Enriching a Treebank to Investigate Relative Clause Extraposition in German

Jan Strunk

Sprachwissenschaftliches Institut, Ruhr-Universität Bochum
44780 Bochum, Germany
strunk@linguistics.rub.de

Abstract

I describe the construction of a corpus for research on relative clause extraposition in German based on the treebank TüBa-D/Z. I also define an annotation scheme for the relations between relative clauses and their antecedents which is added as a second annotation level to the syntactic trees. This additional annotation level allows for a direct representation of the relevant parts of the relative construction and also serves as a locus for the annotation of additional features which are partly automatically derived from the underlying treebank and partly added manually. Finally, I also report on the results of two pilot studies using this enriched treebank. The first study tests claims made in the theoretical literature on relative clause extraposition with regard to syntactic locality, definiteness, and restrictiveness. It shows that although the theoretical claims often go in the right direction, they go too far by positing categorical constraints that are not supported by the corpus data and thus underestimate the complexity of the data. The second pilot study goes one step in the direction of taking this complexity into account by demonstrating the potential of the enriched treebank for building a multivariate model of relative clause extraposition as a syntactic alternation.

1. Introduction

Relative clauses in German and other languages are normally realized *integrated* in the noun phrase that they modify but they can also be separated from their antecedent by intervening material and occur further to the right in *extraposed* position, mostly at the end of the matrix clause; cf. examples (1) and (2). It is usually assumed that the integrated and extraposed variants of relative clauses are semantically equivalent.

(1) *Ich habe [$_{NP}$ alle diesbezüglichen Threads*
I have all related threads
[$_{RC}$ die ich finden konnte]] gelesen.
that I find could read
"I have read all the related threads that I could find."

(2) *Ich habe [$_{NP}$ alle Bücher _] gelesen*
I have all books read
[$_{RC}$ die ich finden konnte.]
that I find could
"I have read all books that I could find."

Relative clause extraposition has mostly been studied within generative grammar using introspective data; cf. e.g. Baltin (2006). Although a few corpus studies have also been published (Shannon, 1992; Uszkoreit et al., 1998; Hawkins, 2004; Francis, 2010), they have mostly concentrated on individual factors and have not tried to account for relative clause extraposition as a syntactic alternation using an integrated (statistical) model – as proposed for example for the English dative alternation by Bresnan et al. (2007).
In this paper, I describe the construction of an enriched German treebank that will allow a more systematic investigation of the following three questions:

1. **Empirical investigation of constraints on relative clause extraposition:** What are the constraints on relative clause extraposition? Are constraints posited in the theoretical literature compatible with corpus data?
2. **Modeling relative clause extraposition as a syntactic alternation:** What factors influence the decision of the speaker to realize a relative clause as integrated or extraposed and how can they be combined in an integrated model?
3. **Disambiguation of the attachment of relative clauses:** What is the degree of attachment ambiguity of integrated and extraposed relative clauses? How is the antecedent of a relative clause identified by the hearer and how can the same be achieved using an automatic disambiguation algorithm?

The remainder of the paper is structured as follows: Section 2. describes the construction of the corpus, starting with the choice of the underlying treebank in section 2.1. and continuing with the additional automatic and manual annotation steps in section 2.2.. In section 3., I report on two pilot studies exemplifying the first two of the above-mentioned intended uses of the enriched treebank: In section 3.1., I test constraints with regard to syntactic locality, definiteness, and restrictiveness that have been proposed in the syntactic literature. In section 3.2., I build a multivariate statistical model including these and additional factors in order to predict whether a relative clause will be extraposed or not. The paper concludes with a summary in section 4..

2. Enriching the treebank

2.1. Original treebank

As a basis for the construction of the corpus of German relative clauses, I chose the third release of the *Tübingen Treebank of Written German* (TüBa-D/Z) (Telljohann et al., 2005),[1] which contains articles from the German newspaper *taz* and comprises 27,125 sentences and 473,747 tokens in all. The TüBa-D/Z treebank is part-of-speech tagged and also provides information about inflectional morphology.

