DoReCo - building a cross-linguistic database of spoken language

Frank Seifart, Leibniz-ZAS, Berlin

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). A main reason for this has been the lack of relevant cross-linguistic data. However, over the past ca. 20 years efforts to document endangered languages have produced vast amounts of annotated spoken language data from a wide variety of languages from around the world, much of which are time-aligned with audio files, typically using ELAN (ELAN developers 2019).

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 at the level of discourse units with audio files. 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 some aspects of an ongoing study using this corpus. We study cross-linguistic vs. language-specific patterns in word-final lengthening as indicative of major vs. minor prosodic boundaries – reflecting potentially species-wide articulatory constraints and cognitive constraints on planning, as well as potentially culture-specific discourse-unit signaling functions. I also address methodological challenges arising from the relatively small size of individual corpora in DoReCo, which is problematic 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).

ELAN developers. 2019. *ELAN (Version 5.7) [Computer software]. (June 14, 2019)*. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.

Jaeger, T. Florian & Esteban Buz. 2017. Signal Reduction and Linguistic Encoding. In E. M. Fernández & H. Smith Cairns (eds.), *The Handbook of Psycholinguistics*, 38–81. Hoboken, NJ: John Wiley & Sons.

Liberman, Mark Y. 2019. Corpus Phonetics. *Annual Review of Linguistics* 5(1). 91–107.

Norcliffe, Elisabeth, Alice C. Harris & T. Florian 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.

Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS. In N. Calzolari, et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 3940–3947. Reykjavik: European Language Resources Association (ELRA).