Self Embedded Relative Clauses in a Corpus of German Newspaper Texts

Christian Korthals
Computational Linguistics
University of the Saaarland
cnkortha@coli.uni-sb.de

Abstract

The distribution of center self-embeddings and extrapositions in German is assumed to reflect a universal performance strategy of minimizing memory load during parsing. Self-embedded relative clauses of embedding depth 2 were semi-automatically analysed in a treebank of German newspaper texts. Clause length and especially extraposition distance are found as the main distinctive parameters between center embeddings and extrapositions.[1]

1 Introduction

The opposition of center self-embedding constructions and extrapositions is an interesting phenomenon of syntax from two points of view: from the perspective of parsing and automata theory, center embeddings force the parser to at least possess context free power (Chomsky 1959). From a psycholinguistic perspective, center embeddings and extrapositions are assumed to differ in their memory load, because the processing of one phrase may have to be delayed until intervening material has been processed.

This paper examines German relative clause (RC) self-embeddings as an example of these kinds of constructions.[2] In section 2, we define and classify relative clauses and explain in which cases the choice between center embedding vs. extraposition arises in

[1]This work was funded within the NEGRA project, which is part of the *Sonderforschungsbereich* 378 of the *Deutsche Forschungsgemeinschaft* at the University of the Saarland. I wish to thank Thorsten Brants, Reinhard Köhler, Valia Kordoni, Lars Konieczny, Daniela Kurz, Oliver Plähn, Christoph Scheepers, Geert-Jan Kruijff, Kristina Striegnitz, Hans Uszkoreit and the anonymous reviewers for their comments and support.

[2]Note that the term "self-embedding" is used to refer to all cases where a RC (or, defining more weakly, also an S) contains another RC. We use the terms "center embedding", "right embedding" and "extraposition" to subclassify self-embeddings.

German syntax. Hypotheses on the acceptability of different constructions are formulated primarily on the basis of Hawkins' performance theory, which is introduced in section 3 together with other preliminaries. The hypotheses were tested against data from a treebank of German newspaper texts. This corpus is described in section 4 together with a fully automatic method of deriving description parameters of each relative clause in the corpus. In section 5, we present the results of the evaluation of three hypotheses: a) Longer relative clauses should tend to be extraposed, while shorter ones should tend to be center embedded. This turns out to be weakly supported only (Section 5.1). b) The potential distance between relative pronoun and antecedent should decide whether a clause is extraposed or not (Section 5.2) or not. This factor will turn out to be the most relevant one. c) Embedding and embedded relative clause should differ in their internal structure. This turns out to be not supported. (Section 5.3).

2 Types of Relative Clauses

A relative clause is a sub-clause that contains at least one relative pronoun. The TIGER annotation scheme (Brants et al. 1997) and the Stuttgart Tübingen tagset (Schiller et al. 1999) classify relative pronouns into four classes based on two criteria: substituting ("the man *who* ...") vs. attributive ("the man *whose* ..."), and proper relative pronoun vs. wh-pronoun ("places *where* ...", "*what* is surprising is ..."). Especially when wh-pronouns are involved, there is difficulty in distinguishing between relative clauses and other clause types (e.g. place holder phrase[3], clausal complement). If there is no antecedent NP, one speaks of a *free relative clause* (Eisenberg 1999):269.

In German, main declarative clauses in the present or past tense have verb second position. If a post-verbal NP contains a relative clause, the relative clause may turn out to be situated at the end of the sentence, directly following its antecedent NP. Such a case of *right embedding* (RE) yields no processing difficulty at all, as in example (1).

(1) *Er kannte den Termin, der für die Konferenz festgesetzt war.*
He knew the deadline which for the conference set was.
"He knew the deadline set for the conference."