[1] http://arbuckle.sfs.uni-tuebingen.de/en_tuebadz.shtml

The syntactic analysis is based on the traditional topological theory of German clause structure in combination with a shallow analysis of constituency structure containing only maximal projections. TüBa-D/Z also labels the grammatical function of phrases and provides coreference relations.
After converting the treebank to TigerXML format, I used TigerSearch (Lezius, 2002) to extract the subset of sentences containing one or more relative clauses (identified by the label R-SIMPX). This yielded a corpus of 2,603 sentences containing 2,789 relative clauses.[2]
TüBa-D/Z forms an ideal basis for my corpus of relative clauses because it identifies them with an unambiguous label and allows for the easy identification of their antecedents and also their position (*integrated* vs. *extraposed*). Moreover, the wealth of morphological, syntactic, and functional features it contains are all potentially relevant for research on relative clause extraposition.

2.2. Adding a specialized secondary annotation layer

In order to facilitate the manual inspection and annotation of the corpus as well as the extraction of feature values for statistical analysis I decided to add a secondary annotation level to mark and connect all the relevant subparts of a relative construction. The individual elements of this additional structural annotation also serve as locus for the annotation of relevant features.
Instead of reinventing the wheel, I adopted the annotation tool SALTO[3] which was "originally developed for the annotation of semantic roles in the frame semantics paradigm" (Burchardt et al., 2006). SALTO is able to complement a syntactic tree in TigerXML format with an additional annotation level in which relations are modeled with so-called *frames*, directed acyclic graphs of depth one.
The scheme that I propose for the annotation of relative constructions is exemplified in figure 1 showing the annotation of the example sentence in (3).

(3) *Es gibt* $[_{NP}$ *einen neuen Kuli* __] *auf*
there is a new pen on
dem Markt, $[_{RC}$ *der heute schon als*
the market that today already as
Rarität zu bezeichnen ist.]
rarity to call is
"There is a new pen on the market that already has to be regarded as a rarity today."

It consists of several different frames that model the relative construction, i.e., the relation between a relative clause and its antecedent, and their respective subparts. The relative clause itself is modeled by the *RelClause* frame. It is anchored to the relative pronoun and its element *RelClause* points to the R-SIMPX node of the relative clause in the original constituent structure tree. Another element called *ExtraP* refers to the left-extracted phrase at the beginning of the relative clause. It is introduced because the grammatical function of the head noun within the relative clause, the presence of pied-piping, etc. can be established by looking at this phrase. The antecedent of the relative clause is modeled using the *Antecedent* frame which is anchored to the corresponding phrase in the syntactic tree, for example, the NP containing the head noun of the relative clause. Its element *Det* points to the determiner closing off that phrase if there is one. Information about the presence and type of a determiner allows for the derivation of a definiteness feature for the antecedent and is also interesting because some determiners like the special demonstrative *derjenige* ("that one") seem to function as a cataphor indicating the presence of a following relative clause. The second element *Nominal* refers to the antecedent NP excluding the determiner. It is introduced because it corresponds to the scope of restrictive relative clauses – as argued e.g. in Kiss (2005). It points to another frame also called *Nominal* which provides the opportunity to distinguish between the head noun itself and its modifiers using the frame elements *Head* and *Attribute*, respectively. The identity of the head noun may be useful to rank possible antecedents in attachment disambiguation, as are of course its morphosyntactic features since relative pronouns have to agree with the head noun in number and gender in German. The presence or absence of further modifiers in addition to a relative clause may likewise be useful for attachment disambiguation or might even influence the likelihood of extraposition. The *RelConstruction* frame, finally, connects the relative clause to its antecedent. In fact, since the corpus is also supposed to be useful for studying attachment disambiguation, all NPs within the matrix clause are annotated with *Antecedent* frames but only the actual antecedent is connected to the relative construction with the *Antecedent* element of the *RelConstruction* frame. Potential alternative antecedents are connected to the *RelConstruction* with the *AlternativeSyn* or *AlternativeSem* elements depending on whether the alternative antecedent is only morphosyntactically or also semantically compatible with the relative clause.
The structural and relational annotation of relative constructions just described is derived automatically from the original annotation of the treebank as follows: First, relative clauses are identified by searching for phrases of the category R-SIMPX. The extracted phrase (*ExtraP*) inside the relative clause is always directly dominated by the topological field node *C*. The relative pronoun is identified by searching for the word with the appropriate part-of-speech within the extracted phrase. These elements are then connected using a *RelClause* frame. Second, the matrix clause of all relative clauses is identified as the next higher sentential node in the syntactic structure and all nominal phrases within that clause are annotated with *Antecedent* frames as potential antecedents. Determiners can be identified as words of appropriate category that are immediately dominated by a nominal phrase. If an NP contains more than a determiner, the remainder is modeled using a *Nominal* frame. The head noun can be identified as the noun inside the NP that has the functional edge label HD. All other phrases dominated by an NP are connected to the *Nominal* frame with *Attribute* pointers. Finally, a *RelConstruction* frame is introduced that connects the relative clause to its actual antecedent. In TüBa-D/Z, the actual antecedent of

