A High Quality Partial Parser for Annotating German Text Corpora

Stefan Klatt

Institute for Intelligent Systems – University of Stuttgart
Universitätsstr. 38, D-70569 Stuttgart, Germany
klatt@iis.uni-stuttgart.de

Abstract

In this paper, a two-stage partial parser for untagged German sentences is presented. In the first stage, the sentence is segmented into better parsable units according to the Topological Field Model. In the second stage, *minimal phrases* of NPs, DPs and PPs as well as nominal multiword units are identified in each of the recognized fields. In this paper, we discuss the results of the second stage. We evaluated 500 parsed sentences of a newspaper corpus. The achieved recall and precision rates are better than the ones of comparable systems as reported in literature so far.

1. Introduction

Because ambiguity is still one of the central problems for full parsing, partial parsing is used for many NLP tasks. For the task of infomation extraction, it is easier to extract domain and scenario specific information with a partial parser instead of determining the correct syntactic analysis from a huge set of ambiguous analyses.

Furthermore, partial parses are of great value for corpus-based computational lexicography. For the extraction of subcategorization frames of content words (e.g. verbs) and the identification of light verb constructions, corpora annotated with partial parses are very fruitful. Partial parses are also helpful for syntactic grammar refinement, for disambiguation tasks like PP-attachment (Hindle and Rooth, 1993), semantic clustering (Riloff and Shepherd, 1997) as well as an input source for building fully parsed treebanks (Skut et al., 1998).

In (Klatt, 1997), we suggested a strategy for parsing German sentences consisting of three stages. The partial parser described here is a slight modification of the first two stages. In the first stage, a sentence is segmented according to the Topological Field Model for German (cf. (Drach, 1937), (Höhle, 1986)), shortly TFM. In the second stage, so-called *minimal phrases* are recognized in each recognized topological field. In the third stage, a fully parsed structure is assigned – first on the field level, then on the sentence level. But here, we're confronted with the problem of ambiguity – one of the major problems in parsing – where disambiguation strategies using treebank information seems to be the best solution. Unfortunately, the size of the existing treebanks for German (e.g. the TIGER project (Brants et al., 2002)) is too small to be applicable for such a task.

This problem is one motivation (beneath the other applications mentioned before) for the construction of our partial parser, that is based on the analysis technique Pattern-Matching Easy-First Planning, shortly PEP (Klatt, 1997). In opposite to mainstream techniques a sentence is not strictly processed from left to right. Instead we prefer an easy-first strategy, doing the easier decisions before the harder ones, as described in (Abney, 1996).

In the next two chapters, we introduce the structures we want to recognize and illustrate how this could be done with PEP. In the fourth chapter, we describe the recognition process for finding *minimal phrases* in the so-far received segmented fields of the first parsing stage[1]. In the fifth chapter, we present the evaluation of the identified structures, before we show in the sixth chapter some worthwile applications and extensions.

2. Structures to be recognized

2.1. Topological Fields

For the segmentation of a sentence into its topological fields, we make use of an extension of the TFM by Rehbein (cf. (Rehbein, 1992)). The extended TFM splits up a sentence into seven fields. A so-called sentence bracket (SK) consisting of a left (LK) and a right part (RK) segments a sentence into a top, middle and bottom field (in German 'Vorfeld' (VF), 'Mittelfeld' (MF) and 'Nachfeld' (NF)). Giving coordinations and punctuations a home, Rehbein extends this model by the fields 'Satzanfangsrahmen' (SAR) and 'Satzenderahmen' (SER).

(1) Sie hoffen zutiefst, dass sie gewinnen werden.
They deeply hope that they will win.

(2)

SAR	VF	LK	MF	RK	SER	NF
	Sie	hoffen	zutiefst		,	(3)

(3)

SAR	VF	LK	MF	RK	SER
		dass	sie	gewinnen werden	.

Because we have to process *real-life* sentences, we extend the TFM by some more fields. E.g. SKEL marks a sentence bracket, where all verbs were elided. Furthermore, we use different SK-annotations with respect to their clausal subtype: SKI marks a verb-first- or a part of a verb-second-clause (cf. (4)[2]).