In contrast, perfect and future tenses are realized as auxiliary constructions with an auxiliary finite verb at second position and a non-finite verb in final position. Additionally, also subordinate clauses (and among them relative clauses) have the (finite) verb in final position. The kind of syntactic constructions which can occur after this "final" verb is heavily restricted (material after the final verb is analysed as residing in the *Nachfeld*, or *extraposition field* in a topological analysis of German, cf. (Eisenberg 1999):388). This

[3]The annotation scheme allows for an RC analysis or a place holder phrase analysis in cases as "*Das* kaufen, *was* es gibt" (to buy what is available)

means that whenever a relative clause is embedded in another relative clause, at the very least the verb of the superordinated relative clause must follow after the subordinated relative clause, if the NP including the RC is to be realized continuously, i.e. without intervening material. These cases will be called *center embeddings* (CE). (2) is a complex example, where a relative clause is center embedded in another relative clause, which is again center embedded in the main clause. If center embedding is to be avoided, the subordinated relative clause must be extraposed over the verb (and maybe some of its arguments). This case is illustrated in example (3) and will be called an *extraposition* (EX).

However, it is possible to extrapose even more material than the relative clause alone, as illustrated in example (4). This again results in a continuous realization, and thus is considered a type of RE, too.

(2) [$_S$ *Er hat den Termin*$_{\mathsf{NN}}$, [$_{RC}$ *der*$_{\mathsf{PREL}}$ *für die Konferenz*$_{\mathsf{NN}}$, [$_{RC}$ *die*$_{\mathsf{PREL}}$ *er*
He has the deadline which for the conference which he
besuchen$_{\mathsf{VINF}}$ *wollte*$_{\mathsf{VFIN}}$,] *festgesetzt*$_{\mathsf{VPART}}$ *war*$_{\mathsf{VFIN}}$,] *gekannt*$_{\mathsf{VFIN}}$.]
attend wanted set was known
"He knew the deadline that was set for the conference he wanted to attend."

(3) [$_S$Er hat den Termin, [$_{RC}$der für die Konferenz festgesetzt war,] [$_{RC}$die er besuchen wollte,] gekannt.]

(4) [$_S$Er hat den Termin, [$_{RC}$der festgesetzt war für die Konferenz, [$_{RC}$die er besuchen wollte,]] gekannt.]

Throughout the paper, we will investigate relative clauses of embedding depth 2, that is the most inner RCs in sentences of the structure $[S\ \alpha\ [RC\ \beta\ [RC]\ \gamma]\ \delta]$. Note that the dominance relation between the clauses will in general be indirect: Normally, at least an NP, the antecedent NP of the RC will occur in α or β; there may also be more material, e.g. coordinated structures or VPs.

3 Theoretical Assumptions

Standard context free phrase structure grammars would indeed generate center self-emedded structures like (2) and render them grammatical. It was early noted, though, e.g. by Chomsky (1959), that humans have difficulty understanding center embedded structures. The existence of alternative syntactic constructions like right extrapositions can therefore be viewed as emerging from a universal principle of preventing center embeddings (Köhler 1999) or restricting them to a certain depth.

Many alternative explanations have been provided to explain humans' parsing difficulties with center embedding. Lewis (1996) discusses some of them. Center embedded relative clauses are predicted to be hard in his model due to memory interference effects

between stacked preverbal NPs that cannot be integrated into a coherent parse tree before the verb is encountered, and therefore have to be stored in short term memory. He predicts center embeddings are incomprehensible only in those cases where more than two NPs before the occurrence of the verb receive the same grammatical function.

In contrast, Hawkins (1994) does not consider the internal structure of embedded material at all but reduces all the factors that may contribute to difficulties in understanding to one simple variable, namely phrase length in terms of the number of word forms. The quality of a preterminal node is simply the number of terminal nodes from the beginning of the phrase to the first terminal of the last immediate constituent, divided by the number of immediate constituent nodes of the phrase. The overall quality of the sentence is the sum of the quality of its nodes. For relative clauses, Hawkins's principle of early immediate constituents (EIC) predicts that a) shorter center embedded relative clauses should be more acceptable than longer ones and b) that extrapositions should be more acceptable if their extraposition distance is small. Thus, a) and b) are competing principles of locality that predict a high memory load when the processing of a phrase has to be delayed until intervening material has been processed.