[2] If this corpus needed to be expanded later, one could add sentences from more recent releases of TüBa-D/Z as well as other German corpora such as the NEGRA and TIGER corpora.
[3] http://www.coli.uni-saarland.de/projects/salsa/

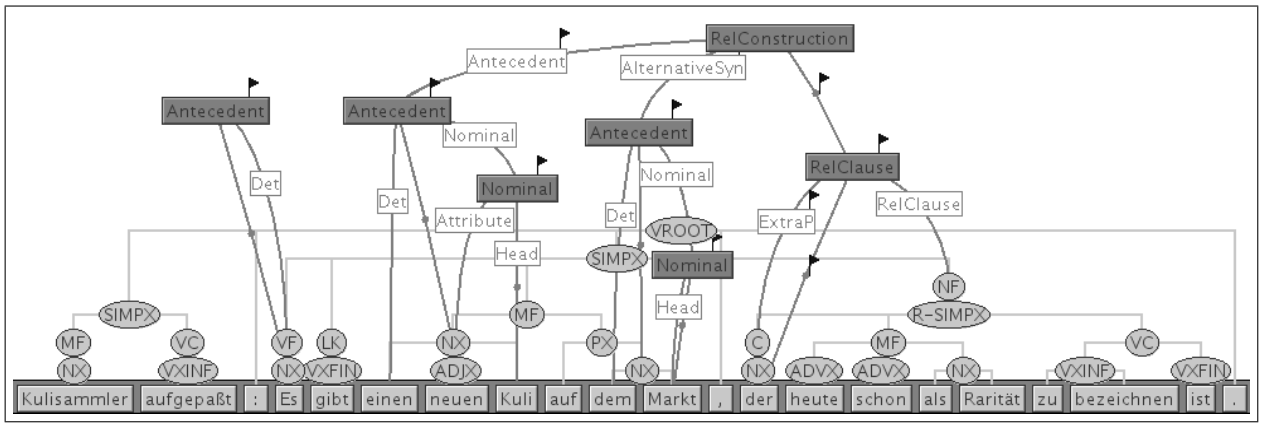

Figure 1: A *RelConstruction* frame connects relative clause and antecedent. A potential alternative antecedent is connected to the *RelConstruction* using the element *AlternativeSyn*.

an integrated relative clause can be identified as the direct sister of the antecedent phrase. The actual antecedent of an extraposed relative clause is either connected to the relative clause with a secondary edge or can be identified by a functional label that indicates the grammatical function of the relative clause's antecedent within the matrix clause.

