Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Restricting Discontinuity

Stefan Müller

DFKI GmbH

Vm

Report 237
November 1999

November 1999

Stefan Müller

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken

Tel.: (0681) 302 - 5295
Fax: (0681) 302 - 5338
e-mail: `Stefan.Mueller@dfki.de`

Gehört zum Antragsabschnitt: 2.4

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 701 V0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt beim Autor.
ISSN 0949-6084

Abstract

The paper deals with the parsing of an HPSG grammar for German that employs the concept of word order domains and discontinuous constituents. I will propose constraints on the continuity of phrases. These constraints reduce the search space for parsing and improve the efficicency by a factor of up to 4.

1 Word Order Domains

During the last years several grammarians have argued for linguistic descriptions of language that use the concept of discontinuous constituents (Reape, 1991; Reape, 1992; Reape, 1994; Pollard, Kasper, and Levine, 1992; Pollard, Kasper, and Levine, 1994; Kathol and Pollard, 1995; Kathol, 1995; Müller, 1995; Müller, 1996b; Müller, 1999a).

Reape introduced word order domains and discontinuity into the HPSG framework to describe clause union phenomena (see below). Kathol (1995), Kathol and Pollard (1995), and I, on the other hand, use word order domains to describe extraposition and permutability of constituents in German.

(1) a. , weil der Mann der Frau das Buch gibt.
because the man the woman the book gives
'because the man gives the woman the book.'

b. , weil der Mann das Buch der Frau gibt.
because the man the book the woman gives

The idea is that in (1) the same lexical entry for *gibt* is used in the parse of both sentences. In the first sentence *gibt* is combined with *der Frau* forming a discontinuous constituent, and then *der Frau gibt* is combined with *das Buch*. In (1b), the elements are combined in the same way, but *der Frau gibt* is not discontinuous.

Parsing long sentences or highly ambiguous input with domain-based grammars can lead to combinatorial explosion if the discontinuity is not restricted.

To give an example consider (2).

(2) , weil der || der || das || gibt.
because the the the gives
'because he gives it to her.'

In (2) the NPs correspond to demonstrative pronouns. These are partly homonymous with the definite articles in German. Parsing (1a) produces the discontinuous phrase (2). In (2) the discontinuous parts are marked by '||'.

In the next section I will show how the search space can be dramatically reduced for grammars with certain properties.

[†]Thanks to Ewan Klein for some discussion. I also want to thank Godmar Back for proof reading and some general comments. This paper appeared as (Müller, 1999b).

2 Restricting Discontinuity

All grammars cited above that use discontinuous constituents license discontinuous phrasal signs that are never needed in any analysis. For an example take (3).

(3) ,	weil	den	Männern	Frauen	Bücher	geben.
	because	the	men$_{dat}$	women$_{nom}$	books$_{acc}$	give

'because the men women books give'

Without additional restrictions, the discontinuous dative NP *den Frauen* will be built. Furthermore, this NP is combined with other material yielding even more phrasal signs that are never used in any analysis. Using binary branching structures, the following phrases will be licensed.

(4) a. [den Frauen]$_{dat}$ geben

b. [den Frauen]$_{dat}$ Bücher$_{acc}$ geben

The phrasal signs in (4) can never be used in an analysis of (3), since (4b) cannot be combined with the bare plural NP *Männer*. This combination should be ruled out by LP-constraints that state that an NP cannot appear in the middle of another NP. Without such constraints, the result would be (5).

(5) a. Männer$_{nom}$ [den Frauen]$_{dat}$ Bücher$_{acc}$ geben

b. [den Frauen]$_{dat}$ Männer$_{nom}$ Bücher$_{acc}$ geben

The phonology of the domain elements in (5) does not correspond to the phonology of the input. Therefore, signs with such domains are not licensed by the grammar.[1]

I formulate the following continuity constraint that rules out discontinuous NPs such as *den Frauen* and therefore also the unwanted structures in (4).