4 Data and Methods

In order to test hypotheses on complex syntactic constructions, syntactically annotated corpora (treebanks) are needed. The NEGRA Millennium corpus (Negra) is a treebank comprising 20,571 sentences semi-manually compiled at the University of the Saarland. Word forms are POS-tagged according to the Stuttgart Tübingen tagset (Schiller et al. 1999). The syntactic structure is annotated according to an annotation scheme that recognizes formal node tags (phrase labels like *sentence*, *noun phrase*) and functional edge labels (like *subject*, *post-modification*, *relative clause*, *clausal complement*). The annotation scheme allows for crossing phrase-structure edges to encode discontinouous constituents. The Negra corpus is a sub-corpus of the 2.4 million sentence *Frankfurter Rundschau* newspaper corpus (Rundschau), which was statistically POS tagged for the course of this study.

Since center embedded relative clauses are quite a rare phenomenon, the data from Negra was not sufficient. Therefore, we heuristically searched the *Frankfurter Rundschau* corpus for possible center embeddings and annotated the findings according to the Negra annotation scheme. The heuristics searched for at least two occurrences of commas followed by an optional preposition and a potential relative pronoun in the sentence. Since word forms that are potential relative pronouns are highly ambiguous between determiner and relative pronoun in German, the statistical tagger used did not yield reliable results, and manual filtering was necessary. At the present point of time, we cannot yet estimate recall and precision accurately, but it is clear that further work is necessary to increase the recall.

The resulting data was automatically analysed with C, Perl and Awk tools operating on the data structures representing the syntactic trees. For each sentence, a characterization

of all the relative clauses it contained was generated fully automatically. The parameters were then filtered and statistically evaluated. Among the parameters were the **embedding depth** within a superordinated relative clause, the **relative pronoun** together with a detailed description of its tag and its function within the clause, the kind of **antecedent phrase** (NP, PP or *free relative clause*), the kind of **embedding** (RE, CE, EX) the **length** of superordinated and subordinated clause and the **(potential) extraposition distance** (cf. section 5.2).

5 Results

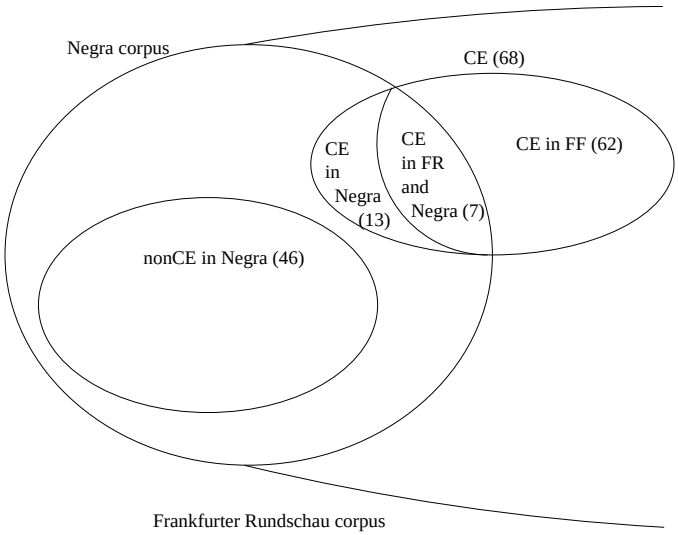

Figure 1: Relations between corpora used

Of 20,571 sentences in Negra, 2352 (11.4%) contained at least one relative clause, and there were 2389 relative clauses in total. 61 sentences (0.3%) contained a relative clause embedding, i.e. two or more relative clauses of which at least two were embedded into each other. 1 of these sentences contained a double RC self-embedding, 1 sentence contained a conjunction of two embedded RCs clauses within another RC. These two instances will be ignored in the rest of the paper. Of the remaining 59 embedded relative clauses, 39 (66%) were extraposed from their embedding RC, 13 (22%) were center embedded, and 7 (12%) were "right embedded", i.e. involved constructions where more material than the relative clause was extraposed.

Table 1 shows the distribution in detail, together with the percantage of subordinated relative clauses extraposed, center embedded and right embedded within their class.

count	RC1 under matrix	RC2 under RC1	
25	RE	EX	69%
7	RE	CE	19%
4	RE	RE	11%
11	EX	EX	69%
3	EX	CE	19%
2	EX	RE	12%
3	CE	EX	43%
3	CE	CE	43%
1	CE	RE	14%
59			

Table 1: Sentences in Negra containing a relative clause self-embedding

The heuristical search in the unannotated *Frankfurter Rundschau* corpus yielded 62 center embedded RCs. 14 of these were CE in CE embeddings, 28 CE in RE embeddings and 17 CE in EX embeddings. There were 3 more complex constructions involving CE relative clauses.

