4

Finite-State Parsing of German

ERHARD W. HINRICHS

4.1 Introduction

There has been a remarkable revival of finite-state methods in linguistics over the last twenty-five years. This renewed interest is a direct consequence of the pioneering work on two-level phonology and morphology by Kimmo Koskenniemi (Koskenniemi, 1983) and of the independently developed approach to finite-state morphology by Ron Kaplan and Martin Kay (Kaplan and Kay, 1994). Based on mathematically rigorous models of finite-state transduction, there are now wide-coverage finite-state accounts of an impressive range of typologically diverse languages available. Inspired by these successes, research of finite-state models for syntactic analysis was revived in the early nineties, notably by Stephen Abney (Abney, 1991) and by Fred Karlsson and his associates (Karlsson et al., 1995)[1]. Their research ended a period of more than three decades of little or no research on finite-state models of syntax under the influence of Chomsky's claim that finite-state automata are inadequate due to their inability to account for center-embedding construction in natural languages (Chomsky, 1963).

The two alternative models of finite-state syntax developed by Abney and Karlsson reflect in an interesting way two leading paradigms for representing syntactic structure. Abney's chunk parser is designed to provide a partial bracketing of an input text. This bracketing identifies non-recursive phrases, so-called *chunks*, which span from the left periphery of a phrase to its phrasal head. The resulting bracketing is partial in that it leaves any structural rela-

[1] Rules in Constraint Grammar are, in isolation, implementable with finite-state methods. Editor's comment.

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

tionships between individuals chunks unresolved.

Karlsson's constraint-grammar formalism is designed to provide a shallow syntactic parse of an input text that identifies the beginnings and ends of non-recursive phrases and the grammatical functions of verbal complements.

The purpose of this paper is to review recent work on finite-state syntactic analysis of German. Rather than comparing the details of individual finite-state parsing systems for German, the discussion will focus on those aspects of German sentence structure that make German an interesting language from a finite-state perspective. Section 4.2 surveys existing finite-state and constraint-based parsers of German. Section 4.3 discusses complex prenominal modifier structures in German which are recursive in nature. Their recursiveness provides an interesting challenge for Abney's conception of what a chunk is. Section 4.4 gives an overview of the main characteristics of German sentence structure This provides the necessary background for the discussion of interesting challenges and opportunities that the sentence structure of German poses for finite-state approaches. This discussion is the topic of section 4.5.

4.2 A Survey of Finite-State and Constraint-Grammar Parsers of German

Most of the research on finite-state parsing of German has utilized Abney's chunk parsing model and produces partial bracketings of the input text. Two recent examples of Abney-style chunk parsers for German are the Dereko parser (Müller and Ule, 2001) and the YAP parser (Kermes, 2002, Kermes and Evert, 2002). In addition to finite-state chunk parsers, Schmid and Schulte im Walde (2000) have developed a statistical chunk parser for German that is based on probabilistic context-free grammars.

There are at least four parsers for German that use finite-state methods internally and produce dependency relations. Connexor, a Finnish language technology company, has developed a syntactic parser called *Machinese Syntax* for a variety of languages, including German, that produce part-of-speech classes, inflectional tags, noun phrase markers and syntactic dependencies for written input. The output representations follow the style of annotation familiar from Constraint Grammar.[2] Schiehlen (2003) has developed a finite-state parser for German that produces dependency relations and that uses underspecification to encode ambiguities that arise from alternative valence frames of verbs and from alternative attachment sites for PP modifiers. Most recently, Trushkina (2004) and Müller (2005) have developed parsers that combine chunk parsing with dependency parsing. Trushkina's GRIP parser is based

[2] Duchier (1999) and Foth et al. (2004) have also developed dependency parsers for German, albeit without explicitly relying on finite-state methods.

on the Xerox Incremental Parsing System (XIP) (Aït-Mokhtar et al., 2002), while Müller's parser uses the suite of lcg tools (Mikheev et al., 1998, 1999). Both parsers are limited to those dependency relations that refer to complements and do not deal with adjuncts.

