Abstract:

Speech rate and pauses provide us with a window into the cognitive-neural and physiological-articulatory basesof the human language production system, but cross-linguistic variation in this domain remain understudied. This project fills this gap by comparative studies of spontaneously spoken language in a diverse sample of 50 languages. For this purpose, we create a multilingual reference corpus of language documentation data (DoReCo) consisting of annotations and associated audio recordings that are archived at repositories such as The Language Archive (TLA), especially from the DOBES collection. DoReCo will be built from data that are already transcribed, translated into a major language, and time-aligned at the level of discourse units with audio files. Within the current project, these data will be time-aligned at the phoneme level. We have identified at least 50 languages, from which corpora of at least 10,000 words can be included in DoReCo, and a subset of at least 30 of these, which are additionally already annotated for morpheme breaks and morpheme glosses. In DoReCo, subcorpora and annotations are treated as citable publications, provided with a permanent identifier and associated with a CC BY 4.0 license. DoReCo will have a lasting effect beyond the specific research goals of the DoReCo project, as a platform for easy access to over one million words of annotated corpus data from over 50 languages for cross-linguistic research on spoken language. This represents an unprecedented contribution to open, reproducible science regarding global linguistic diversity and cultural heritage. Both of DoReCo's two specific research goals address the universality of constraints on human language arising from species-wide articulatory and cognitive properties: Firstly [in Berlin], we investigate patterns of phonetic lengthening with the aim towards establishing universal vs. language-specific patterns in (i) the degree to which different types of phonological segments undergo variation in duration (e.g. vowels vs. different types of consonants)–reflecting articulatory and perceptual constraints–and (ii) word-final lengthening as indicative of major vs. minor prosodic boundaries–reflecting cognitive constraints on planning and potentially signalling discourse units. Secondly [in Lyon], we investigate universal vs. language-specific patterns in the temporal distribution of morphemes regarding (i) information rate in terms of morphemes per second and (ii) the number of morphemes in inter-pausal units–both reflecting cognitive constraints on language use. The project will be carried out by an interdisciplinary team bringing together expertise on documentary linguistics, phonetics, typology, and quantitative linguistics, with strong institutional support from two leading research centres in Germany and France.