Principle 1 (Continuity Constraint) *A saturated phrasal element (a maximal projection) has to be continuous.*

This principle is too strong if extraposition is described via a complex word order domain formation process as suggested by Kathol and Pollard (1995). Kathol and Pollard assume that the marked material in (6) is a constituent.

(6) ,	weil	der	Mann	*ein*	*Lied*	gesungen	hat,	*das*	*jeder*	*kennt*.
	because	the	man	a	song	sung	has	that	everybody	knows

'because the man sang a song which everybody knows'

When the NP *ein Lied, das jeder kennt* is combined with its head *gesungen hat*, the object *ein Lied, das jeder kennt* is split into two parts that are inserted into the domain of the head. The two parts can then be serialized independently. If a grammar used this extraposition analysis, the Continuity Constraint as stated above would be too strong since it would rule out (6). Principle 2 is the adapted version.

[1] In (Müller, 1999a, p. 423–424) and (Müller, 1999c), I claimed that grammars that do not have continuity constraints as the one proposed below overgenerate. Ewan Klein pointed out to me that this claim is wrong for the reasons discussed above.

Principle 2 (Continuity Constraint) *The non-extraposed part of a saturated phrasal element has to be continuous.*

So, in (6) *ein Lied* has to be continuous, and *das jeder kennt* has to be continuous too, since it is a maximal relative clause.

The formulation of such a constraint might appear to be trivial, but it is not. It is dependent upon linguistic analyses for a number of phenomena. The constraint is compatible with grammars such as the one proposed by Kathol (1995) and the one argued for in (Müller, 1999a). It is not compatible with grammars such as the one proposed by Reape (1992).

Reape's grammar is not compatible because he explicitly assumes discontinuous sentential maximal projections in order to describe phenomena such as clause union and modification by adverbs.

(7)	, weil	der	Fritz	die	Maria	seit langem	zu	lieben	scheint.
	because	the	Fritz	the	Maria	for a long time	to	love	seems

'because Fritz seems to have loved Maria for a long time.'
'because it has seemed for a long time that Fritz loves Maria.'

He assumes that in (7) *scheint* takes a sentential complement *der Fritz die Maria zu lieben* which is discontinuous in the reading where the adverbial phrase *seit langem* scopes over *scheint*. In this reading, the whole phrase *Maria den Mann zu lieben* is modified by the adverbial phrase *seit langem*. If one analyzes *zu lieben scheint* as a verbal complex, as suggested by Hinrichs and Nakazawa (1989) and Kiss (1995), one does not have to assume discontinuous maximal sentential projections. In my analysis of (7) *scheint* is combined with either *seit langem zu lieben* (narrow scope) or *zu lieben* (wide scope). In the wide scope reading the adverbial phrase modifies the verbal complex *zu lieben scheint*. *Scheint* raises the arguments of the verbal projection it embeds, i.e., *zu lieben* or *seit langem zu lieben*. These arguments are then combined with the verbal complex.

This principle can be strengthened to further improve the efficiency.

Principle 3 (Continuity Constraint (strong version)) *The non-extraposed part of a saturated phrasal element has to be continuous. The non-extraposed part of a phrasal element may contain no more discontinuous parts than the number of elements in the phrasal element's subcat list plus one.*

This means that projections of the lexical element *schlafen* (*sleep*), which is an intransitive verb, may contain at most two gaps. Projections of the ditransitive verb *geben* (*give*) may contain at most four gaps.

(8)	weil	*oft*	ein Mann	*geschickt*	einer Frau	*schnell*	ein	Buch
	because	often	a man	clever	a woman	fast	a	book
	mit Absicht	*gibt*.						
	with intention	gives						

In the artificial example (8), the verb *gibt* is combined with its adjuncts (the phrases in *italics*) before the complements are inserted into the gaps. As there is no complement that could fill another gap the number of gaps is limited to three.

The one extra gap is needed to account for examples such as (9).