4.3 Prenominal Modifiers and Recursive Chunk Structures

One of the syntactic constructions that make German an interesting language from a chunk parsing perspective are complex prenominal modifiers such as the participial construction as in (1a).

(1) a. der seinen Sohn liebende Vater
the his son loving father
'the father who loves his son'

b. $[_{NC}$ der $[_{NC}$ seinen Sohn] liebende Vater]

Such examples are interesting since they do not simultaneously satisfy the two defining properties that Abney associates with the term *chunk*. Abney (1996) defines the notion of chunk as " ... the non-recursive core of an intra-clausal constituent, extending from the beginning of the constituent to its head." In the case of (1a), the article *der* and the nominal head *Vater* seem to represent the left and right periphery of a nominal chunk. However, chunks are also defined as non-recursive structures. This seems to suggest that only the substring *seinen Sohn* qualifies as a noun chunk (NC) and seems rule out the structure in (1b), where the entire string is a nominal chunk as well. In fact, Abney appeals to the "no chunk within a chunk"-constraint to explain the ungrammaticality of English NPs as in (2).

(2) * the proud of his son father

For cases like (1), there seem to be two solutions to this impasse: one may argue that only the NP inside the premodifier, or one considers the complex NP as a nominal chunk and gives up.

A telling piece of evidence in favor of the latter solution is provided by the grammaticality of (3), the German counterpart of (2).

(3) der auf seinen Sohn stolze Vater
the on his son proud father
'the father who is proud of his son'

This seems to suggest that Abney's "no chunk within a chunk"-constraint is not universally applicable across languages, even though it does seem to hold for English. However, the assumption that chunks are non-recursive in nature is not only motivated by examples such as (2). Notice that once one accepts recursive bracketings shown in (1), one allows center-embedding constructions. The fact that natural languages allow for such constructions was

identified by Chomsky as the key argument for rejecting finite-state models for natural language analysis. If one allows center-embeddings to an arbitrary level of embedding, then the analysis of such constructions lies beyond the expressive power of regular grammars. Chomsky's argument crucially rests on the assumption that there is in principle no depth bound on the number of embeddings inside a center-embedding construction. Chomsky readily admits that there are, of course, processing limitations by language users that limit center-embeddings to two or at most three for a given utterance. However, such upper bounds, he argues, should be considered aspects of performance grammar, not of competence grammar. If one accepts this argument then the inadequacy of finite-state grammars seems to refer to competence grammar only. Thus, if one views a finite-state parser as a model of performance grammar, then one can simply impose a reasonable depth bound on center-embedding constructions in a finite-state grammar. This is precisely what Kermes (2002) and Müller (2005) have done in order to be able to treat complex prenominal modifiers as part of chunks that exhibit limited, i.e. depth-bounded, recursion. In addition to complex, prenominal modifiers, Kermes' YAC parser also admits a limited number of post-head nominal modifiers as in (4).

(4) a. die Köpfe der Apostel
the heads of the apostles
'the heads of the apostles'

b. Jahre später
years later
'year later'

In order to accommodate examples such as (1), (3), and (4), Kermes (2002) modifies Abney's definition of a chunk as in (5).

(5) A chunk is a continuous part of an intra-clausal constituent including recursion, pre-head as well as post-head modifiers, but no PP-attachment or sentential elements.

4.4 The Macro-structure of German: topological fields

One of the characteristic features of German syntax is the placement of the finite verb in different clause types. Consider the finite verb *wird* in (6) as an example.

(6) a. Peter wird das Buch gelesen haben.
Peter will the book read have
'Peter will have read the book.'

b. Wird Peter das Buch gelesen haben?
Will Peter the book have read
'Will Peter have read the book?'

c. dass Peter das Buch gelesen haben wird.
that Peter the book read have will
'... that Peter will have read the book.'

In non-embedded assertion clauses, the finite verb occupies the second position in the clause, as in (6a). In yes/no questions, as in (6b), the finite verb appears clause-initially, whereas in embedded clauses it appears clause finally, as in (6c). Regardless of the particular clause type, any cluster of non-finite verbs, such as *gelesen haben* in (6a) and (6b) or *gelesen haben wird* in (6c), appears at the right periphery of the clause.

