kathrin Beck, Erhard Hinrichs, Heike Telljohann
Universität Tübingen
Tübingen, Germany
{kbeck,eh,telljohann}@sfs.uni-tuebingen.de

Abstract

This paper describes a CoNLL-style chunk representation for the Tübingen Treebank of Written German, which assumes a flat chunk structure so that each word belongs to at most one chunk. For German, such a chunk definition causes problems in cases of complex prenominal modification. We introduce a flat annotation that can handle these structures via a stranded noun chunk.

1 Introduction

The purpose of this paper is to investigate how the annotation of noun phrases in the Tübingen Treebank of Written German (TüBa-D/Z) can be transformed into chunks with no internal structure, as proposed in the CoNLL 2000 shared task (Tjong Kim Sang and Buchholz, 2000). Chunk parsing is a form of partial parsing, in which non-recursive phrases are annotated while difficult decisions, such as prepositional phrase attachment, are left unsolved. Flat chunk representations are particularly suitable for machine learning approaches to partial parsing and are inspired by the IOB approach to NP chunking first proposed by Ramshaw and Marcus (1995). They are particularly relevant for approaches that require an efficient analysis but not necessarily a complete syntactic analysis.

German allows a higher degree of syntactic complexity in prenominal modification of the syntactic head of an NP compared to English. This is particularly evident in written texts annotated in the TüBa-D/Z. The complexity of German NPs that causes problems in the conversion to CoNLL-style chunks also affects PCFG parsing approaches to German.The complexity of NPs is one of the phenomena that have been addressed in tree transformation approaches for German parsing (Trushkina, 2004; Ule, 2007; Versley and Rehbein, 2009).

2 Defining Chunks

The notion of a chunk is orginally due to Abney (1991), who considers chunks as non-recursive phrases which span from the left periphery of a phrase to the phrasal head. Accordingly, the sentence "The woman in the lab coat thought you had bought an expensive book." is assigned the chunk structure: "[S [NP The woman] [PP in [NP the lab coat]] [VP thought]] [S [NP you] [VP had bought] [NP an [ADJP expensive] book]] .". Abney-style chunk parsing is implemented as cascaded, finite-state transduction (cf. (Abney, 1996; Karlsson et al., 1995)).

Notice that cascaded, finite-state transduction allows for the possibility of chunks containing other chunks as in the above sentence, where the prepositional chunk contains a noun chunk within. The only constraint on such nested chunks is the prohibition on recursive structures. This rules out chunks in which, for example, a noun chunk contains another noun chunk. A much stricter constraint on the internal structure of chunks was subsequently adopted by the shared task on chunk parsing as part of the Conference for Natural Language Learning (CoNLL) in the year 2000 (Tjong Kim Sang and Buchholz, 2000). In this shared task, chunks were defined as non-overlapping, non-recursive phrases so that each word is part of at most one chunk. Based on this definition, the prepositional phrase in the sentence above would be chunked as "[Prep in] [NP the lab coat]". Since the prepositional chunk cannot have an embedded noun chunk, the definition of the CoNLL shared task assumed that the prepositional chunk only contains the preposition, thus taking the definition seriously that the chunk ends with the head. The noun chunk remains separate. Additionally, the noun phrase "an expensive book" is annotated as a noun chunk without internal structure.

The CoNLL shared task definition of chunks is

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 147–151,
Uppsala, Sweden, 15-16 July 2010.

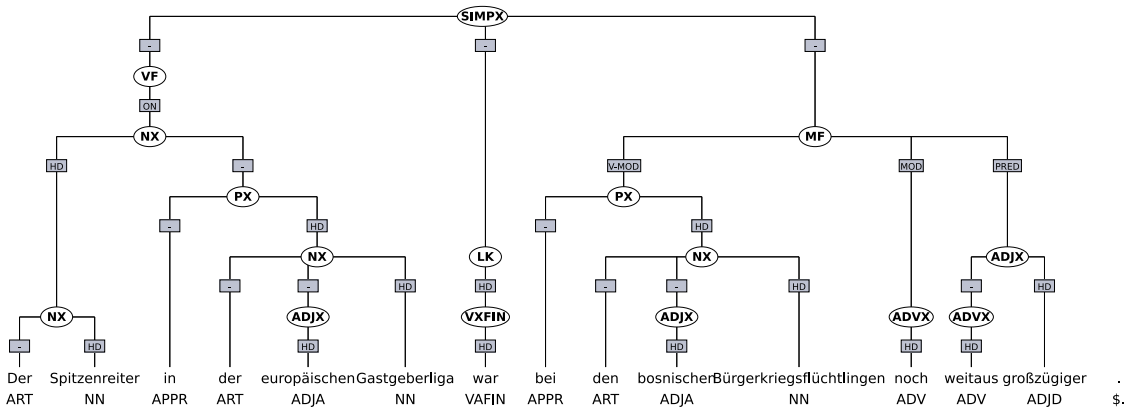

Figure 1: Treebank annotation for the sentence in (2).

useful for machine learning based approaches to chunking since it only requires one level of analysis, which can be represented as IOB-chunking (Tjong Kim Sang and Buchholz, 2000). For English, this definition of chunks has become standard in the literature on machine learning.

