Data-driven approach to rare phenomena in the syntax of embedded clauses

Edyta Jurkiewicz-Rohrbacher
Universität Regensburg

Rare phenomena are a problematic part of language use, since they are often denied or are missing in language descriptions, norms, or linguistic theories. It is hard to make any predictions about them. For this reason, they pose a problem for natural language processing and artificial intelligence tools, which rapidly develop nowadays.

The domain of syntax, in particular complementation, is very rich in such structures, because complex sentences are generally less studied and less frequent than simple sentences. In the talk, I will discuss the problem of rare phenomena using examples from Slavic:

1. (The lack of) clitic climbing in BCS, further called diaclisis,
2. Infinitival complementation with accusative controllers in Polish,
3. Double dative argument combinations in Russian.

Obtaining appropriate empirical material for such studies is challenging. As often argued in psycholinguistic research, certain structures are impossible to be obtained from corpora. My studies challenge this claim by working with massive corpora. I put forward that corpus-driven research is a crucial and necessary step to:

1. draw conclusions about phenomena that are infrequent, or at least
2. formulate the precise hypothesis which can be passed for further research (e.g., experimental), without this step designing experiments based on own intuition might turn very costly and risky.