The discontinuous positioning of the verbal elements in verb-first and verb-second clauses is the traditional reason for structuring German clauses into so-called *topological fields* (Erdmann, 1886, Drach, 1937, Höhle, 1986). The positions of the verbal elements form the *Satzklammer* (sentence bracket) which divides the sentence into a *Vorfeld* (initial field), a *Mittelfeld* (middle field), and a *Nachfeld* (final field). The Vorfeld and the Mittelfeld are divided by the *linke Satzklammer* (left sentence bracket), which is realized by the finite verb or (in verb-final clauses) by a complementizer field. The *rechte Satzklammer* (right sentence bracket) is realized by the verb complex and consists of verbal particles or sequences of verbs. This right sentence bracket is positioned between the Mittelfeld and the Nachfeld. Thus, the theory of topological fields states the fundamental regularities of German word order.

The topological field structures in (7) for the examples in (6) illustrate the assignment of topological fields for different clause types.

(7) a. $[_{VF}$ $[_{NC}$ Peter]] $[_{LK}$ wird] $[_{MF}$ $[_{NC}$ das Buch]]
$[_{RK}$ $[_{VC}$ gelesen haben.]]

b. $[_{LK}$ Wird] $[_{MF}$ $[_{NC}$ Peter] $[_{NC}$ das Buch]]
$[_{RK}$ $[_{VC}$ gelesen haben?]]

c. $[_{LK}$ $[_{CF}$ dass]] $[_{MF}$ $[_{NC}$ Peter] $[_{NC}$ das Buch]]
$[_{RK}$ $[_{VC}$ gelesen haben wird.]]

(7a) and (7b) are made up of the following fields: LK (linke Satzklammer) is occupied by the finite verb. MF (Mittelfeld) contains adjuncts and complements of the main verb. RK (rechte Satzklammer) is realized by the verbal complex (VC). Additionally, (7a) realizes the topological field VF (Vorfeld), which contains the sentence-initial constituent. The left sentence bracket (LK) in (7c) is realized by a complementizer field (CF) and the right sentence bracket (RK) by a verbal complex (VC) that contains the finite verb *wird*.

4.5 Finite-state Parsing of German

The structure of topological fields delineates the borders and the composition of a clause and thus reveals the overall anatomy of a sentence. It turns out

that topological fields together with chunked phrases provide a solid basis for a robust analysis of German sentence structure. All chunk parsing systems mentioned in section 4.2 adopt an annotation strategy which annotates the topological fields for the left and right sentence brackets before identifying any other fields or chunks. To my best knowledge, this strategy was first proposed by Braun (1999) and by Neumann et al. (2000) as a means of identifying sentence boundaries for German.

It turns out that the advantages of topological field annotation go significantly beyond sentence boundary detection. Robust identification of topological fields can help reduce the search space for subsequent chunk annotation since chunks can only occur within the boundaries of a given topological field.

(8) [$_{VF}$ [$_{NC}$ Außenminister Joschka Fischer]] [$_{LK}$ hat] [$_{MF}$ [$_{NC}$ die Abgeordneten]] [$_{RK}$ gebeten] [$_{NF}$ [$_{MF}$ [$_{NC}$ die Entscheidung]] [$_{RK}$ zu verschieben.]]

'Foreign Minister Joschka Fischer asked the members of parliament to postpone the decision.'

(8) is a V2-clause with an extraposed *zu*-infinitive that is governed by the verb form *gebeten*. Such extraposed constituents are positioned in the topological field Nachfeld (NF). By locating the noun chunks *die Abgeordneten* and *die Entscheidung*, which occurs the Mittelfeld of the V2-clause, in different topological fields, it becomes clear that they modify the verbs *gebeten* and *verschieben*, respectively.

Recognition of topological fields can also effectively reduce potential ambiguities that can arise if only local syntactic context is taken into account.