A corpus of all CE relative clauses was compiled by merging the CE relative clauses in Negra with those from *Frankfurter Rundschau*, which yielded 68 CE relative clauses in total. Since we do not yet have more data from the *Rundschau* corpus, we ignored the type of embedding of the self-embedded RC within the top level S for the compilation of the corpus of center embeddings.

In analogy, a corpus of non center embedded RCs (nonCE) was computed. It included the 40 extrapositions and the 7 right embeddings from Negra. Figure 1 shows the structure of the data.

5.1 Clause Length

Based on the principle of Early Immediate Constituets (EIC, Hawkins (1994)), one would expect to find a preference for the extraposition of longer phrases, while shorter phrases we would expect to appear center embedded more often, because they would not delay the processing of their matrix clause for very long.

As a comparison, we computed the overall distribution of relative clause length in the Negra corpus, i.e. the length of all 2389 relative clauses in the Negra corpus, be they embedded in other clauses or not. Within Negra, we find an average relative clause length of 10.2. We compared these results to all 68 center embedded relative clauses in the corpus of center embeddings (CE sub-corpus in figure 1) as well as with all 46 non-center embedded relative clauses from Negra (nonCE sub-corpus). Note that extrapositions and "right embeddings" were treated as one class on the basis of the arguments under section

2.

Table 2 shows the frequencies of relative clauses of different length in the corpus of center embeddings (CE) and in the corpus of extrapositions and right embeddings (nonCE) together with the overall distribution of relative clause length in Negra.

length	2	3	4	5–6	7–10	11–14	15–24	25–47	Total	Average
Negra	.00	2.2	5.0	19.3	36.7	20.4	13.3	2.9	2389	10.2
nonCE	0	2.1	2.1	23.4	38.3	17.0	17.0	0	46	9.4
CE	0	7.4	10.3	35.3	25.0	20.6	1.5	0	68	7.5

Table 2: Distribution of relative clause length for all RCs in Negra, center embedded RCs and non center embedded RCs

There is no significant difference between the distribution of non-center embedded relative clauses we investigated and the overall distribution in Negra ($\chi^2_{(df=7)} = 5.09$). That is, extraposed or right embedded relative clauses are quite representative in their length of relative clauses in total. However, there is a significant difference between the distributions of center embedded relative clauses and the distribution in Negra ($\chi^2_{(df=10)} = 34.0$; $p = 0.001$). There is also a significant difference between the distribution of CE and nonCE ($\chi^2_{(df=6)} = 68.9$; $p = 0.001$).

We conclude that a subordinated relative clause in a relative clause extraposition is not significantly shorter than the average of relative clauses in total, namely around 10 word forms. In contrast, a subordinated relative clause center embedded within another relative clause is significantly shorter, namely about 8 word forms.

5.2 Extraposition Distance

One of the parameters that was fully automatically generated from the corpus data is the *(potential) extraposition distance* of a relative clause. In the case of an extraposition, this is the number of word forms between the end of the annotated antecedent NP and the relative clause (ex). In example (3) above, e.g., the RC is extraposed over the two word forms *festgesetzt war* ($ex = 2$). In the case of a center embedding it is the number of remaining word forms of the superordinated clause after the end of the subordinated clause (rd). Note that potentially every center embedding can be transformed into an extraposition and vice versa, and that always $rd = ex$, as in example (2).

A hypothesis following from Hawkins (1994) is that extrapositions over a small number of word forms are more acceptable than those over a far distance.