(9) *Schlief* im Wohnzimmer *gestern* ein Mann *lange*?
slept in.the sitting room yesterday a man long

In the analysis of (9) the adverbial phrases are combined with the verbs first. The adverbs are combined from right to left.[2] In the intermediate phrase *Schlief gestern lange*, there are two gaps. The gap between *schlief* und *gestern* has to be filled by an adverbial phrase since no adverbial phrase is allowed to be placed to the right of *gestern*. Therefore, the gap between *gestern* and *lange* has to be filled by an argument, e.g., *ein Mann*. The gap between *schlief* and *gestern* can be filled by an arbitrary number of adverbial phrases. These adverbial phrases have to border *gestern* since otherwise the number of gaps would be increased.

3 Results

The first of the two tables below shows the parse times and the number of passive edges in the chart that are constructed without the strong version of the Continuity Constraint. The second shows the results with the strong version of the Continuity Constraint.

I have chosen example sentences like the one in (10) that are very complex and have a high lexical ambiguity.

(10) Hat der Mann oft die Frau Hans das Nilpferd füttern
has the man often the woman Hans the hippo feed
helfen lassen?
help let
'Did the man let the woman help Hans feed the hippo often?'

The examined sentences have a verbal complex with four ambiguous verbs. *Helfen* (*help*) can take a dative NP, or an infinitive with and without *zu* (*to*), it can form a verbal complex or take a VP complement. Furthermore, there is a lexical entry for the form of *helfen* that triggers auxiliary flip. The grammar contains lexical entries both for the auxiliary and the main verb reading of *hat* (*have*), and finally the case of the NPs in the example sentences is ambiguous in most of the sentences. For the purpose of comparison, I have added the simple sentence (11). Section 4 contains more examples.

(11) Ich erwarte, daß Karl der Frau das Buch gibt, die schläft.
I expect that Karl the woman the book gives who sleeps.
'I expect that Karl gives the book to the woman who sleeps.'

The times were measured on a 300 MHz Sparc Ultra 2 under Solaris 2.5.1 with SWI-Prolog 2.9.10.

[2] Bartsch and Vennemann (1972) and Kasper (1994) claimed that certain adverbial phrases have different scopings. However, previous work by Koster (1975) and Reis (1980) has demonstrated that the examples by Bartsch and Vennemann are cases of extraposition. See (Müller, 1999a, Ch. 11) for a more detailed discussion of the linearization of adverbs and scope.

Sentence	Words	entries	sec	edges
Hat John Cecilia Hans das Nilpferd füttern helfen lassen?	9	59	17	1389
Hat John die Frau Hans das Nilpferd füttern helfen lassen?	10	62	19	1445
Hat der Mann die Frau John das Nilpferd füttern helfen lassen?	11	69	10	848
Hat John oft Cecilia Hans das Nilpferd füttern helfen lassen?	10	60	123	4597
Hat John oft die Frau Hans das Nilpferd füttern helfen lassen?	11	63	130	4742
Hat der Mann oft die Frau Hans das Nilpferd füttern helfen lassen?	12	70	55	2687
Ich erwarte, daß Karl der Frau das Buch gibt, die schläft.	11	48	2	231

Sentence	Words	sec	edges
Hat John Cecilia Hans das Nilpferd füttern helfen lassen?	9	7	701
Hat John die Frau Hans das Nilpferd füttern helfen lassen?	10	8	697
Hat der Mann die Frau John das Nilpferd füttern helfen lassen?	11	5	501
Hat John oft Cecilia Hans das Nilpferd füttern helfen lassen?	10	35	2141
Hat John oft die Frau Hans das Nilpferd füttern helfen lassen?	11	32	1992
Hat der Mann oft die Frau Hans das Nilpferd füttern helfen lassen?	12	19	1302
Ich erwarte, daß Karl der Frau das Buch gibt, die schläft.	11	2	189

The introduction of the strong version of the Continuity Constraint reduces the number of passive edges by a factor of up to 2.4. The parse time is reduced by a factor of 4 for the most complex example.