(9) [$_{VF}$ [$_{NC}$ Man]] [$_{LK}$ sah] [$_{MF}$ [$_{PC}$ in [$_{NC}$ der Öffentlichkeit]] [$_{ADVC}$ nur] [$_{NC}$ Männer] [$_{PC}$ mit [$_{NC}$ Zigarette]]] [$_{KOORD\_F}$ und] [$_{VF}$ [$_{NC}$ rauchende Frauen]] [$_{LK}$ waren] [$_{MF}$ [$_{NC}$ ein Thema] [$_{PC}$ für Karikaturen]]

'In public, you saw only men with cigarettes, and smoking women were a topic for caricatures.'

In (9) the coordination *und* forms a coordination field (KOORD_F) with two V2 clauses as sentential conjunctions. The parallelism between the two clauses can be easily detected in terms of the their left and right sentence brackets and their Vorfeld constituents. However, if only local context is taken

into account, the coordination may be misanalysed as an NP conjunction between the two NP chunks *Männer mit Zigarette* and *rauchende Frauen*.

Identifying the left and right sentence bracket of a clause prior to any other syntactic chunk annotation follows the principle of "easy first" parsing advocated by Abney since these two sentence brackets can be detected with great reliability for any clause type of German.

As shown by Müller and Ule (2001), Hinrichs et al. (2002), Müller and Ule (2002), another class of ambiguities that can be resolved by topological field information concerns potential ambiguities in part-of-speech assignments to lexical tokens. Two classes of common tagging errors in German concern the distinction between finite and non-finite verb forms and the distinction between homonymous prepositions and subordinating conjunctions.[3] The token *seit* in (10) is ambiguous between a preposition (APPR) or a subordinating conjunction (KOUS).[4]

(10) $[_{VF}$ $[_{LK}$ $[_{CF}$ Seit]] $[_{MF}$ Banting und Best Insulin zum ersten Mal] $[_{RK}$ $[_{VC}$ isolieren konnten]]], $[_{LK}$ haben] $[_{MF}$ die Mediziner lebenserhaltende Kontrolle über Diabetiker] $[_{RK}$ $[_{VC}$ gewinnen können]].

'Ever since Banting and Best have been able to isolate insulin for the first time, physicians have been able to win life-preserving control of diabetes.'

The theory of topological fields helps to determine the correct tag for *seit* in such cases. The entire clause is a verb-second clause with an embedded clause occupying the clause-initial position. The embedded clause has to adhere to the constraints on how the left and right sentence bracket have to be realized for a verb-final clause. In particular, the left sentence bracket (LK) has to consist of a complementizer field (CF) which can be realized by a coordinating conjunction (KOUS), but crucially not by a preposition (APPR).

Sentence (11) provides an example of a potential part-of-speech ambiguity between a finite (VVFIN) and a non-finite (VVINF) verb for the verb form *nehmen*.

(11) $[_{VF}$ $[_{NC}$ Libyen]] $[_{LK}$ kann] $[_{MF}$ $[_{NC}$ keinen Einfluss] $[_{PC}$ auf $[_{NC}$ die Politik]] $[_{NC}$ Marokkos]] $[_{RK}$ nehmen]

'Libya can exert no influence on the politics of Marocco.'

[3] See Brants (1999) for more detailed discussion.

[4] The part-of-speech tags used for the annotation are taken from the Stuttgart-Tübingen tagset (STTS) Schiller et al. (1995).

Once again, the topological field assignment shown in (11) uniquely determines that *nehmen* has to be a non-finite verb (VVINF) since the right sentence bracket in a verb-second clause may only contain non-finite verbs.

Notice that the type of topological field information that resolves the two types of part-of-speech ambiguities illustrated by examples (10) and (11) are non-local in nature. The crucial clues for disambiguating the lexical tokens in question span essentially the entire clause. It is for this very reason that such examples pose a serious challenge for both rule-based and statistical taggers.[5]

4.6 Conclusion

This paper has presented a survey of finite-state parsing systems for German and has discussed two aspects of German sentence structure that are of general interest from a finite-state perspective: the treatment of complex prenominal modifiers and the characterization of German clauses structure in terms of topological fields.