For German, chunk parsing has been investigated by Kermes and Evert (2002) and by Müller (2004). Both approaches used an Abney-style chunk definition. However, there is no corresponding flat chunk representation for German because of the complexity of pre-head modification in German noun phrases. Sentence (1) provides a typical example of this kind.

(1) $[_{NC}$ der $[_{NC}$ seinen Sohn] liebende Vater]
the his son loving father
'the father who loves his son'

The structure in (1) violates both the Abney-style and the CoNLL-style definitions of chunks – Abney's because it is recursive and the CoNLL-style definition because of the embedding. A single-level, CoNLL-style chunk analysis will have to cope with the separation of the determiner "der" and the head of the outer phrase. We will discuss an analysis in section 5.

3 The Treebank: TüBa-D/Z

The Tübingen Treebank of Written German (TüBa-D/Z) is a linguistically annotated corpus based on data of the German newspaper 'die tageszeitung' (taz). Currently, it comprises approximately 45 000 sentences. For the syntactic annotation, a theory-neutral and surface-oriented annotation scheme has been adopted that is inspired by the notion of topological fields and enriched by a level of predicate-argument structure. The annotation scheme comprises four levels of syntactic annotation: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields, which characterize the word order regularities among different clause types of German, and which are widely accepted among descriptive linguists of German (cf. (Drach, 1937; Höhle, 1986)). Below this level of annotation, i.e. strictly within the bounds of topological fields, a phrase level of predicate-argument structure is applied with its own descriptive inventory based on a minimal set of assumptions that has to be captured by any syntactic theory. The context-free backbone of phrase structure (Telljohann et al., 2004) is combined with edge labels specifying the grammatical functions and long-distance relations. For more details on the annotation scheme see Telljohann et al. (2009).

(2) Der Spitzenreiter in der europäischen Gastgeberliga war bei den bosnischen Bürgerkriegsflüchtlingen noch weitaus großzügiger.
'The front-runner in the European league of host countries was far more generous with the Bosnian civil war refugees.'

Figure 1 shows the tree for the sentence in (2). The main clause (SIMPX) is divided into three topological fields: initial field (VF), left sentence bracket (LK), and middle field (MF). The finite

verb in LK is the head (HD) of the sentence. The edge labels between the level of topological fields and the phrasal level constitute the grammatical function of the respective phrase: subject (ON), ambiguous modifier (MOD), and predicate (PRED). The label V-MOD specifies the long-distance dependency of the prepositional phrase on the main verb. Below the lexical level, the parts of speech are annotated. The hierarchical annotation of constituent structure and head (HD) / non-head (-) labels capture phrase internal dependencies. While premodifiers are attached directly on the same level, postmodifiers are attached higher in order to keep their modification scope ambiguous. The PP "in der europäischen Gastgeberliga" is the postmodifier of the head-NX and therefore attached on a higher phrase level.

4 General Conversion Strategy

The conversion to CoNLL-style chunks starts from the syntactic annotation of the TüBa-D/Z. In general, we directly convert the lowest phrasal projections with lexical content to chunks. For the sentence in (2) above, the chunk annotation is shown in (3). Here, the first noun phrase[1], "Der Spitzenreiter", as well as the finite verb phrase and the adverbial phrase are used as chunks.

(3) [NX Der Spitzenreiter] [PX in der europäischen Gastgeberliga] [VXFIN war] [PX bei den bosnischen Bürgerkriegsflüchtlingen] [ADVX noch] [ADJX weitaus großzügiger].

This sentence also shows exceptions to the general conversion rule: We follow Tjong Kim Sang and Buchholz (2000) in including ADJPs into the NCs, such as in "den bosnischen Bürgerkriegsflüchtlingen". We also include premodifying adverbs into ADJCs, such as in "weitaus großzügiger". But we deviate from Tjong Kim Sang and Buchholz in our definition of the PCs and include the head NP into this chunk, such as in "in der europäischen Gastgeberliga".

(4) a. Allerdings werden wohl Rationalisierungen mit der Modernisierung der Behördenarbeit einhergehen.

