A typological paradox: Information structure and linguistic corpora

Dejan Matić
University of Münster

matic@uni-muenster.de

Research on information structure (IS) displays an interesting discrepancy between its proclaimed aims and the methods it applies. While the basic idea is to capture how interactional aspects of communication are linguistically encoded, IS is nevertheless investigated with the help of introspection and a couple of constructed sentences, to the full exclusion of the actual language use. IS categories are routinely identified through a limited number of litmus contexts, such as question-answer pairs, which have been taken as a basis for higher-level theorising on the nature of IS, in the form of Alternative Semantics, QUD, or similar (e.g., Rooth 1992, Roberts 2012). It has been argued that the results of this kind of approach are flawed, as they suppress linguistic variation and impose preconceived ideas of how interactive communication works on how speakers actually communicate and how this is reflected in the language (e.g., Matić & Wedgwood 2013, Ozerov 2018).

The use of spoken language corpora to better understand IS is doubtless a huge step forward. However, I am going to argue that it would be naïve to believe that simply switching from introspection to corpus is going to produce more satisfying descriptions and typologies of IS. The reason for this is that the notional apparatus of IS is based on language- and discourse-independent categories (Aristotelian theory of judgment, Gestalt psychology, Hamblin semantics of questions, etc.) rather than on empirical investigation. Searching for 'topic', 'focus', or some similar category from the IS repertoire in corpora is bound to result in finding the same kinds of structures that the introspective IS research has found, as the categories themselves pose a limit to the possible outcomes of the investigation, not unlike glasses that allow a person to see only in one resolution, as it were.

I assume – and there are good reasons to do so – that the structuring of linguistic information does not pertain to sentences but is rather a matter of the way utterances are incorporated in discourse. I also assume that there is not only one possible way to organise larger stretches of discourse but rather a potentially infinite number of ways, with different categories and different kinds of relevant factors. On these assumptions, the only way to investigate IS is to perform a qualitative analysis of individual corpora with as few research presuppositions as possible and try to find the relevant interpersonal, epistemic, processing, etc. factors to which the structures and utterances in the given corpus are sensitive. Only after this kind of analysis has been performed can we try and compare corpora in order to see to what extent the factors found in one are relevant in the other, and vice versa. The final result of this endeavour may turn out to be a tentative typology of IS systems. The proposed methodology of investigating IS in language corpora will be illustrated with case studies from two languages of Siberia, Even (North Tungusic) and Yukaghir (isolate). Certain aspects of the intricate IS systems of Even and Yukaghir will be compared to corpora from better known European languages.

References: Matić, D. & D. Wedgwood (2013). The meanings of focus: The significance of an interpretation based category in crosslinguistic analysis. Journal of Linguistics 49, 127–163. Ozerov, P. (2018). Tracing the sources of Information Structure: Towards the study of interactional management of information. Journal of Pragmatics 138, 77–97. Roberts, C. (2012). Information Structure: Towards an integrated formal theory of pragmatics. Semantics & Pragmatics 5.