References

Abney, Steven. 1991. Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny, eds., *Principle-Based Parsing*. Dordrecht: Kluwer Academic Publisher.

Abney, Steven. 1996. Chunk stylebook. Unpublished manuscript, University of Tübingen.

Aït-Mokhtar, Salah, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering* 8(2–3):121–144.

Brants, Thorsten. 1999. *Tagging and Parsing with Cascaded Markov Models*, vol. 6 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI, Universität des Saarlandes.

Braun, Christian. 1999. *Flaches und robustes Parsen deutscher Satzgefüge*. Diplomarbeit, Universität. des Saarlandes, Saarbrücken.

Chomsky, Noam. 1963. Formal properties of grammars. In R. D. Luce, R. R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology*, vol. II, pages 323–418. John Wiley.

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

Duchier, Denys. 1999. Axiomatizing Dependency Parsing Using Set Constraints. In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL 6)*, pages 115–126. Orlando, FL.

Erdmann, Oskar. 1886. *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt*. Stuttgart: Verlag der Cotta'schen Buchhandlung. Erste Abteilung.

[5] For a more detailed discussion see Trushkina and Hinrichs (2004).

Foth, Kilian, Michael Daum, and Wolfgang Menzel. 2004. A broad-coverage parser for German based on defeasible constraint. In *KONVENS 2004, Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache*, pages 45–52. Vienna.

Hinrichs, Erhard W., Sandra Kübler, Frank H. Müller, and Tylman Ule. 2002. A hybrid architecture for robust parsing of German. In *Proceedings of LREC 2002*. Las Palmas, Gran Canaria.

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

Kaplan, Ronald M. and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20(3):331–378.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Kermes, Hannah. 2002. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, University of Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Kermes, Hannah and Stefan Evert. 2002. YAC – a recursive chunker for unrestricted German text. In M. G. Rodriguez and C. P. Araujo, eds., *Proceedings of the Third International Conference on Language Resources and Evaluation*, vol. V, pages 1805–1812.

Koskenniemi, Kimmo. 1983. Two-level model for morphological analysis. In *IJCAI-83*, pages 683–685. Karlsruhe, Germany.

Mikheev, Andrei, Claire Grover, and Marc Moens. 1998. Description of the LTG system used for MUC-7. In *Proceedings of 7th Message Understanding Conference (MUC-7)*.

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

Müller, Frank Hendrik. 2005. *A Finite-State Approach to Shallow Parsing and Grammatical Functions Annotation of German*. Ph.D. thesis, University of Tübingen.

Müller, Frank Henrik and Tylman Ule. 2001. Satzklammer annotieren und Tags korrigieren. Ein mehrstufiges Top-Down-Bottom-Up-System zur flachen, robusten Annotierung von Sätzen im Deutschen. In H. Lobin, ed., *Proceedings der GLDV-Frühjahrstagung 2001*, pages 225–234. Gießen. Sig.:sb.

Müller, Frank H. and Tylman Ule. 2002. Annotating topological fields and chunks – and revising pos tags at the same time. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan.

Neumann, Günter, Christian Braun, and Jakub Piskorski. 2000. A divide-and-conquer strategy for shallow parsing of German free texts. In *Proceedings of ANLP-2000*, pages 239–246. Seattle, WA.

Schiehlen, Michael. 2003. Combining deep and shallow approaches in parsing German. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo.

Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Tech. rep., Universität Stuttgart and Universität Tübingen.

Schmid, Helmut and Sabine Schulte im Walde. 2000. Robust German Noun Chunking with a Probabilistic Context-Free Grammar. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 726–732. Saarbrücken, Germany.

Trushkina, Julia and Erhard W. Hinrichs. 2004. A hybrid model for morpho-syntactic annotation of German with a large tagset. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 238–246. Barcelona, Spain.

Trushkina, Julia S. 2004. *Morpho-syntactic Annotation and Dependency Parsing of German*. Ph.D. thesis, University of Tübingen.