Corpus-based cross-linguistic research on the temporal dynamics of speech

Donnerstag,
05.03.2020
09:00–10:00
VMP5 2101/2105

Frank Seifart (Keynote)
Leibniz-Zentrum Allgemeine Sprachwissenschaft Berlin
seifart@leibniz-zas.de

Local speech rate variation and pauses provide us with a window into the cognitive-neural and physiological-articulatory bases of the human language production system (e.g., Jaeger & Buz 2017), but cross-linguistic variation in this domain remains understudied (Norcliffe et al. 2015). However, over the past 20 years, efforts to document endangered languages have produced vast amounts of annotated spoken language data from a wide variety of languages, which are time-aligned with audio.

In the first part of this talk, I will present an effort to tap into these resources by creating a multilingual reference corpus (DoReCo) from language documentation collections that are archived at repositories such as The Language Archive (TLA), especially from the DOBES collection. DoReCo extracts from such collections narrative texts that are already transcribed, translated into a major language, and time-aligned with audio files at the level of discourse units. Within the DoReCo project, these data are being converted to a common file format and time-aligned at the phoneme level using the MAUS software (Strunk et al. 2014). Corpora from at least 50 languages will be included, a subset of at least 30 of which are fully annotated for morpheme breaks and morpheme glosses. A minimum of 10,000 words per language words is set as a realistic corpus size for the short- or mid-term fieldwork-based projects from which most DoReCo corpus donations stem.

In the second part of this talk, I will present preliminary results of analyses of this corpus. One set of studies investigates cross-linguistic vs. language-specific patterns in utterance-final lengthening as indicative of prosodic boundaries – reflecting potentially species-wide articulatory constraints and cognitive constraints on planning, as well as potentially culture-specific discourse-unit signaling functions. Another set of studies investigates cross-linguistic vs. language-specific patterns in the temporal distribution of morphemes regarding information rate in terms of morphemes per second (following Pellegrino et al. 2011) and in the number of morphemes in inter-pausal units – both reflecting cognitive constraints on language production. I also address methodological challenges arising from the relatively small size of individual corpora in DoReCo, given the large number of varied factors that are known influence speech rate and pauses, including individual speaker variation and word token frequencies (Liberman 2019).

References: Jaeger, T. F. & E. Buz (2017). Signal Reduction and Linguistic Encoding. In E. M. Fernández & H. Smith Cairns (eds.), The Handbook of Psycholinguistics. Hoboken, NJ: Wiley, 38–81. Liberman, M. Y. (2019). Corpus Phonetics. Annual Review of Linguistics 5(1), 91–107. Norcliffe, E., A. C. Harris & T. F. Jaeger (2015). Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. Language, Cognition and Neuroscience 30(9), 1009–1032. Pellegrino, F., C. Coupé & E. Marsico (2011). A Cross-Language Perspective on Speech Information Rate. Language 87(3), 539–558. Strunk, J., F. Schiel & F. Seifart (2014). Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS. In N. Calzolari et al. (eds.), LREC 2014. Reykjavik: ELRA, 3940–3947.