There is striking evidence for this hypothesis. In fact, extrapositions over a distance of more than three word forms do not occur in subordinated embedded relative clauses at all. In contrast, there is massive data for center embeddings whose potential extraposition

distance is between 4 and 17. Table 3 shows the distributions in detail. Note that RE embeddings were ignored in this analysis.

distance	1	2	3	4	5	6	7	8	9	10–13	14-17	total
CE rd	11	13	8	12	6	6	2	2	2	4	2	68
EX ex	24	11	4									39

Table 3: Distribution of (potential) extraposition distance of center embedded and extraposed embedded relative clauses

As can be seen from the table, relative clause embeddings with a potential extraposition distance between 1 and 3 do occur as well center embedded as extraposed. If center embedding is assumed to be the preferred option for relative clauses with a potential extraposition distance greater than 3, but center embeddings and extrapositions are both possible for RCs with small potential extraposition distance, one would expect length differences not only between EX and CE but also between "close" CEs and "far" CEs.

The analysis confirmed this hypothesis. While extrapositions are 9.4 word forms long on average (see above, table 2), "far" center embeddings (whose potential extraposition distance rd is greater than 3) are 8.6 word forms long, and "close" center embeddings ($rd \leq 3$) are 5.7 word forms long (The χ^2 values reached $p = 0.001$ for EX vs. close CE, $p = 0.02$ for EX vs. far CE, and $p = 0.004$ for far CE vs. close CE).

The data suggests that not all center embeddings lead to problems, but that especially those center embeddings are acceptable that would lead to large extraposition distances, if the relative clause would be extraposed. An example is (5). Sentences with a small (potential) extraposition distance, on the other hand, occur center embedded as well as extraposed. Examples of both cases are provided in (6) and (7).

(5) [$_S$ *Der Regierungschef*$_{\mathsf{NN}}$, [$_{RC}$ *der*$_{\mathsf{PREL}}$ *sich wegen Pensions- und*
The head of government who himself due to pensions and
Ausgleichszahlungen$_{\mathsf{NN}}$, [$_{RC}$ *die*$_{\mathsf{PREL}}$ *sein Gehalt jahrelang monatlich um einige tausend*
compensations which his salary for years monthly by some thousand
Mark aufstockten$_{\mathsf{VFIN}}$], *seit zwei Wochen unangenehmen Fragen stellen*$_{\mathsf{VINF}}$ *muß*$_{\mathsf{VFIN}}$],
marks increased for two weaks embarrassing questions answer must
forderte$_{\mathsf{VFIN}}$ *den Jubel und die unkritische Solidarität seiner Genossen im*
called for the chearing and the uncritical solidarity of his comrades in
Oskar-Land ein].
Oskar land (verb particle)
"The head of government, who has been exposed for two weaks to anwering embarrassing questions about pensions and compensation money that have been increasing his monthly salary by some thousand marks for years called for his comrades' applause and uncritical solidarity in the 'Oskar' territory." *(far CE under far CE)*

(6) [$_S$ *Im selben Zuge soll der Rabatt*$_{\mathsf{NN}}$ *von 20 Pfennig pro Kubikmeter,* [$_{RC}$ *der*$_{\mathsf{PREL}}$
By the same token shall the reduction of 20 Pfenning per cubic metre which

bisher *denjenigen* *Großverbrauchern*$_{\mathsf{NN}}$, [$_{RC}$ *die*$_{\mathsf{PREL}}$ *mehr* *als* *10,001* *Kubikmeter*
until now those large consumers who more than 10,001 cubic metres
im *Jahr* *verbrauchen*$_{\mathsf{VFIN}}$], *zugestanden*$_{\mathsf{VPART}}$ *wurde*$_{\mathsf{VFIN}}$], *entfallen*$_{\mathsf{VFIN}}$].
per year consume allowed was be abolished

"By the same token, the reduction of 20 Pfennig per cubic metre which was granted to those large customers who consume more than 10,001 cubic metres a year, is to be abolished." *(close CE under close CE)*

(7) [$_{S}$ *Dieser* *Fonds* *wird*$_{\mathsf{VFIN}}$ *beispielsweise* *mit* *dem* *Flaschenpfand*$_{\mathsf{NN}}$ *gefüllt*$_{\mathsf{VPART}}$],
This fund is for example with the bottle deposit filled
[$_{RC}$ *das*$_{\mathsf{PREL}}$ *Annemarie* *Roth* *für* *die* *Flaschen*$_{\mathsf{NN}}$ *erlöst*], [$_{RC}$ *die*$_{\mathsf{PREL}}$ *sie* *bei* *den*
which Annemarie Roth for the bottles gets which she during her
regelmäßigen *Spaziergängen* *im* *Grüneburgpark* *findet*].
frequent walks in the Grüneburg park finds

