Relative Clause Adjacency as a Characteristic of 18th Century German

Sophia Voigtmann (Saarland University) & Katrin Ortmann (Ruhr-Universitaaet Bochum)

Keywords:
Relative Clauses, Extraposition, Automatic Annotation, New High German, Information Density

In German, about half of all attributive relative clauses (RCs) are separated from their head noun (1), but the explanations for this variation are diverse and only partly underpinned by evidence – especially for Early Modern German.

(1) Der Herr [...] wird [den] nicht ungestraft lassen, **der seinen Namen misbrauchet.**
The Lord [...] will him not unpunished leave **who his name abuses.**
'The lord, as a fair judge, will not leave him unpunished who abuses his name.' (Christliches Bedencken von dem vorsetzlichen Meineid, J. H. Benner, 1739, DTA)

Zifonun et al. (1997) and Uszkoreit et al. (1998), among others, agree that information disentanglement is one of the key factors driving extraposition. According to Information Theory (Shannon 1948), information is defined as the probability of occurrence of a word in its context, the so-called surprisal of a word. The higher the surprisal, the higher the cognitive load of a word and vice versa (Hale 2001). Surprisal can therefore be related to processing difficulties (Hale 2001, Levy 2008) and its effects can also be found in corpus data (Levy & Jaeger 2007, Jaeger 2010, among others).
In the case of RC extraposition, we expect a correlation between extraposition and high surprisal values because moving the RC to the end of the sentence frees up cognitive capacities and distributes information more evenly across the sentence (Gibson 1998, Hawkins 1992). However, Voigtmann & Speyer (2021) have shown the opposite for RCs in Early Modern German of the 18th century. In this time period, RCs with high surprisal values are placed adjacent to their head nouns.
One possible explanation for this observation comes from the Language Dependent Structural Forgetting hypothesis (Futrell et al. 2021) which states that the frequent usage of structures can facilitate their processing even if principles of successful communication are violated. According to this, readers and writers of the 18th century might have been so familiar with long and complex middle fields that a dense writing style and RC adjacency was considered prestigious (Konopka 1996). But since Voigtmann & Speyer (2021) focus on scientific texts, another explanation for their findings could also be that the embedding of RCs despite high surprisal values is a style characteristic of scientific writing of this century.
To test these hypotheses, we look at RC extraposition in non-scientific genres from the GerManC Corpus (Bennett et al. 2007). Our data set includes 132 texts (148.705 tokens) from the 18th century from the genres of humanities, narrative prose, newspaper, and sermons. The corpus is provided with automatically created POS tags, lemmas, and orthographic normalization. Using a constituency parser that was trained on modern and historical German treebanks (Ortmann 2021), we automatically identify the RCs in our data sample and classify them as embedded or extraposed with the help of a topological field parser (Ortmann 2020).
We then train a 2-skipgram language model on lemmatized date taken from the German Text Archive (DTA) from 1700 to 1800 and run a linear regression analysis (R, The R Core Team 2018) similar to the procedure in Voigtmann & Speyer (2021). The regression shows a significant correlation between RC embedding and high cumulative surprisal values ($z=2.043$, $p < 0.01$) independent of the genre under investigation.
These results are in line with previous findings and lead us to conclude that the embedding of highly informative RCs in a position where they should be harder to understand is not only a phenomenon of scientific writing in this time period but a characteristic of 18th century German. Our study, thus, supports the hypothesis of habituation (Futrell et al 2021) to a dense writing style in which embedding is considered prestigious and is preferred over extraposition even in texts closer to orality (Koch & Oesterreicher 2007).

References:

Bennett, P., M. Durrell, A. Ensslin, S. Scheible, R. Whitt. 2007. GerManC (Version 1.0). University of Manchester.

DTA. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. Berlin, 2007-2021.

Futrell, R., E. Gibson & R. Levy. 2021. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. Cogn Sci. 45(5).

Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. COGNITION 68. 1–76.

Hale, J. 2001. A probabilistic early parser as a psycholinguistic model. Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics.

Hawkins, J. A. 1992. Syntactic weight versus information structure in word order variation. Informationsstruktur und Grammatik. Linguistische Berichte Sonderhefte 4. 196–219.

Jaeger, T. F. 2010. Redundancy and reduction: speakers manage syntactic information density. Cogn. Psychol. 61:2.

Koch, P. & W. Österreicher. 2007. Schriftlichkeit und kommunikative Distanz. Zeitschrift für germanistische Linguistik 35(3). 346–375.

Konopka, M. 1996. Strittige Erscheinungen der deutschen Syntax im 18. Jahrhundert. Tübingen: dissertation.

Levy, R. & F. Jaeger. 2007. Speakers optimize information density through syntactic reduction. In B. Schölkopf J. C. Platt T. Hoffman (ed.) Advances in neural information processing systems 19, 849–856. MIT Press.

Levy, R. 2008. Expectation-based syntactic comprehension. Cognition 106(3). 1126–1177.

Ortmann, K. 2020. Automatic Topological Field Identification in (Historical) German Texts. Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL). 10–18.

Ortmann, K. 2021. Automatic Phrase Recognition in Historical German. Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021). 127–136.

R Core Team. 2018.R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/

Shannon, C. E. 1948. A mathematical theory of communication. The Bell System Technical Journal 27(3). 379–423.

Uszkoreit, H., T. Brants, D. Duchie B. Krenn, L. Konieczny & S. Oepen. 1998. Studien zur Performanzorientierten Linguistik. Aspekte der Relativsatzextraposition im Deutschen. Kognitionswissenschaft 7:65.

Voigtmann, S. & A. Speyer. 2021. Information density and the extraposition of German relative clauses. Frontiers in Psychology. 1–18.

Zifonun, G., L. Hoffmann & B. Strecker. 1997. Grammatik der deutschen Sprache. Bd 2. Berlin: Niemeyer.