LiME
Linguistics @ Meertens

Towards the automatic detection of syntactic differences

Martin Kroon

Utrecht University

Can syntactic differences between languages be detected automatically, and if so, how? With the enormous number of natural languages and dialects, the very high level of variation they exhibit between one another, and the technically infinite number of possible sentences per language or dialect, systematic manual comparison is a hugely daunting task. The field of comparative syntax would therefore significantly benefit from the (partial) automatization of the process, as it would increase the scale, speed, systematicity, and reproducibility of research.

In this talk, which centers around my recent PhD research, I show through case studies involving English, Dutch, German, Czech, and Hungarian that correct hypotheses on syntactic differences between languages can be generated automatically from parallel corpora through the use of the minimum description length principle, counting mismatches between part-of-speech pattern occurrences, word alignment, and mapping annotation from an annotated language onto another unannotated language. The tools developed for the purposes of this research work well and can aid a linguist significantly in their search for differences or similarities, but do not replace the human researcher, whose instinct and interpretation remain crucial in the process.