

Where to place a phrase? An informational and generative approach to phrasal extraposition

The German surface clause structure is characterized by the so-called ‚sentence brackets‘, i.e. the phenomenon that verbal material is distributed over two positions in the clause, one hosting the finite part of the verbal form in the left periphery (‚left sentence bracket‘, corresponding to C°, e.g. Dürscheid 1989), and the base position of the verb at the right edge of the clause (‚right sentence bracket‘, RSB). While in most cases the end of the clause coincides with the RSB, sometimes clauses in which material follows the RSB can be found (1). In descriptive terms, the position to the right of the RSB is called ‚postfield‘ (PoF).

- (1) Gestern hat Paul Susie Geld t₁ [RSB gegeben] [PoF für ein Eis]₁.
yesterday has Paul Susie money given for an ice-cream
,Yesterday, Paul gave Susie some money for ice-cream.’

The structural analysis of the German PoF has been a matter of debate from the 1980s on. The theories fall basically in four categories: (a) extraposition accounts in the sense of movement of a phrase and its subsequent adjunction to vP or TP (e.g. v. Stechow & Sternefeld 1988), (b) base generation (e.g. Haider 2010), which involves a variable base with respect to headedness (right-branching in the case of PoF constituents), (c) post-syntactical accounts that view extraposition as a PF phenomenon (e.g. Truckenbrodt 1995), (d) mixed accounts, e.g. Inaba (2007) who assumes extraposition in the case of non-sentential constituents, base generation in the sense of sentential constituents, and post-syntactical extraposition of relative clauses, or Frey (2015), who basically adheres to a base-generation account, but assumes post-syntactical extraposition of all attributive material.

While the usage of the PoF is rather restricted in Modern German, it was far more common in older stages of German, e.g. Early New High German (ENHG). A study of PoF filling in ENHG is consequently fruitful in order to investigate the conditions under which the PoF is utilized, and to draw conclusions about the underlying structure. We want to present a corresponding study. On a corpus of extraposed and embedded Prepositional Phrases (PP), Nominal Phrases, Adverbial Phrases, Adjective Phrases, comparative elements and Relative Clauses (RC) taken from medical and theological texts from the Deutsche Textarchiv (DTA) from 1600 to 1900 we aim to find out whether the placement to the PoF is influenced by Information Density (ID) in terms of Shannon (1948) and whether ID can help to decide on the structure of this material. ID can be defined as the “amount of information per unit comprising the utterance” (Levy & Jaeger 2007, 1). It is measured as surprisal, calculated via $-\log_2 P(\text{word}|\text{context})$. So frequent combinations have lower surprisal values than rare combinations. Since surprisal is connected to perceiving difficulties (Hale 2001), the impact of frequent combinations with minor surprisal-values on the working memory is lower than it is for rare combinations with higher surprisal-values (Levy & Jaeger 2007, Hale 2001). This corresponds to the Uniform Information Density Hypothesis (UID) by Levy and Jaeger (2007), that claims that producers distribute information as evenly as possible across a discourse, and to findings by Haider (2012, 192): „DPs at the right edge are well formed only if they are adverbial ... or if they are so ‚heavy‘, viz. lengthy, that the working memory relief overrides the restrictions of grammar”. The heaviness of extraposed material is included in another value derived from surprisal: the cumulative surprisal. This is the sum of the surprisal values of all

words in a complex syntactic unit (e.g. a constituent or a clause). This measurement correlates slightly with the length of the material, because more (possibly difficultly perceivable) words in the PoF create a higher cumulative surprisal value.

This is visible in the results for the placement of PPs in the PoF. Having manually determined the phrases mentioned above and calculated the cumulative surprisal values using a bigram language model for every 50 years with Kneser-Ney smoothing (Chen & Goodman 1999) on the lemmata of similar texts from the DTA, we then used a logistic regression (Generalized Linear Mixed-Effects Model, lme4 package, R) over the summed surprisal values. The results for 312 PPs from 1650 to 1700 and 1850 to 1900 was highly significant ($|z| = -5,461$, $p < 0,001$).

This talk will concentrate on ENHG data that might give hints to the structural analysis of the PoF in ENHG, using, e.g. the diagnostics of Haider (2010). As e.g. data like in (1) in which the PoF constituent is moved out of another constituent is very rare in ENHG (0.2% of instances in our corpus; one example in (2)), we take this as an indication that the conditions for a left-headed VP that Haider (2010) formulates are not fulfilled, so this does not seem to be an adequate account of ENHG. We will discuss whether a 'real' structural change has taken place or rather a quantitative redistribution of available variants.

(2) Ich will [Feindschaft t₁] setzen
 I want_to hostility create
 [zwischen der Schlangen Samen und des Weibes Samen]₁
 between the snake's seed and the woman's seed.
 (Arndt, 1610, sentence 287)
 'I want to create hostility between the snake's and the woman's seed.'

References:

- Dürscheid, Christa (1989): Zur Vorfeldbesetzung in deutschen Verbzweit-Strukturen. Trier: Wissenschaftlicher Verlag Trier.
- Frey, Werner (2015): Zur Struktur des Nachfelds im Deutschen. In: Vinckel-Roisin, Hélène (ed.): Das Nachfeld im Deutschen. Theorie und Empirie. Berlin: de Gruyter, 53-76.
- Haider, Hubert (2010): The Syntax of German. Cambridge: CUP.
- 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.
- Inaba, Jiro (2007): Die Syntax der Satzkomplementierung. Berlin: Akademie Verlag.
- Levy, R. & Jaeger, F. 2007. Speakers optimize information density through syntactic reduction. In: Advances in Neural Information Processing Systems 19. pp. 849-856.
- Shannon, C. E. 1948. A Mathematical Theory of Communication. Bell System Technical Journal 27. pp. 379-423; 623-656.
- Truckenbrodt, Hubert (1995): Phonological phrases: their relation to syntax, focus, and prominence. PhD dissertation, MIT.
- von Stechow, Armin & Wolfgang Sternefeld (1988): Bausteine syntaktischen Wissens. Opladen: Westdeutscher Verlag.

Example taken from: Arndt, Johann (1610). Vom wahren Christenthumb. Bd. 1. Magdeburg. (http://www.deutschestextarchiv.de/book/show/arndt_christentum01_1610)