"This fund will be financed from the deposit which Annemarie Roth gets back for the bottlse she finds during her frequent walks in the Grüneburg park." *(close EX under close EX)*

5.3 Structural Parallelism or Anti-Parallelism

If more complex categories than S or NP are assumed for syntactic descriptions, the concept of self-embedding becomes unclear: there might be differences between an object RC embedding a subject RC, for example. Also Lewis (1996) would predict differences in acceptability depending on the noun phrases involved. As a matter of fact, Lewis (1996) predicts an influence of *all* noun phrases occurring within the matrix clause and all embedded clauses. For reasons of simplicity, we concentrated on two factors only: the antecedent NPs and the phrase containing the relative pronoun. We considered these variables as possible criteria of structure similarity of the clauses involved and hypothesized there might be a trend towards anti-parallelism. There is still not enough data for more variables to be investigated.

Within the Negra corpus, there is a ratio of roughly 70% antecedent NPs and 30% antecedent PPs of all relative clauses in total. Unexpectedly, embedded relative clauses have 57% antecedent NPs and 43% antecedent PPs, be they center embedded or not ($\chi^2_{(df=1)} = 9.34$; $p = 0.002$). There is no significant difference between CE and nonCE.

The hypothesis that the distribution of the antecedent NP of the subordinated RC might depend on the antecedent NP of the superordinated RC is not supported: There are four possible classes (RC with NP or PP antecedent embedded within RC with NP or PP antecedent), whose distribution does not differ significantly between CE and nonCE.

Our second criterion of internal clause structure was the function of the relative pronoun within the RC. Again, the distribution of RC pronoun functions within the entire Negra corpus was computed as a comparison.

Table 4 shows an abbreviated characterisation of the pronoun function together with the original Negra tags for reference. Note that subject RCs are by far the most frequent class of relative clauses in German, and that object RCs are quite rare.

63%	substituting pronoun as subject (SB:PRELS ...)
11%	substituting pronoun as accusative object (OA:PRELS ...)
1.4%	substituting pronoun as dative object (DA:PRELS ...)
14%	substituting pronoun within adverbial PP (NK:PRELS MO:PP ...)
1.7%	attributive pronoun within subject NP (NK:PRELAT SB:NP ...)
1.7%	attributive pronoun within other non-subject phrase (NK:PRELAT ...)
4.5%	adverbial wh-pronoun (MO:PWAV ...)
1.2%	substituting wh-pronoun as subject in free RC (SB:PWS ...)

Table 4: Function of relative pronoun within all relative clauses in Negra

Significance tests were possible for the main classes *subject relative, object relative, others* only. There is only a marginal difference in the distribution of these classes between nonCE and Negra ($\chi^2_{(df=2)} = 5.1;\ p = 0.07$). In contrast, there are significant differences between the distributions of CE and of Negra ($\chi^2_{(df=2)} = 21.9;\ p = 0.0001$), and between nonCE and CE ($\chi^2_{(df=2)} = 43;\ p = 0.0001$). Center embedded RCs were significantly more often subject RCs (74%) than nonCEs (45%) or RCs in general (58%).

While center embedded RCs may have a tendency to be subject RCs in general, it could be shown that it does not have any influence on the type of subordinated relative clause whether the superordinated relative clause is a subject RC, an object RC or any other type of RC. The matrix clauses of subject RCs are also subject RCs in 63% of the cases and others in the remaining 37%. This is true of center embedded RCs as well as of non-center embedded RCs ($p < 0.001$).

Recall, however, that center embeddings were mainly taken from the *Frankfurter Rundschau* corpus to enhance data, while nonCE are solely from Negra. Since the material was heuristically extracted from the *Frankfurter Rundschau*, and the recall was low, it may well be argued that the difference is due to properties of the extraction heuristics.