(4) [VF Portugal] [SKI [LK wird] [MF im Finale Spanien] [RK schlagen]] und [SKEL [MF Frankreich davor]].
Portugal will beat Spain in the final and France before.

(5)

SAR	VF	LK	MF	RK	SER
und			Frankreich davor		.

[1] A detailed description and evaluation of this parsing stage is beyond the scope of this paper and will be given in a own publication.

[2] Note that empty fields are omitted in some of the examples.

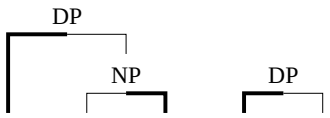
2.2. Intra-clausal minimal phrases

In the second parsing stage, we determine so-called *minimal phrases* of selected categorial heads (nouns, determiners, prepositions) of each recognized VF, MF and NF as well as phrases with an adjectival head modified by special adverbials. A *minimal phrase* could be considered as a structured chunk comparable to the chunk definition in (Kermes, 2003). The traditional notion of a chunk is that of a flat, non-recursive structure (Abney, 1991) (cf. (6)). Kermes extends this definition by two aspects (i) recursive embedding and (ii) post-head embedding (cf. (7)).

(6) [NP das erste Tor] [NP des Turniers]
the first goal of the tournament

(7) [NP [NP das [AP erste] Tor] [NP des Turniers]]
the first goal of the tournament

In (8), we see the *minimal phrases*, we assign to this text. First, we consider a determiner as a governor of a NP according to the DP hypothesis (Abney, 1987)[3]. Second, we make no post-head embedding, since such a decision is dependent from the lexical verb (cf. (9) and (10)).

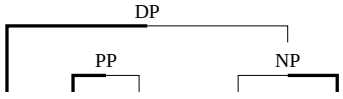


(8) das erste Tor des Turniers

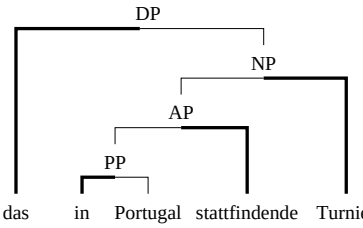
(9) Ich habe [[das Auto] [meiner Frau]] gefahren.
I have driven the car of my wife

(10) Ich habe [das Auto] [meiner Frau] geschenkt.
I have donated the car to my wife

Third, for simple constructions like the two DPs in (8), we assign a full parse to each of them. For complex constructions like the one in (11), we assign a so-called closure, a not fully and also – strictly speaking – ill-formed structure. But this can be easily corrected later (e.g. if this structure is recognized as the only part of a VF) by two tree operations (cf. (12)).



(11) das in Portugal stattfindende Turnier
The tournament taking place in Portugal



(12) das in Portugal stattfindende Turnier

We define a *minimal phrase* as a parse tree reaching from the head of a phrase to its governed element that (i) leaves out post-head embedding and (ii) models recursive embedding of the same categorial head by a closure. These are the structures, we want to annotate in the second parser stage – as well as nominal multiword units (N-MWUs) as shown in (13).

(13) [DP die Dissidentin] [N-MWU Aung San Suu Kyi]
the dissident Aung San Suu Kyi

[3]Syntactic heads are printed by thick lines.

3. The Analysis Technique PEP

Pattern-Matching Easy-First Planning (Klatt, 1997), shortly PEP, is an analysis technique that can be used for several analyis tasks. We built a tokenizer (Klatt and Bohnet, 2004), a POS-tagger (Klatt, 2002) and this partial parser for German which all outperform existing systems. The analysis process of PEP is driven by traversing a transition network deterministically. A transition network is defined by several states that are linked by directed arcs. Every arc is associated with a so-called network function (NWF). PEP has a lot of NWFs, the most important ones are one corresponding pattern for finding adjacent elements (the NWF `nb`, cf. (14)) and one corresponding pattern for finding non-adjacent patterns with a left and right border element (the NWF `seq`, cf. (15)).

(14) Lisbon is a [nb [NB$_1$=ADJA nice] [NB$_2$=NN town]].

(15) Lisbon is [seq [LB=DET a] nice [RB=NN town]].

