Automatic Detection of Syntactic Differences through the Minimum Description Length principle and feature mapping

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 aspects of my recent PhD research, I will discuss a systematic approach to detect and rank hypotheses about possible syntactic differences for further investigation by leveraging parallel data and using the Minimum Description Length (MDL) principle. We deploy the SQS-algorithm ('Summarising event seQuenceS'; Tatti and Vreeken 2012) – an MDL-based algorithm – to mine 'typical' sequences of Part of Speech (POS) tags for each language under investigation, and create a shortlist of potential syntactic differences based on the number of parallel sentences with a mismatch in pattern occurrence. The approach proved useful in both retrieving POS building blocks of a language as well as pointing to meaningful syntactic differences between languages.

Additionally, I will briefly discuss tools that I developed in light of the question of whether extensive linguistic knowledge about a language can be leveraged in order to detect grammatical properties of a less well-described language and differences between the two languages. To this end, word alignment is used to map source language words onto target language words with the aim of detecting syntactic features of the target language and differences between source and target language by semi-automatically analyzing this mapping.