'However, rationalizations will accompany modernization in the workflow of civil service agencies.'

b. [ADVX Allerdings] [VXFIN werden] [ADVX wohl] [NX Rationalisierungen] [PX mit der Modernisierung] [NX der Behördenarbeit] [VXINF einhergehen].

In cases of complex, post-modified noun phrases grouped under the prepositional phrase, we include the head noun phrase into the prepositional chunk but group the postmodifying phrase into a separate phrase. The sentence in (4a) gives an example for such a complex noun phrase. This sentence is assigned the chunk annotation in (4b). Here, the head NP "der Modernisierung" is grouped in the PC while the post-modifying NP "der Behördenarbeit" constitutes its own NC.

The only lexical constituent in the treebank that is exempt from becoming a chunk is the *named entity* constituent (EN-ADD). Since these constituents do not play a syntactic role in the tree, they are elided in the conversion to chunks.

5 Complications in German

While the conversion based on the phrasal annotation of TüBa-D/Z results in the expected chunk structures, it is incapable of handling a small number of cases correctly. Most of these cases involve complex NPs. We will concentrate here on one case: complex premodified NPs that include the complement of a participle or an adjective, as discussed in section 2. This is a non-trivial problem since the treebank contains 1 497 cases in which an ADJP within an NP contains a PP and 415 cases, in which an ADJP within an NP contains another NP. Sentence (5a) with the syntactic annotation in Figure 2 gives an example for such an embedded PP.

(5) a. Die teilweise in die Erde gebaute Sporthalle wird wegen ihrer futuristischen Architektur auch als "Sport-Ei" bezeichnet.

'The partially underground sports complex is also called the "sports egg" because of its futuristic architecture.'

b. [sNX Die] [ADVX teilweise] [PX in die Erde] [NX gebaute Sporthalle] [VXFIN wird] [PX wegen ihrer futuristischen Architektur] [ADVX auch]

[1] For the sake of convenience, we will use acronyms in the remainder of the paper. Since we use the same labels in the treebank annotation and in the chunk representation (mostly ending in X), we will use labels ending in P (e.g. NP, PP) to talk about phrases in the treebank and labels ending in C (e.g. NC, PC) to talk about chunks.

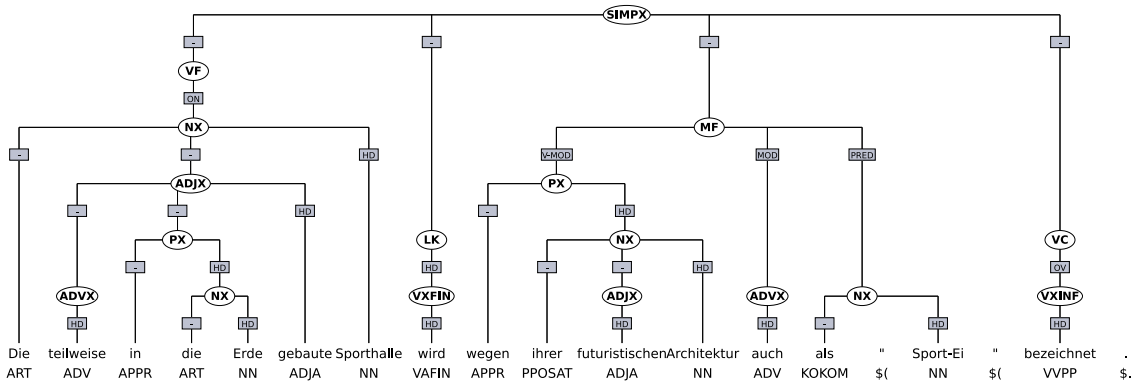

Figure 2: Treebank annotation for the sentence in (5a).

[NX als “ Sport-Ei] ” [VXINF bezeichnet].

Since we are interested in a flat chunk annotation in which each word belongs to at most one chunk, the Abney-style embedded chunk definition shown in sentence (1) is impossible. If we decide to annotate the PP “in die Erde” as a chunk, we are left with two parts of the embedding NP: the determiner “Die” and the ADVP “teilweise” to the left of the PP and the ADJP “gebaute” and the noun on the right. The right part of the NP can be easily grouped into an NC, and the ADVP can stand on its own. The only remaining problem is the treatment of the determiner, which in German, cannot constitute a phrase on its own. We decided to create a new type of chunk, stranded NC (sNX), which denotes that this chunk is part of an NC, to which it is not adjacent. Thus the sentence in (5a) has the chunk structure shown in (5b).

The type of complex NPs shown in the previous section can become arbitrarily complex. The example in (6a) with its syntactic analysis in Figure 3 shows that the attributively used adjective “sammelnden” can have all its complements and adjuncts. Here, we have a reflexive pronoun “sich” and a complex PP “direkt vor ihrem Sezessions-Standort am Karlsplatz”. The chunk analysis based on the principles from section 4 gives us the analysis in (6b). The complex PP is represented as three different chunks: an ADVC, and two PCs.

