Where do all the NPs go? – A corpus linguistic study on NP extraposition in German scientific writing from 1650 to 1900

Although in modern German, it is highly marked to place an NP in the postfield,[1] the phenomenon is not as rare as expected in early New High German (1650-1900) data (ex. A).

A. …weil er […] von den meisten Medicis [genennet wird]$_{RSB}$ **ein Schmid aller Kranckheiten**.
... as he ... by the most doctors called is a forger of.all sicknesses.
"...as he is called a forger of sicknesses by most physicians." (Abel 1699, 225)

However, studies concerned with extraposition in diachrony treat the placement of NP as a marginal phenomenon that can nearly exclusively be explained by the length of the NP (Ebert 1980, Sapp 2014) or pragmatic factors like givenness (Light 2011).

Although it is not mentioned as such in these studies, both explanations can be linked to processing difficulties which are resolved by extraposition. Processing difficulties can be rather objectively investigated using Information Density, namely Surprisal (ID; Shannon 1948). Levy and Jaeger (2007; 1) define ID as the "amount of information per unit comprising the utterance". It is calculated as the likelihood with which a word occurs in a context (P(word) = -log2(word|context)). More expected combinations of words result in lower surprisal values and, thus, in lower perceiving difficulties (Hale 2001), as low surprisal values reduce the impact of the working memory (Levy & Jaeger 2007, Hale 2001, Levy 2008). We claim that the surprisal value of NPs is also relevant for their placement in the postfield. Therefore, we propose that NPs with high surprisal values are more likely to be extraposed.

To investigate this claim, we built a corpus of medical and theological texts from 1650 to 1900 taken from the Deutsches Textarchiv (DTA, BBAW 2019). We manually extracted extraposed and embedded NP and the sentence brackets using WebAnno (Eckart de Castilho et al. 2016). Then, we calculated a 2-Skip-Bigram-Language Model (Guthrie et al. 2016) to gain surprisal values for every word in the context. These surprisal values were used to calculate the mean Skipgram surprisal on lemmata for every annotated NP. Furthermore, we determined the length of the NP, the text genre (medical vs. theological), and the Orality Score (COAST, Ortmann & Dipper 2022) since extraposition is claimed to be more likely in conceptionally oral texts (Koch & Oesterreicher 2007) and the time of publication, the period. To determine the most influential factor for extraposition, logistic regression was performed with R (The R Core Team 2022).

As a result, we find that extraposition is indeed linked to high surprisal values ($z=-2.44$, $p<.05$ *) and that length is not significant ($z=-0.48$, $p<0{,}63$), in contrast to the aforementioned literature. However, both the genre ($z=-2.58$, $p<.001$**) and the interaction between Orality Score and the period ($z=-2.68$, $p<.001$**) are more significant. That suggests an influence of genre and a change over time. The latter is furthermore supported by a slightly significant result for the interaction between length and period ($z=-1.75$, $p<.1$).

Following Speyer (2015: 499), we suggest that there are more processing capacities available behind the right sentence bracket because the main verb is eventually processed at this point. Thus, there is no uncertainty about the constituent function of the extraposed phrases, which causes further strain on the working memory. This leaves more capacities to process lexical difficulties, represented by the surprisal values. In our corpus, the effect is more pronounced than the influence of length. Furthermore, we detect indications of language change in the interactions and an influence of the genre, suggesting a difference in writing style that could yield further investigations.

[1] The postfield is the position behind the right sentence bracket (RSB) and the RSB is the position late in the clause where verbal material, which is distributed over two positions in the clause in German, occurs (Wöllstein 2014).

Example taken from:

Abel, H. (1699). *Wohlerfahrner Leib-Medicus Der Studenten*. Leipzig: Groschuff.

References:

BBAW (2019). Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften; http://www.deutschestextarchiv.de/. [last accessed: 2023-01-19]

Ebert, R. P. (1980). Social and stylistic variation in early new high german word order: The sentence frame (›satzrahmen‹). 102. Jahresband, 357–398.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016): A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the LT4DH workshop at COLING 2016*, Osaka, Japan.

Guthrie, D., B. Allison, W. Liu, L. Guthrie, and Y. Wilks (2006). A closer look at skip-gram modelling. *Proceedings of the Fifth International Conference on Language Resources and Evaluation.*

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 Computional Linguistics*.

Koch, P. & Oesterreicher, W. 2007. Schriftlichkeit und kommunikative Distanz. *ZGL 35*, 346-375.

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

Levy, R. & Jaeger, F. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems 19*. 849-856.

Light, C. (2011). The information structure of subject extraposition in early new high german. In S. Müller (Ed.), *Proceedings of the HPSG 2011 Conference*

Ortmann, K. and S. Dipper (2022a). Coast (conceptual orality analysis and scoring tool). https://github.com/rubcompling/COASTcoast-conceptual-orality-analysis-and-scoring-tool [last accessed: 2023-01-18]

R Core Team (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Sapp, C. D. (2014). Extraposition in middle and new high german. *The Journal of Comparative Germanic Linguistics 17(2)*, 129–156.

Speyer, A. (2015a). Auch früher wollte man informieren – Zum Einfluss der Informationsstruktur auf die Syntax in der Geschichte des Deutschen. *Zeitschrift für germanistische Linguistik 43(3),* 485–515.

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

Wöllstein, A. (2014). *Topologisches Satzmodell* (2 ed.). Heidelberg: Winter.