PEP allows a bi-directional search for patterns from every token position inside the input text. There exist three different ways for searching a pattern: a *categorial-driven* search, a *positional-driven* search and a *token-driven* search. (17) is an example of a categorial-driven search pattern. Here, all adjacent determiners and nouns in (16) are detected and merged to a NP constituent (see (18)). (19) is an example of a positional-driven search pattern. Assuming that the word position pointer `*s-top*` is positioned at the beginning of the first word in (16), only the first determiner-noun-pair is detected (see (20)). In (19) the left context, which is associated with `*s-top*` (`:LC *s-top*`), is chosen as the anchor point of the pattern (`:AP LC`). For the *token-driven* search as well as additonal features of PEP see (Klatt, 1997).

(16) The ball hit the bar.

(17)
```
(nb ((m-cat DET)) ((m-cat N))
 :match-as ((m-cat NP)))
```

(18) [nb [DET The] [N ball]] hit [nb [DET the] [N bar]].

(19)
```
(nb ((m-cat DET)) ((m-cat N))
 :AP LC :LC *s-top*
 :match-as ((m-cat NP)))
```

(20) [nb [DET The] [N ball]] hit the bar.

4. Minimal phrase recognition

In this chapter, we describe the recognition of *minimal phrases* in the second parsing stage in more detail. Usually every NP, DP and most PPs consist of a right-peripheral noun in German. We make use of this property by choosing each noun inside a VF, MF or NF as the starting point for the recognition of the *minimal phrases*. We process a field from left to right, go to the next noun (with a seq-pattern) and apply the following stages to it.

4.1. Recognizing fully parsed subtrees

Determining the left adjacent element of the noun (with a nb-pattern), we build a binary branching tree[4] , if the left

[4]The merged structures are printed in boldface.

neighbour is a proper determiner, preposition or attributive adjective (cf. (21)). If the left neighbour is an uppercase written word, but no known first name, we apply a corpus-based test to recognize N-MWUs in the surrounding context. This test identifies *Aung San Suu Kyi* in (22) as a N-MWU. After that, we iterate the process for the newly built structure and its left neighbour (cf. (23) and (24)).

(21) mit dem [NP **ersten Friedensnobelpreis**]
with the first Nobel Peace Prize

(22) die Dissidentin [N-MWU **Aung San Suu Kyi**]
the dissident Aung San Suu Kyi

(23) mit [DP **dem [NP** ersten Friedensnobelpreis **]**]

(24) [PP **mit [DP** dem [NP ersten Friedensnobelpreis] **]**]

This substage also leads to a partial recognition of closures (cf. (25)) that will be continued in the next substage.

(25) mit der [PP in Birma] [NP **lebenden Frau**]
with the woman living in Birma

4.2. Recognizing closures of subtrees

In this substage, we process all topological fields in a second run by identifying the first motherless NP with a left adjacent attributive adjective (ADJA-N-pair) from left to right. Next, we identify the nearest proper non-adjacent possible governor (determiner or preposition) to its left. If the governor is *motherless*, we mark these two elements as the border elements and continue with the strategy of the substage described before to merge left adjacent elements with the closure (cf. (26)).

(26) [PP mit [DP **der** [PP in Birma] **[NP lebenden Frau]]**]

After that, we start an iteration of this process with the next right ADJA-N-pair. If the governor stands in a **non-motherless** relation (cf. (27-(29)), we build a closure, too. But in this case, we assign a lower confidence factor to it preferring the alternative analysis. Sometimes, the higher ranked analysis is correctly eliminated by the ongoing analysis process, e.g. by identifying the closure in a VF-position (cf. (29)) or in respect to the subcat frame constraints of the lexical verb.

(27) dass er [DP **die** Blumen] **[NP** singenden Frauen **]** gab
that he gave the flowers to the singing women

(28) dass er [DP **die** Lieder] **[NP** singenden Frauen **]** traf
that he met the women singing songs

(29) [DP **Die** Lieder **[NP** singenden Frauen **]**] tanzten
The women singing songs were dancing

4.3. Picking up the rest

At last, we're picking up the rest, e.g. pronouns with left-peripheral prepositons (cf. (30)), NPs with a prenominal genitiv (cf. (31)), first names that followed by an uppercase written word. In the latter case, we don't prefer one of the two assigned analyses, since both could be correct (cf. (32) and (33)).