(6) a. Sie “thematisierten” auf Anraten des jetzigen Staatskurators Wolfgang Zinggl die sich direkt vor ihrem Sezessions-Standort am Karlsplatz sammelnden Fixer.

’On the advice of the current state curator Wolfgang Zinggl, they “broach the issue” of the junkies who gather right in front of their location of secession at the Karlsplatz .’

b. [NX Sie] “ [VXFIN thematisierten] ” [PX auf Anraten] [NX des jetzigen Staatskurators] [NX Wolfgang Zinggl] [sNX die] [NX sich] [ADVX direkt] [PX vor ihrem Sezessions-Standort] [PX am Karlsplatz] [NX sammelnden Fixer].

6 Conclusion

In this paper, we have shown how a CoNLL-style chunk representation can be derived from TüBa-D/Z. For the complications stemming from complex prenominal modification, we proposed an analysis in which the stranded determiner is marked as such. For the future, we are planning to make this chunk representation available to license holders of the treebank.

References

Steven Abney. 1991. Parsing by chunks. In Robert Berwick, Steven Abney, and Caroll Tenney, editors, *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, Dordrecht.

Steven Abney. 1996. Partial parsing via finite-state cascades. In John Carroll, editor, *ESSLLI Workshop on Robust Parsing*, pages 8–15, Prague, Czech Republic.

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

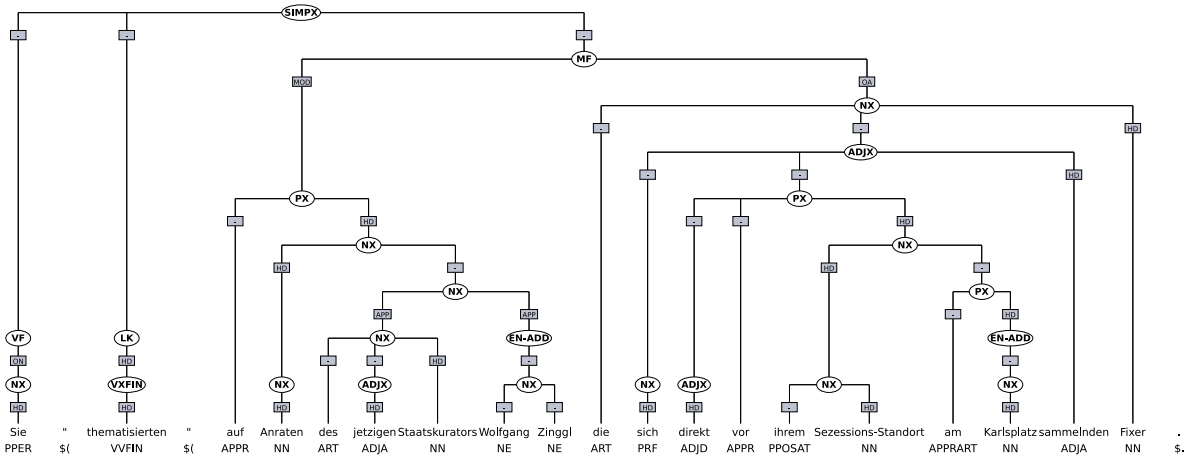

Figure 3: Treebank annotation for the sentence in (6a).

Tilman Höhle. 1986. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Fred Karlsson, Atro Voutilainen, J. Heikkilä, and Atro Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Hannah Kermes and Stefan Evert. 2002. YAC – a recursive chunker for unrestricted German text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Gran Canaria.

Frank H. Müller. 2004. Annotating grammatical functions in German using finite-state cascades. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the ACL 3rd Workshop on Very Large Corpora*, pages 82–94, Cambridge, MA.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2229–2235, Lisbon, Portugal.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck, 2009. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

Erik Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL shared task: Chunking. In *Proceedings of The Fourth Conference on Computational Language Learning, CoNLL'00, and the Second Learning Language in Logic Workshop, LLL'00*, pages 127–132, Lisbon, Portugal.

Julia S. Trushkina. 2004. *Morpho-Syntactic Annotation and Dependency Parsing of German*. Ph.D. thesis, Eberhard-Karls Universität Tübingen.

Tylman Ule. 2007. *Treebank Refinement: Optimising Representations of Syntactic Analyses for Probabilistic Context-Free Parsing*. Ph.D. thesis, Eberhard-Karls Universität Tübingen.

Yannick Versley and Ines Rehbein. 2009. Scalable discriminative parsing for German. In *Proceedings of the International Conference on Parsing Technology (IWPT'09)*, Paris, France.