Figure 1 shows the results of parsing 1,022 utterances form the Verbmobil CD 32 with and without the condition. The average parse time for the Verbmobil sentences was reduced by factor 3. The sentences were parsed with an edge limit of 25,000. As the figure shows, this edge limit is reached for 18 word sentences already without the restriction.[3]

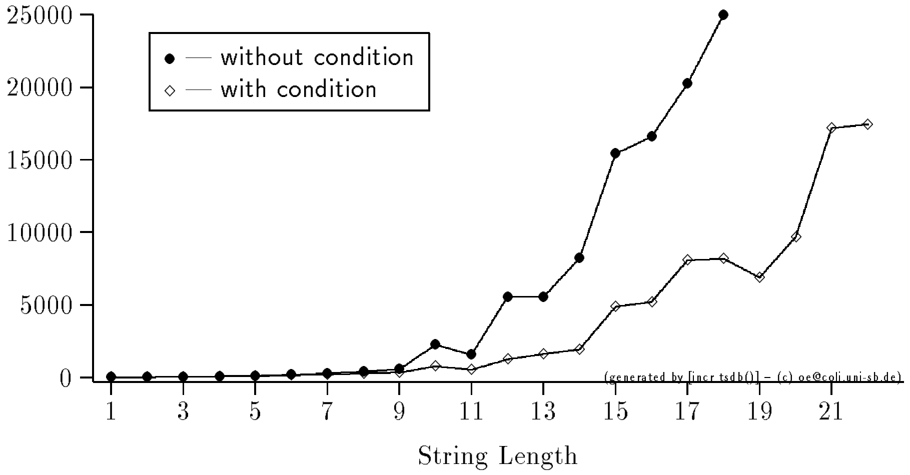

Figure 1: Passive edges generated by the grammar during a full parse

[3]The figure was created with TSDB++ by Stephan Oepen (Oepen and Flickinger, 1998).

4 Comparison of the Results

As far as I know, Kasper, Calcagno, and Davis (1998) (KCD) are the only ones that have published parse times of a system that processes a word order domain based HPSG grammar. In their paper they claim that parsing with discontinuous constituents has never been applied to HPSG grammars for nontrivial fragments. This is not true. The Babel-system has been available on the web[4] since 1995 and its existence is documented in several publications (See for instance (Müller, 1995; Müller, 1996a; Müller, 1996b)). The linguistic analyses are documented in a 486 page book (Müller, 1999a). The grammar covers a wide range of phenomena, a list of which can be found on the web.[5] The test suite that I use is also provided and can be used for comparison with other systems.

The actual grammar that is processed by the system is much bigger than the one described by KCD. The system currently has over 10,500 entries in the stem lexicon and over 240,000 full forms. I use 21 phrase structure rules. In order to make the systems comparable, I reduced the amount of phrase structure rules to such an extent that the reduced grammar covers those phenomena that the grammar described by KCD can handle. From the description of their grammar and from the knowledge about general properties of HPSG grammars it can be concluded that even this reduced Babel grammar has a broader coverage than the grammar KCD describe. For instance, my grammar can handle extraposition using Kathol and Pollard's (1995) approach; it also treats the German verbal complex correctly.

KCD state that they measured their parsing times on a SPARCstation 5. Paul Davis informed me that it was a 170 MHz machine. For the purpose of comparison I tested the sentences on a SPARCstation 5 with 110 MHz and extrapolated to the larger clock rate. I measured the times under Solaris 2.5 with SWI-Prolog 2.9.10.

The list on the next page shows the examples that were given by KCD.[6] I tested additional sentences because the word *schnell* is both an adjective and an adverb. The adjective *schnell* can appear with a predicative meaning like *roh* in (12).

(12) a. Er ißt das Fleisch roh.
he eats the meat raw

b. Er ißt das Fleisch nackt.
he eats the meat naked

Such predicative adjectives are serialized similar to adverbs. They can refer to the subject or to the direct object. In T7, for instance, *schnell* can refer to *er* or to *ihn*. To exclude these readings from the statistics, I have introduced the examples marked with an apostrophe. *Oft* is a true adverb and does not produce these readings. The results are shown in table 1. For Babel, I state the number of lexical entries and the number of passive edges in a parse.