(30) Er wartet [PP auf sie].
He's waiting for her.

(31) [DP Birmas [NP erste Friedensnobelpreisträgerin]]
Birmas first Nobel Peace Prize laureate

(32) dass sich [NP Gottfried Dienst] irrte
that Gottfried Dienst made a mistake

(33) dass [N Gottfried] [N Dienst] hatte
dass Gottfried was on duty

For the sake of evaluation[5], we applied a longest-match strategy to the recognized structures whose confidence factor matches a parameterizable threshold.

5. Evaluation

For the evaluation of the *minimal phrase* recognition, we've chosen the first 500 sentences of the REFD-corpus[6]. Table 1 shows a frequency distribution respective to the token length and the construction type of the identified NPs, DPs and PPs. Note that we didn't count XPs of the word length 1.

	token length						
XP	constr	2	3	4	5	≥ 6	$\sum$
NP	nb	574	132	32	9	6	753
	mwu	27	3	4	1	3	38
DP	nb	932	269	43	9	7	1260
	seq	–	–	11	10	14	35
PP	nb	334	379	134	32	10	889
	seq	–	–	4	3	4	11

Table 1: Frequency Distribution of selected XPs

The evaluated results in terms of precision and recall are shown in Table 2. Recall and precision were computed as follows: $Rec = \frac{corr*100\%}{corr+miss}$, $Prec = \frac{corr*100\%}{corr+spur}$.

	constr	freq	corr	miss	spur	Prec.	Rec.
NP	mwl	39	37	2	1	98.09	96.74
	nb	763	739	24	14		
DP	nb	1260	1249	11	21	98.39	99.00
	seq	35	32	3	–		
PP	nb	890	874	16	15	98.22	98.00
	seq	12	10	2	1		

Table 2: Precision and Recall of selected XPs

5.1. Discussion of NP results

We identified 38 N-MWUs in a rule-based fashion by special suffixes (e.g. *Rudolf Hell GmbH*) as well as by a corpus-based strategy (e.g. *Assurances Generales de France*). Only one of the N-MWUs wasn't detected properly. Instead of assigning *Heidelberger Zement AG* a MWU-reading, we did this only for *Zement AG* assuming that *Heidelberger* has the reading of an adjective of origin. In (34) and (35), we see the most complex structured NPs. In (36), we see two spurious and one missing NP annotations, since we wrongly recognized *Jahren wie Pilze* as a noun coordination.

(34) `<pp> in <dp> der <np>` **Außen- , <np> Sicherheits- , <np> Sozial- , <np> Wirtschafts- und Finanzpolitik** `</np> </np> </np> </np> </dp> </pp>`

(35) `<dp> die <np> <adj> klapprigen und stinkenden </adj>` **<np> alten <np> Laster und Busse** `</np> </np> </np> </dp>`

(36) `<pp> in <dp> den <np>` **letzten <np> Jahren wie Pilze** `</np> </np> /dp> </pp>`

[5] In our full parsing strategy, we regret of such a disambiguation, hoping that the ongoing analysis process will dismiss some ambiguous ill-formed structures.

[6] Thanks to the Institute for Natural Language Processing (IMS) of Stuttgart for making the corpus available to us.

5.2. Discussion of DP and PP results

(36) is also an example of a propagating error for the DP and PP recognition, which was the most frequent error source (in total: 10 missing and 5 spurious DPs, 13 missing and 9 spurious PPs). 23 of the 35 DP closures, recognized with a seq-pattern, possess one XP between the border elements (cf. (37)). In this case, we can easily generate the correct structure by labelling the adjective as governor of the left adjacent XP. Such a correction isn't possible by two or more XPs in between the border elements (cf. (38)). In (39), we see two correctly annotated PPs.