In summary, the hypotheses that there might be a dependency between the internal structure of the superodinated RC and the subordinated RC could be falsified. For both variables considered, antecedent and relative pronoun function, such a dependency was not significant. There was neither a tendency towards anti-parallelism between embedding RC and embedded RC, nor towards parallelism. There were two unpredicted effects: an affinity of embedded (center embedded and extraposed) RCs to have PP antecedents, and a tendency of center embedded relative clauses to be subject RCs. These effects may well be due to our extraction heuristics.

6 Discussion

The data is compatible with models that explain the distribution of center embeddings and extrapositions as following from restrictions on short term memory. Two parameters,

embedded clause length and extraposition distance that are predicted to be relevant by Hawkins's EIC principle were confirmed. However, extraposition distance seems to be the more prominent factor, whereas clause length is of secondary importance. The data may also be interpretable within related models, such as Gibson's SPLT theory (Gibson 1998), which assumes higher "integration cost" with increasing extraposition distance and/or clause length.

Predictions following from a phrase structure architecture assuming that the internal clause structure of the participating RCs matters could be falsified. Assumptions that parallelism or anti-parallelism between the RCs could simplify processing could not be confirmed with the present data. In order to test predictions of a more elaborated interference model like Lewis (1996), more data is needed.

The question may be raised whether the material investigated – newspaper texts – reflects production or perception preferences. Konieczny (2000), for instance, found significant acceptability asymetries between perception and production data.

7 Conclusion and Future Work

The paper demonstrated how a syntactically annotated corpus together with appropriate querying and processing tools can be used to semi-automatically prove or disprove hypotheses derived from a model of performance. This technique allows for an easy repetition of the study on different data or with additional parameters.

The data can be interpreted as following from a tradeoff between different principles of minimizing memory effort during parsing (or generation, respectively). Such principles are the principle of Early Immediate Constituents and constraints on embedding depth and phrase length (see e.g. Köhler (1999) for a wider framework). The main results are the quantification of the tendency to extrapose over short distances only, and to center embed especially shorter phrases whose potential extraposition distance is high. We also found that self-embedded relative clauses do not seem to deviate in their structure from the average of relative clauses.

The study confirms and supplements a prior study by Uszkoreit et al. (1998) with a new method (automatical parameter generation), more data (20,000 instead of 12,000 sentences) and, especially, a shift towards RC self-embeddings instead of simple RC in S embeddings. For the future, an automatic calculation of EIC scores is planned. A cross-check with functional equivalents of relative clauses (Köhler 1999) and psycholinguistic acceptability tests are two possible ways to supplement the study. We are confident to be able to increase the recall from the *Rundschau* corpus to enhance the evaluation at those points where data sparseness prohibited significant statements.

References

Brants, T., R. Hendriks, S. Kramp, B. Krenn, C. Preis, W. Skut, and H. Uszkoreit (1997). Das NEGRA-Annotationsschema. Negra Project Report, Universität des Saarlandes, Computerlinguistik, Saarbrücken, Germany.

Chomsky, N. (1959). On Certain Formal Properties of Grammars. *Information and Control* (2), 137–167.

Eisenberg, P. (1999). *Grundriß der deutschen Grammatik*, Volume 2. Der Satz. Stuttgart: Metzler.

Gibson, E. (1998). Linguistic Complexity: Locality of Syntactic Dependencies. *Cognition* (68), 1–76.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge UP.

Köhler, R. (1999). Syntactic Structures: Properties and Interrelations. *Journal of Quantitative Linguistics 6*(1), 46–57.

Konieczny, L. (2000). Locality and Parsing Complexity. *Journal of Psycholinguistic Research 29*(6), 627–645.

Lewis, R. L. (1996). Interference in Short-term Memory: The Magical Number Two (or Three) in Sentence Processing. *Journal of Psycholinguistic Research 25*, 93–115.

Schiller, A., S. Teufel, C. Stöckert, and C. Thielen (1999). Guidelines für das Tagging: Kleines und großes Tagset. Stuttgart: Institut für maschinelle Sprachverarbeitung.

Uszkoreit, H., T. Brants, D. Duchier, B. Krenn, L. Konieczny, S. Oepen, and W. Skut (1998). Studien zur performanzorientierten Linguistik. Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft 7*(3), 129–133.