Figure 2 shows the Babel results with the apostrophe examples, parsed with the full grammar. The examples without apostrophe are drawn with a dotted line.

[4] http://www.dfki.de/~stefan/Babel/e_index.html
[5] http://www.dfki.de/~stefan/Babel/e_phaenomene.html
[6] Due to space limitations I do not give a translation here.

T1	Er hilft ihr.
T2	Hilft er seiner Freundin?
T3	Er hilft ihr schnell.
T3'	Er hilft ihr oft.
T4	Hilft er ihr schnell?
T4'	Hilft er ihr oft?
T5	Ließ er ihr ihn helfen?
T6	Er ließ ihn ihr schnell helfen.
T6'	Er ließ ihn ihr oft helfen.
T7	Ließ er ihn ihr schnell helfen?
T7'	Ließ er ihn ihr oft helfen?
T8	Der Vater ließ seiner Freundin seinen Sohn helfen.
T9	Sie denkt, daß er ihr hilft.
T10	Sie denkt, daß er ihr schnell hilft.
T10'	Sie denkt, daß er ihr oft hilft.
T11	Sie denkt, daß er ihr ihn helfen ließ.
T12	Sie denkt, daß er seiner Freundin seinen Sohn helfen ließ.

Sent.	KCD98	Babel (reduced)			Babel (full)			Babel (full)
	time	time	lex. e.	edges	time	lex. e.	edges	time on Ultra 2, 300Mhz
T1	1.8	0.2	5	14	0.5	11	25	0.11
T2	3.6	0.2	5	17	0.7	10	29	0.16
T3	4.9	0.3	6	24	1.0	17	61	0.22
T3'		0.3	6	24	0.8	12	41	0.18
T4	5.2	0.3	6	29	1.9	17	73	0.25
T4'		0.3	6	29	0.9	12	49	0.20
T5	6.2	0.4	7	29	2.9	39	106	0.70
T6	32.4	0.8	8	64	7.7	45	374	1.66
T6'		0.8	8	64	4.5	40	226	1.90
T7	26.6	0.8	8	65	7.2	45	386	1.71
T7'		0.8	8	65	4.7	40	228	1.81
T8	52.1	0.6	10	29	3.5	48	113	0.82
T9	48.0	0.4	9	45	2.6	47	121	0.61
T10	253.8	0.5	10	45	3.5	53	176	0.85
T10'		0.5	10	45	3.6	48	136	0.70
T11	176.5	0.6	11	46	5.0	75	204	1.18
T12	528.1	0.7	12	49	5.2	76	207	1.25

Table 1: Parse times of T1–T12 in seconds, lexical entries and passive edges during a parse with Babel