(37) <**dp**> **einen** durchaus <**np**> logischen <np> nächsten Schritt </np> </**np**> </**dp**>

(38) <**dp**> **das** <pp> am Freitag </pp> <pp> im Bundestag </pp> <**np**> angenommene Gesetz </**np**> </**dp**>

(39) <dp> die <**pp**> **von ihm** </**pp**> <np> beauftragten Wirtschaftsprüfer </np> </dp> <**pp**> **von** <**np**> Price Waterhouse </**np**> </**pp**>

6. Possible Applications and Extensions

In principle, our parser can be used for any NLP task where partial parsing is involved. It could help building fully parsed treebanks in a semi-automatically way. To make our tree format compatible to existings formats, a simple post-processing strategy is necessary. Especially the TFM-segmentation allows further worthwile applications. For instance, our parser is very suitable for the acquisition of light verb constructions (LVCs) at the end of MF and RK. For the case that RK is empty, we take the verb in LK. We can use it also for the task of subcat frame recognition of content words (incl. LVCs). Furthermore, it supports syntactic and semantic grammar refinement. If there are two constituents in a VF-position and our grammar rules don't allow us to merge them, the *One-consituent-in-VF*-constraint tells us that the two constituents must belong together. So we can identify the noun *Friedensnobelpreisträgerin*, denoting a female person, as a kind of a title.

(40) [VF Friedensnobelpreisträgerin [N-MWU Aung San Suu Kyi]] [SKI hat ...]
Nobel price laureate Aung San Suu Kyi has ...

A worthwile extension would be the integration of more corpus-based tests. In the case of the seldomly occuring conjunction *wie*, we can substitute the conjunction candidate by a more frequently used one, e.g. *und*, expecting to find more occurences of the latter coordination in the corpus, what is actually the case (cf. (41)). The ratio for the constellation *Jahren wie/und Pilze* is 2:0, what implies that this isn't a noun coordination.

(41) Männer wie/und Frauen – (13/321 occ.)
Men as/and women

7. Summary

We presented a partial parser for German, that combines rule-based with corpus-based decisions. In a first step, it segments and annotates sentences in corpora into better parsable units according to the Topological Field Model. In a second step, it recognizes so-called *minimal phrases* in the previously segmented fields. An evaluation of the second stage demonstrated the high quality of the parser, that can be used for many analysis tasks as well as for the task of corpus annotation.

8. References

Steven P. Abney. 1987. *The English Noun Phrase in its Sentential Aspect*. Ph.D. thesis, MIT, Cambridge, Mass.

Steven P. Abney. 1991. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle–based Parsing*. Kluwer, Dordrecht.

Steven P. Abney. 1996. Partial parsing via finite–state cascades. In *Workshop on Robust Parsing*, Prague. 8th European Summer School in Logic, Language and Information.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Erich Drach. 1937. *Grundgedanken der deutschen Satzlehre*. Diesterweg, Frankfurt.

Donald Hindle and Mats Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1):103–120.

Tilman Höhle. 1986. Der Begriff Mittelfeld. Anmerkungen über die Theorie der topologischen Felder. In W. Weiss, Wiegand E. H., and M. Reis, editors, *Textlinguistik contra Stilistik/Wortschatz und Wörterbuch/Grammatische oder pragmatische Organisation von Rede*. Niemeyer, Tübingen.

Hannah Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, Universität Stuttgart.

Stefan Klatt and Bernd Bohnet. 2004. You don't have to think twice if you carefully tokenize. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Hainan (China).

Stefan Klatt. 1997. Pattern-matching Easy-first Planning. In Alice Drewery, Geert-Jan M. Kruijff, and Richard Zuber, editors, *The Proceedings of the Second ESSLLI Student Session*, Aix-en-Provence. 9th European Summer School in Logic, Language and Information.

Stefan Klatt. 2002. Combining a Rule-Based Tagger with a Statistical Tagger for Annotating German Texts. In Stephan Busemann, editor, *KONVENS 2002. 6. Konferenz zur Verarbeitung natürlicher Sprache*, Saarbrücken (Germany).

Jochen Rehbein. 1992. Zur Wortstellung in komplexen deutschen Sätzen. In Ludger Hoffmann, editor, *Deutsche Syntax: Ansichten und Aussichten*. Walter de Gruyter, Berlin, New York.

Ellen Riloff and Jessica Shepherd. 1997. A Corpus-Based Approach for Building Semantic Lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*.

W. Skut, T. Brants, B. Krenn, and H. Uszkoreit. 1998. A Linguistically Interpreted Corpus of German Newspaper Text. In *Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany. 10th European Summer School in Logic, Language and Information.