In addition to the structural and relational annotation of relative constructions, potentially interesting features are also derived automatically from the original annotation of the treebank and added as *flags* to the relevant frames or frame elements. Automatically derived features include the position of the relative clause with the possible values *integrated*, *extraposed*, and *edge* (if the relative clause and its antecedent are not separated but occur together at the right edge of the matrix clause). *Antecedent* frames are automatically annotated with information about syntactic category, definiteness (derived from the presence and type of determiner), grammatical function, case, number, and gender. The *ExtraP* element of the *RelClause* frame is also adorned with flags providing case and grammatical function of the extracted phrase within the relative clause. Flags with information about the case, number, and gender of the relative pronoun are also added to the anchor element of the *RelClause* frame. Various lengths and distances are also measured automatically, e.g. the length of the relative clause and the actual and potential antecedents and the distance between the relative clause and the antecedent's head and right edge. Moreover, the depth of embedding of the antecedent in the matrix clause (measured in crossed maximal projections) and the actual path of crossed phrase boundaries is also determined automatically from the original syntactic annotation of the treebank.

The annotation tool SALTO makes it possible to easily inspect, correct, and further enrich the automatically derived annotation with information added manually. So far, only the feature *restrictiveness* of the relative clause with the values *restrictive*, *appositive*, and *unkown* has been manually annotated. Other types of features that will likely be added manually include information structure, givenness, and animacy, which have often been identified as relevant factors in studies on syntactic alternations; cf. e.g. Bresnan et al. (2007). Shannon (1992) also stresses the importance of information structure for extraposition. The relational annotation of relative constructions is also further expanded manually: Potential antecedents that both match the morphosyntactic features of the relative pronoun and are semantically plausible alternatives are connected to the *RelConstruction* frame using the element *AlternativeSem*. Alternative antecedents that match morphosyntactically but do not make sense semantically are instead connected with the *AlternativeSyn* element.

The annotation scheme presented here provides a basic scaffold for the annotation of relative constructions and can easily be expanded with additional features. It is well-suited both for manual and automatic annotation and allows the annotated features to be extracted relatively easily for subsequent statistical analysis, as the following pilot studies demonstrate.

3. Using the treebank to study relative clause extraposition

3.1. Testing constraints from the theoretical literature

The following three univariate studies illustrate the use of the treebank to test theoretical claims made in the literature on relative clause extraposition.

3.1.1. Syntactic locality

Generative theories of syntactic locality predict that relative clauses cannot be extraposed from antecedents embedded arbitrarily deeply inside the matrix clause. For example, Chomsky (1973)'s *Subjacency* principle predicts that a relative clause cannot be extraposed out of an NP that is embedded in another NP. Baltin (2006)'s theory of *Generalized Subjacency* is even more restrictive. He claims that extraposition can cross at most one maximal projection, which means that the antecedent of an extraposed relative clause has to be a direct subconstituent of the matrix clause. However, Müller (2004) and Kiss (2005) have presented both authentic and invented German counterexamples to these predictions. As the embedding of the antecedents could be automatically measured in terms of maximal projections in the treebank, the predictions of generative theories of subclausal locality can be tested in a systematic manner on actual corpus data. Table 1 shows the percentage of clearly

extraposed relative clauses depending on the depth of embedding of the antecedent in the matrix clause. This information is also rendered graphically in the associated diagram next to the table itself.
Table 1 does show that the likelihood of extraposition decreases with increasing depth of embedding. This is also confirmed by comparing a binary logistic regression model predicting extraposition that includes the factor *depth of embedding* to a baseline model without this factor ($\chi^2 = 22.64$, $df = 1$, $p < 0.001$). Still, this locality effect is quite gradual. Generalized Subjacency, for example, would predict a sharp decline in the likelihood of extraposition at a depth of embedding of one. However, almost the same percentage of relative clauses whose antecedent has a depth of embedding of one as those whose antecedent is a direct subconstituent of the matrix clause are extraposed: 24% of the former vs. 25% of the latter. There is even one relative clause whose antecedent is embedded four levels down inside the matrix clause. A manual inspection of the 43 extraposed relative clauses whose antecedent was embedded two levels deep also revealed that 40 of them violate Chomsky's original formulation of Subjacency. It thus seems that while generative theories of Subjacency do correctly predict an influence of syntactic locality, such locality constraints should not be regarded as categorical. The corpus data rather show syntactic locality to have a gradual, probabilistic nature (which could possibly be explained as a processing effect); cf. also Strunk and Snider (forthcoming).

