

Silent pauses and interjections as cues marking out direct speech

A cross-linguistic study on twelve language documentation corpora

Extending earlier work on silent pauses as prosodic cues used for marking direct speech (Gentti 2011, Malibert & Vanhove 2015), this paper compares narrative speech in twelve areally and family diversified languages. Our investigation focuses on silent pauses and interjections and their coordination as speech devices marking out segments of speech as direct speech. The role of interjections as indicating the beginning of direct speech has already been reported for English (Norrick, 2015) and discussed during a typological workshop on direct speech (see <https://blogs.helsinki.fi/speech-representation/data-workshop-online-sessions/>), but a proper cross-linguistic study is still to be done. The data presented in this communication come from subsets of language documentation corpora representing monologic narrative texts in :

Beja, (beja1238, Afro-Asiatic, Vanhove 2021)	Bora (bora1263, Boran, Seifart 2021)
Nisvai (nisv1234, Austronesian/Oceanic, Aznar 2021)	Dolgan (dolg1241, Turkic, Arkhipov 2021)
Arapaho (Algonquian, Cowell 2021)	Movima (movi1243, Isolate, Haude, 2021)
Sanzhi (sanz1248, Caucasian/Daghestanian, Forker 2021)	Ruuli (ruul1235, Bantu, Witzlack-Makarevich, 2021)
Mojeño Trinitario (trin1278, Arawakan, Rose 2021)	Fanbyak (Austronesian/Oceanic, orko1234, Franjeh 2021)
Northern Alta (nort2875, Austronesian, Garcia-Laguia 2021)	Nuu (nngg1234, Tuu, Güldemann, Ernszt, Siegmund & Witzlack-Makarevich, 2021)

Our study strongly relies on multiple automatic processing programs, notably Multitool for parsing annotation files, Jupyter (Kluyver et al. 2016) for documenting the processing steps, and corpora2-corpus for compiling annotation files. By using Free software and open-access corpora, and by documenting the processing, we facilitate the repeatability and the reproducibility of the study.

After introducing our methodology, we produce a cross-corpora description of silent pauses, inter-pausal units, interjections and direct speech. This description provides a context for interpreting the occurrences of silent pauses within the direct speech sequences. Using visualization technics and statistics, we then describe and compare the forms the direct speech sequences can take in the corpora. We will see that, within this variety, the position of silent pauses and interjections used by speakers to mark the beginning or the end of the direct speech can be correlated with the typological characteristics of these languages. The recurrence of similar usages of interjections and silent pauses following typological characteristics is an argument supporting that silent pauses and interjections are linguistically motivated cues for the production of direct speech.

The study presented here has been developed within QUEST (QUALity - ESTablished, see <https://cutt.ly/quest>), an initiative funded by the German Federal Ministry of Education and Research (BMBF) to promote quality standards, curation criteria and methods of quality assurance for language corpora, and in cooperation with DoReCo (Paschen et al. 2020), which gather open-access

language documentation corpora and provides transcription and glossing feedback to the corpus creators in order to be able to produce corpora where transcriptions are at time-aligned.

References

- Arkhipov, A. 2021. Dolgan DoReCo data set. In Seifart et al.
- Aznar, J. 2021. Nisvai DoReCo data set. In Seifart et al.
- Cowell, A. 2021. Arapaho DoReCo data set. In Seifart et al.
- Forker, D. 2021, Sanzhi DoReCo dataset, in Seifart et al.
- Franjeh, M. 2021, Fanbyak DoReCo dataset, in Seifart et al.
- Garcia-Laguia, A. Northern Alta DoReCo dataset, in Seifart et al.
- Genetti, C.. 2011. Direct speech reports and the cline of prosodic integration in Dolakha Newar. *Himalayan Linguistics* 10(1). 55–76.
- Güldemann T. , Ernszt, M. , Siegmund S. , Witzlack-Makarevich, A. , 2021, Nuuk DoReCo dataset, in Seifart et al.
- Haude, K. 2021, Movima DoReCo dataset, in Seifart et al.
- Kluyver, T., B. Ragan-Kelley, F. Pérez, M. Bussonnier, J. Frederic, J. Hamrick, J. Grout, et al. 2016. ‘Jupyter Notebooks—a Publishing Format for Reproducible Computational Workflows’.
- Malibert, I.-I. & M. Vanhove. 2015. Quotative constructions and prosody in some Afroasiatic languages: Towards a typology. In A. Mettouchi, M. Vanhove & D. Caubet (eds.), *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages*, 117–169. Amsterdam: John Benjamins.
- Norrick, Neal R. 2015. « Interjections ». In *Corpus Pragmatics*. Cambridge: Cambridge University Press. www.cambridge.org/9781107015043.
- Paschen, L., F. Delafontaine, C. Draxler, S. Fuchs, M. Stave, & F. Seifart. 2020. ‘Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo)’. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* 12: 2657–66..
- Rose, F. 2021, Mojeño Trinitario DoReCo dataset, in Seifart et al.
- Seifart, F, L. Paschen & M. Stave (eds.). 2021. *Language Documentation Reference Corpus (DoReCo) 0.1*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
- Seifart, F. 2021. Bora DoReCo data set. In Seifart et al.
- Vanhove, M. 2021. Beja DoReCo data set. Originally annotated within the projects CorpAfroAs and CORPORA, reannotated within DoReCo. In Seifart et al.
- Witzlack, A. 2021 Ruuli DoReCo dataset, in Seifart et al.