Speech rate cross-linguistically: The DoReCo database, final lengthening, and pause probabilities

Frank Seifart

In the first part of this talk, I will introduce DoReCo, an initiative to create a multilingual reference corpus, consisting of at least 10,000 words for at least 50 languages. DoReCo extracts from fieldwork-based language documentation collections narrative texts that are already transcribed, translated into a major language, and morphologically analyzed. Within DoReCo, these data are being converted to a common file format and time-aligned at the phoneme level using the MAUS software. In the second part of this talk, I will present two ongoing, cross-linguistic studies on a subset of this corpus: One study investigates phonetic lengthening as a function of utterance-final position. Another study investigates pause probabilities before nouns vs. verbs and relates findings to the fact that, typologically, there are fewer prefixes on nouns vs. verbs.