ENRICHING A TREEBANK TO INVESTIGATE RELATIVE CLAUSE EXTRAPosition IN GERMAN

Jan Strunk
strunk@linguistics.rub.de

Relative Clause Extraposition

- Relative clauses in German can be realized as part of the head noun phrase (integrated) or at the end of the matrix clause (extraposed) without an apparent semantic difference.

- Integrated Relative Clause
  - 
  -Threads (de ich [finden konnte]) gelenken (de ich [finden konnte]) gelesen.
  - Ich habe alle Bücher gelesen (die ich [finden konnte]).

- Extraposed Relative Clause
  - Ich habe alle Bücher gelesen (die ich [finden konnte]).
  - Ich habe alle Bücher gelesen (ich habe [finden konnte]).

Basic Corpus

- Tübinger Baumbank des Deutschen / Schriftpraphic (TüBa-D/Z) (Tübingen Treebank of Written German) (Teipolham et al., 2005)

- Annotated with a relatively flat syntactic structure including topological fields, part-of-speech and morphological tags.

- Sub-corpus including all sentences that contain a relative clause (e-seen) extracted using TIGERSearch, 2603 sentences with 2789 relative clauses.

- Intended uses of the treebank:
  - Empirical investigation of constraints on relative clause extraposition
  - Building a statistical model of relative clause extraposition as a syntactic alternation
  - Training an attachment disambiguation system for (extraposed) relative clauses

Enriching the Treebank with Special-Purpose Annotation

Locality (Depth of Embedding of the Antecedent)

- Generative theories of locality predict that the antecedent of an extraposed relative clause cannot be embedded arbitrarily deep.

- Chomsky’s (1975): Subjacency principle rules out extraposition from an NP/DP that is embedded inside another NP.

- Baltin’s (2008): Generalized Subjacency predicts that the relative clause must be addressed to the next higher maximal projection.

- These theories predict a sharp decline in extraposition likelihood for all antecedents that are embedded at least one level deep.

- Extraposition likelihood does decrease with increasing depth of embedding (likelihood ratio test: $X^2 = 22.84, df = 1, p < 0.001$) but much more gradually predicted.

Definiteness of the Antecedent

- Guerlin & May (1994): correct extraposition to quantifier raising

- This predicts that extraposition should only be possible from indefinite or quantified antecedents but not from definite ones

A Multivariate Model of Relative Clause Extraposition

- A binary logistic regression model of relative clause extraposition (trained on the integrated and extraposed cases from the corpus, edge cases were excluded)

- The initial model included the three factors from the univariate studies above and additional factors that could be derived from the treebank automatically:
  - Antecedent case: case and cataphoric vs. embedding vs. gender and grammatical number: length + length of modifiers + number of modifiers + part-of-speech of proper noun vs. syntactic category vs. syntactic category of immediately dominating phrase vs. topological features
  - Relative Clause: length + restrictiveness
  - Restrictiveness: case + embedding + gender and grammatical number + number of pronoun-like elements

- Extracted Phrase in the Relative Clause: grammatical function + length + syntactic category

Main Findings

- After manual model comparison and selection, the final model contained 12 factors.

- Among them also definiteness, embedding, and restrictiveness.

- It achieves an error rate of 16.47% (3D-fold cross-validation) against a baseline of 38.73%.

- Factors relating to the internal structure of the antecedent and the relative clause are not important.

- The most important factor is the position of the antecedent in the matrix clause (Vorfeld position discourages extraposition).

- Vorfeld (1992, Hawkins 2004) and also the antecedent’s case and definiteness

Final Model

- factor coeff. std. err. z value p value
  - Intercep -1.59 0.30 -5.23 <0.001
  - Embedding -0.34 0.14 -2.37 0.018
  - indef. DP 1.44 0.23 6.28 <0.001
  - acc. RC -0.83 0.21 -3.90 0.002
  - length RC 0.15 0.03 5.70 <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 0.19 11.8 0.004
  - cataphoric 0.61 0.52 1.18 0.238
  - Vorfeld -20.76 554.35 -0.04 0.970
  - Vorfeld -5.13 0.40 -10.55 <0.001

Conclusion

- Corpus data show that intuitions from the generative literature about constraints on relative clause extraposition go in the right direction but too far by assuming categorical constraints

- A preliminary logistic regression model to predict extraposition achieved an error rate of as low as 15.47% and indicated that multiple noncategorical interacting factors are necessary