3.1.2. Definiteness

The theory of Guéron and May (1984) connects extraposition to the phenomenon of quantifier raising. This predicts that relative clauses can only be extraposed from indefinite or quantified NPs, but not from NPs containing a definite article or a demonstrative as determiner – cf. also Baltin (2006). As information about the definiteness of the antecedent of relative clauses could be derived automatically from the original treebank annotation and has already been checked manually, this prediction can also be tested using the corpus of German relative clauses. Calculating the percentage of extraposed relative clauses for definite and indefinite antecedents shows that the prediction is not really borne out; cf. table 2.

	extraposed	integrated	edge
definite antecedent	252	590	480
(%)	19%	45%	36%
indefinite antecedent	335	334	453
(%)	30%	30%	40%

Table 2: Likelihood of relative clause extraposition depending on the definiteness of the antecedent

While extraposition from definite NPs indeed occurs less often – namely in 252 out of 1,322 cases (19%) – than extraposition from indefinite or quantified NPs – which occurs in 335 out of 1,122 cases (30%) – this is again only a tendency, albeit a statistically significant one ($\chi^2 = 67.53$, $df = 2$, $p < 0.001$). These results show that accounts that explain extraposition in terms of quantifier raising of the antecedent cannot be the whole story, at least not for relative clause extraposition in German.

3.1.3. Restrictiveness

Ziv and Cole (1974) make the (somewhat obscure) claim that only restrictive but not appositive relative clauses can be extraposed. This can easily be refuted with counterexamples from the corpus; cf. the sentence in (4).

(4) *Damit wies der BGH die Klage*
thereby dismissed the BGH the lawsuit
[$_{NP}$ der Erben Melchiors __] zurück,
of.the heirs Melchior's back
[$_{RC}$ die das Gut nach der Wende
who the property after the reunification
zurückverlangt hatten.]
demand.back had
"The BGH thereby dismissed the lawsuit of Melchior's heirs, who had demanded back the property after the reunification." (TüBa-D/Z s973)

A systematic evaluation of the corpus data shows that there is again a grain of truth in Ziv and Cole's claim, but only in the form of a slightly higher percentage of extraposition of restrictive relative clauses compared to appositive relative clauses: 28% (334 of all 1,207 restrictive relative clauses) vs. 17% (180 of all 1,023 appositive relative clauses); cf. table 3.[4] There is thus again a statistically significant difference ($\chi^2 = 32.93$, $df = 2$, $p < 0.001$), but no justification for positing a categorical constraint.

	extraposed	integrated	edge
restrictive RC	334	450	423
(%)	28%	37%	35%
appositive RC	180	457	423
(%)	17%	45%	38%

Table 3: Likelihood of relative clause extraposition depending on the restrictiveness of the relative clause

3.2. A multivariate model of relative clause extraposition as a syntactic alternation

In this section, I explore the potential of the enriched treebank for modeling relative clause extraposition as a syntactic alternation that is potentially conditioned by a multitude of different factors simultaneously.
I use binary logistic regression to model the probability that a particular relative clause modifying a particular antecedent will be extraposed – specifically, the *glm* function from the R statistical environment (R Development Core Team, 2010). Even though I have distinguished between three positional variants (*extraposed*, *integrated*, and *edge*)

[4] The distinction between restrictive and appositive relative clauses is not always easy to make during annotation, especially with indefinite antecedents. 559 cases have therefore been left undecided. Moreover, the annotation of restrictiveness has so far only been carried out by the author so that interannotator agreement could not be determined. The results with regard to restrictiveness should therefore be regarded as preliminary.