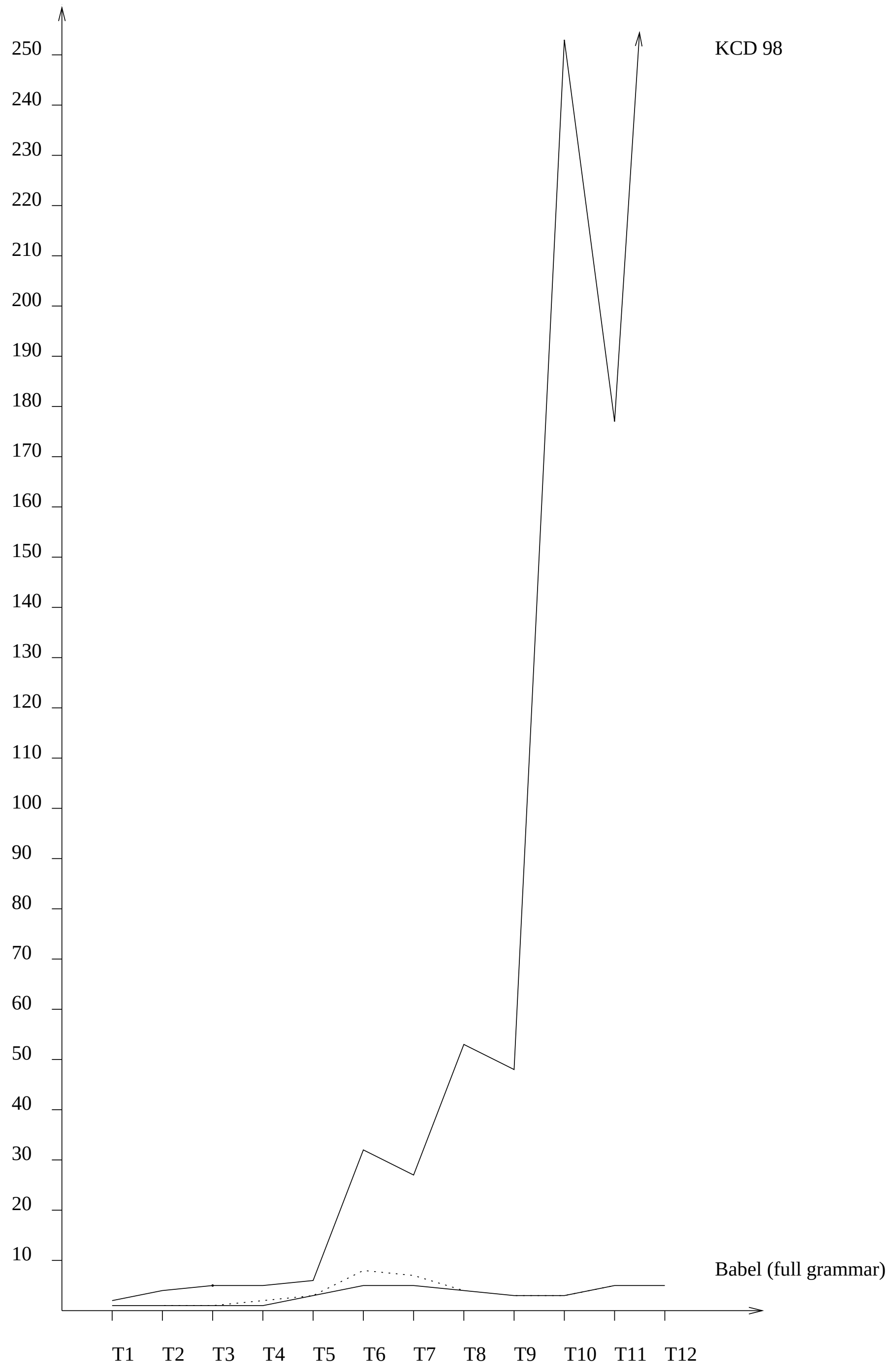

Figure 2: Comparison of Results

5 Conclusion

In this paper, I proposed a constraint on the continuity of maximal projections. In order to speed up the processing of input sentences, I strengthened the Continuity Constraint to restrict the number of discontinuous parts in a sentence. I implemented the strong constraint in the Babel system and demonstrated a reduction in parse times by up to a factor of 4.

I showed that the parse times of the Babel system rise only moderately as the length of the sentence increases. I compared Babel to the system presented in Kasper, Calcagno, and Davis (1998), and demonstrated that Babel can parse sentences up to 754 times faster.

References

Bartsch, Renate and Theo Vennemann. 1972. *Semantic Structures. A Study in the Relation between Semantics and Syntax*. Athenäum-Skripten Linguistik, volume 9. Frankfurt/Main: Athenäum.

Hinrichs, Erhard W. and Tsuneko Nakazawa. 1989. Subcategorization and VP structure in German. In *Aspects of German VP Structure*, SfS-Report-01-93. Eberhard-Karls-Universität Tübingen.

Kasper, Robert T. 1994. Adjuncts in the Mittelfeld. In John Nerbonne, Klaus Netter, and Carl J. Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, CSLI Lecture Notes, number 46. Center for the Study of Language and Information, Stanford, pages 39–70.

