The statistical analysis of cooccurrences: From collocations to arbitrary structures

Thomas Proisl

GSCL Research Talks, 2022-02-17

The study of cooccurrences, i. e. the analysis of linguistic units that occur together, has had a profound impact on our view of language. In this talk, I will discuss how we can generalize established methods for the statistical analysis of two-word cooccurrences and cooccurrences of words and constructions, i.e. collocations and collostructional analysis, to analyze cooccurrences of arbitrary linguistic structures. Starting from collocations, I will first discuss collostructional analysis, focussing on simple collexeme analysis and one of its methodological problems. The usual approach to simple collexeme analysis, i.e. the cooccurrence of a construction and a lexeme that occurs in one of its slots, requires the researcher to classify either all constructions in the corpus or all instances of a suitably defined class of constructions. In practice, it is often not possible or feasible to identify these constructions (this is sometimes referred to as "the problem of the bottom-right cell"). The insights gained from the suggested solution to this problem lead towards a generalized cooccurrence model for the statistical analysis of cooccurrences of arbitrary linguistic structures.