	Depth of Embedding of the Antecedent							
	0	1	2	3	4	5	6	8
extraposed	423	177	43	11	1	0	0	0
(%)	25%	24%	16%	13%	5%	0%	0%	0%
integrated	628	260	133	35	11	3	1	2
(%)	38%	35%	48%	43%	50%	75%	33%	100%
edge	614	297	101	36	10	1	2	0
(%)	37%	40%	36%	44%	45%	25%	67%	0%

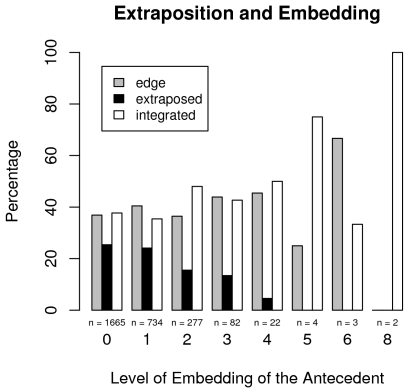

Table 1: Likelihood of relative clause extraposition depending on the depth of embedding of the antecedent

in the preceding section, I will disregard the *edge* variant for the purpose of statistical modeling because in the *edge* cases, the relative clause could not have been extraposed in principle since its antecedent is already located at the right edge of the matrix clause (unless one assumes vacuous extraposition) and it is therefore not clear theoretically whether these cases should be counted among the *integrated* class or should be regarded as a separate class; cf. also Uszkoreit et al. (1998). This simplification also has the benefit of allowing the use of an ordinary dichotomous logistic regression model, which is easier to interpret than a polytomous model.

The first model given in table 4 combines the three factors depth of *embedding* of the antecedent, *definiteness* of the antecedent, and *restrictiveness* of the relative clause, which have already been studied using univariate statistics in the preceding section, in order to test whether they all are significant predictors for the likelihood of relative clause extraposition and whether they are needed independently of one another. Factors with positive coefficients favor extraposition, factors with negative coefficients disfavor extraposition. The p-value indicates whether the coefficient of the given factor is significantly different from zero. As table 4 shows, all three factors make a significant contribution: First, increased depth of *embedding* of the antecedent makes extraposition less likely. Second, *indefinite* antecedents favor extraposition in comparison to *definite* antecedents. And third, *appositive* relative clauses are less often extraposed than *restrictive* ones. Testing

Factor	Coeff.	Std. Err.	z value	p value
(Intercept)	-0.55	0.11	-5.25	<0.001
embedding	-0.16	0.072	-2.27	0.023
indef. NP	0.75	0.13	5.83	<0.001
app. RC	-0.37	0.13	-2.89	0.004

Table 4: Logistic regression model of relative clause extraposition with the factors from section 3.1.

whether each of the three factors can be dropped using the log-likelihood ratio test also indicates that all three factors should be kept in the model: *embedding* ($\chi^2 = 5.36$, $df = 1$, $p = 0.021$), *definiteness* ($\chi^2 = 34.05$, $df = 1$, $p = <0.001$), and *restrictiveness* ($\chi^2 = 8.40$, $df = 1$, $p = 0.004$). Both the coefficients and the log-likelihood ratio tests suggest that *definiteness* is the most important of the three factors followed by *restrictiveness* and *embedding* as the weakest of the three. However, an evaluation of this first model in terms of prediction error using tenfold cross-validation shows that its error rate of 36.32% is only slightly lower than the error rate of a baseline model using only the *a priori* probability of extraposition (36.73%).