Kasper, Robert T., Mike Calcagno, and Paul C. Davis. 1998. Know when to hold 'em: Shuffling deterministically in a parser for nonconcatenative grammars. In *Proceedings of the COLING-ACL 98*, pages 663–669, Montreal.

Kathol, Andreas. 1995. *Linearization-Based German Syntax*. Ph.D. thesis, Ohio State University.

Kathol, Andreas and Carl J. Pollard. 1995. Extraposition via complex domain formation. In *Proceedings of the Thrirty-Third Annual Meeting of the ACL*, Boston. Association for Computational Linguistics. http://linguistics.berkeley.edu/~kathol/Papers/ACL95.ps.gz. 06.29.99.

Kiss, Tibor. 1995. *Infinite Komplementation. Neue Studien zum deutschen Verbum infinitum*. Linguistische Arbeiten, number 333. Tübingen: Max Niemeyer Verlag.

Koster, Jan. 1975. Dutch as an SOV language. *Linguistic Analysis*, 1(2):111–136.

Müller, Stefan. 1995. Scrambling in German – extraction into the *Mittelfeld*. In Benjamin K. T'sou and Tom B. Y. Lai, editors, *Proceedings of the tenth Pacific Asia Conference on Language, Information and Computation*, pages 79–83. City University of Hong Kong. http://www.dfki.de/~stefan/Pub/e_scrambling.html. 27.01.2000.

Müller, Stefan. 1996a. The Babel-System—an HPSG Prolog implementation. In *Proceedings of the Fourth International Conference on the Practical Application of Prolog*, pages 263–277, London. http://www.dfki.de/~stefan/Pub/e_babel.html. 27.01.2000.

Müller, Stefan. 1996b. Yet another paper about partial verb phrase fronting in German. In *Proceedings of COLING 96*, pages 800–805, Copenhagen. COLING 96. http://www.dfki.de/~stefan/Pub/e_pvp.html. 27.01.2000.

Müller, Stefan. 1999a. *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. Linguistische Arbeiten, number 394. Tübingen: Max Niemeyer Verlag. http://www.dfki.de/~stefan/Pub/e_hpsg.html. 27.01.2000.

Müller, Stefan. 1999b. Parsing of an HPSG grammar for German: Word order domains and discontinuous constituents. In Jost Gippert and Peter Olivier, editors, *Multilinguale Corpora. Codierung, Strukturierung, Analyse. 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung*, pages 292–303, Prag. enigma corporation. http://www.dfki.de/~stefan/Pub/e_restricting.html. 27.01.2000.

Müller, Stefan. 1999c. Restricting discontinuity. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium 1999 (NLPRS'99)*, pages 85–90, Peking.

Oepen, Stephan and Daniel P. Flickinger. 1998. Towards systematic grammar profiling. Test suite technology ten years after. *Journal of Computer Speech and Language*, 12(4):411–436. (Special Issue on Evaluation).

Pollard, Carl J., Robert T. Kasper, and Robert D. Levine. 1992. Linearization grammar. Research Proposal to the National Science Foundation.

Pollard, Carl J., Robert T. Kasper, and Robert D. Levine. 1994. Studies in constituent ordering: Toward a theory of linearization in Head-Driven Phrase Structure Grammar. Research Proposal to the National Science Foundation. ftp://csli-ftp.stanford.edu/linguistics/sag/linearization-prop.ps.gz. 06.29.99.

Reape, Mike. 1991. Word order variation in Germanic and parsing. DYANA Report Deliverable R1.1.C, University of Edinburgh.

Reape, Mike. 1992. *A Formal Theory of Word Order: A Case Study in West Germanic*. Ph.D. thesis, University of Edinburgh.

Reape, Mike. 1994. Domain union and word order variation in German. In John Nerbonne, Klaus Netter, and Carl J. Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, CSLI Lecture Notes, number 46. Center for the Study of Language and Information, Stanford, chapter 5, pages 151–198.

Reis, Marga. 1980. On justifying topological frames: 'positional field' and the order of nonverbal constituents in German. *Documentation et Recherche en Linguistique Allemande Contemporaine*, 22/23:59–85.