I therefore fitted a second model to the corpus data incorporating various factors that could be automatically deduced from the underlying treebank (cf. section 2.2.) in addition to the three factors from the first model. The following is a list of all factors used in the second model (ordered by the part of the relative construction they belong to):

(5) **antecedent:** case + cataphoric + definiteness + embedding + gender + grammatical function + length + length of modifiers + number + number of modifiers + part-of-speech of determiner + proper + syntactic category + syntactic category of immediately dominating phrase + topological field; **relative clause:** length + restrictiveness; **relative pronoun:** case + embedding + gender + grammatical function + number + pronominal form; **extracted phrase inside the relative clause:** grammatical function + length + syntactic category

For all of these 26 factors, I performed model comparisons between the full model and a model without the respective factor. The list in (6) contains all factors that I retained in the third and final model to be discussed below because for them the log-likelihood ratio test was at least marginally significant ($p < 0.1$).

(6) **antecedent:** case + cataphoric + definiteness + embedding + syntactic category + topological field; **relative clause:** length + restrictiveness

The first thing to note about this reduced set of predictors in the final model is that it does not contain any factors relating to the relative pronoun or the whole extracted and possibly pied-piped phrase inside the relative clause, nor

does it contain any factors relating to the internal structure, complexity or length of the antecedent. The final model thus suggests that the factors governing relative clause extraposition mostly have to do with "external" properties of the antecedent and the relative clause rather than their "internal" structure; however, cf. also Strunk and Keßelmeier (2010).

Table 5 provides the coefficients and p-values for the final model. It shows that the original factors *embedding*, *definiteness*, and *restrictiveness* are still significant. The most important factor is the position of the antecedent within the topological structure of the German clause: Specifically, the likelihood of extraposition decreases dramatically if the antecedent is located in the *Vorfeld* ("prefield") in front of the finite verb; cf. also Shannon (1992). However, contra Shannon (1992), extraposition from the *Vorfeld* is not categorically ruled out; cf. also Strunk and Snider (forthcoming). These observations are also in accordance with Uszkoreit et al. (1998), who found that the linear distance between the antecedent and the relative clause in words was the strongest factor influencing extraposition in their studies; cf. also the theory by Hawkins (2004).[5]

Another strong factor is the *case* of the antecedent: Compared to the *nominative*, the other three cases all favor extraposition. This could either be a genuine effect of case or could be traced back to the antecedent's grammatical function or maybe ultimately to its position again since subjects tend to precede objects even within the *Mittelfeld* ("middle field") (cf. also footnote 5).

As predicted, for example, by Hawkins (2004) and Wasow (2002), the length of the relative clause also plays a role in that longer relative clauses are more likely to be extraposed.

The occurrence of a special cataphoric demonstrative with *jene/jenige* (factor *cataphoric*) was selected as significant by the model comparison but only seems to have a relatively small influence. I will not discuss the remaining two factors *complex name* and *Vorfeld* here because the former occurred only 11 times in the data and the estimation of the latter's coefficient seems to be problematic.

The final model decreases the prediction error rate dramatically to only 15.47% compared to the baseline error rate of 36.73%, again evaluated using tenfold cross-validation. The final model is thus already quite successful in predicting whether a relative clause will be extraposed or not.

Factor	Coeff.	Std. Err.	z value	p value
(Intercept)	-1.59	0.30	-5.23	<0.001
embedding	-0.34	0.14	-2.37	0.018
indef. NP	1.44	0.23	6.28	<0.001
app. RC	-0.63	0.21	-3.03	0.002
RC length	0.15	0.03	5.70	<0.001
acc. case	1.81	0.26	7.05	<0.001
dat. case	1.22	0.28	4.35	<0.001
gen. case	2.16	0.46	4.72	<0.001
complex name	2.11	1.06	1.99	0.046
cataphoric	0.61	0.52	1.18	0.238
Nachfeld	-20.76	554.35	-0.04	0.97013
Vorfeld	-5.13	0.49	-10.55	<0.001

Table 5: Final logistic regression model of relative clause extraposition

4. Summary

I have described a syntactic treebank enriched with a second annotation level specific to research on relative constructions and relative clause extraposition. I hope that the pilot studies presented in this paper have already demonstrated the value of this enriched corpus. The univariate pilot studies have shown that generalization about extraposition from the theoretical literature with regard to syntactic locality, definiteness, and restrictiveness go in the right direction but go too far in positing categorical constraints.

The preliminary multivariate model of relative clause extraposition that I have presented also shows great promise in accounting for multiple noncategorical interacting factors that influence the likelihood of extraposition simultaneously. Even though I plan to further improve the current annotation of the treebank by checking and correcting automatically annotated features and by adding potentially relevant additional features such as information structure or animacy manually (also evaluating interannotator agreement whenever possible), the preliminary model can already predict extraposition with an error rate as low as 15.47%. I plan to make the enriched corpus publically available in the future once the annotation of the treebank has reached a reasonably complete state.

5. Acknowledgements

The author was partly supported by a scholarship from the German National Academic Foundation.

6. References

Mark R. Baltin. 2006. Extraposition. In Martin Everaert and Henk C. van Riemsdijk, editors, *The Blackwell Companion to Syntax*, volume 2, pages 237–271. Blackwell, Malden, Massachusetts.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam, The Netherlands.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Padó. 2006. SALTO – A versatile multi-level annotation tool. In Proceedings of LREC 2006, Genoa, Italy.

Noam Chomsky. 1973. Conditions on transformations. In Stephen R. Anderson and Paul Kiparsky, editors, *A Festschrift for Morris Halle*, pages 232–286. Holt, Rinehart and Winston, New York.

[5] Linear distances in words have also been automatically measured for the present corpus but have not yet been included in the models because they still have to be manually checked and corrected.

Elaine J. Francis. 2010. Grammatical weight and relative clause extraposition in English. *Cognitive Linguistics*, 21(1):35–74.

Jacqueline Guéron and Robert May. 1984. Extraposition and logical form. *Linguistic Inquiry*, 15(1):1–31.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford, UK.

Tibor Kiss. 2005. Semantic constraints on relative clause extraposition. *Natural Language and Linguistic Theory*, 23:281–334.

Wolfgang Lezius. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, volume 8, number 4.

Stefan Müller. 2004. Complex NPs, subjacency, and extraposition. *Snippets*, 8:10–11.

R Development Core Team. 2010. R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Thomas F. Shannon. 1992. Toward an adequate characterization of relative clause extraposition in modern German. In Irmengard Rauch, Gerald F. Carr, and Robert L. Kyes, editors, *On Germanic Linguistics. Issues and Methods*, pages 253–281. Mouton de Gruyter, Berlin/New York.

Jan Strunk and Katja Keßelmeier. 2010. Using methods from contrastive corpus analysis to study lexical aspects of alternations. Poster at Linguistic Evidence 2010, Tübingen, Germany.

Jan Strunk and Neal Snider. forthcoming. Subclausal locality constraints on relative clause extraposition. In Heike Walker, Gert Webelhuth, and Manfred Sailer, editors, *Rightward Movement from a Cross-linguistic Perspective*. John Benjamins, Amsterdam.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2005. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

Hans Uszkoreit, Thorsten Brants, Denys Duchier, Brigitte Krenn, Lars Konieczny, Stephan Oepen, and Wojciech Skut. 1998. Studien zur performanzorientierten Linguistik: Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft*, 7(3):129–133.

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications, Stanford, USA.

Yael Ziv and Peter Cole. 1974. Relative extraposition and the scope of definite descriptions in Hebrew and English. In Michael W. La Galy, Robert A. Fox, and Anthony Bruck, editors, *Papers from the Tenth Regional Meeting of the Chicago Linguistic Society, April 19–21, 1974*, pages 772–786. Chicago Linguistic Society, Chicago.