Corpus phonetics for under-documented languages: a vowel harmony example

Corpus phonetics is transforming the analysis of acoustic and articulatory recordings in the same way that corpus linguistics has transformed the analysis of transcriptions. This "semiautomatic analysis of digital speech collections" (Liberman, 2019) is opening up new opportunities in phonetics and phonology and inspiring "a movement from the study of small, mostly artificial datasets to the analysis of published corpora of natural speech that are thousands of times larger" (Yuan, Lai, Cieri, & Liberman, 2018). Although many linguists want to use corpus phonetics in their own research, there are still barriers, such as the particular skills needed to use the software and some of the challenges in applying the tools to under-documented languages. In this paper, we develop a language-independent workflow and reveal the results of applying this technology to a vowel harmony and vowel reduction research problem in a recently documented language.

The reduction of vowels is a popular topic for research, but little has been said about the effects of vowel harmony on vowel reduction. In Kera (Chadic) it has been shown (Pearce, 2012), that not only is phonetic reduction linked to the phonetic duration of the vowel, but also that reduction is blocked in vowel harmony domains.

A pilot study of several other languages suggests this same hypothesis concerning vowel reduction is true for them too. For the study in Kera, the statistical results are compelling, but for the pilot study, which included several under-documented languages, the task of collecting enough data, applying the correct harmony rules and then measuring formants manually meant that the data were insufficient to give a confident result. A corpus phonetics approach allows us to test the hypothesis in several languages in a more robust manner. We start with Kera, where we know the expected results, to validate and develop the automated workflow.

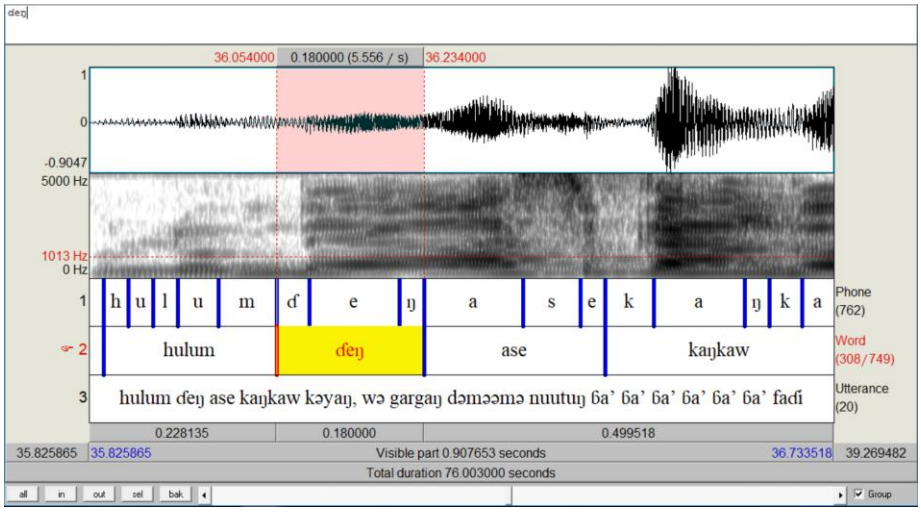

Figure 1: Automatic forced alignment of Kera displayed in Praat

Our approach to the workflow (with an indication of the free software used) is as follows:

- Initial analysis of corpus transcription (Phonology Assistant or visual inspection)

- Decision on whether to use target-language forced alignment, or cross-language forced alignment (our scripts or visual inspection)
- Forced alignment (Montreal forced aligner or Phnrec e.g. see Figure 1)
- Evaluation of forced alignment performance (our scripts)
- Phonetic queries returning acoustic information (EMU-SDMS and our scripts)
- Plots and statistical tests (R)

Figure 2: Formant plots of short duration /i/ and /u/ tokens in Kera. Left: tokens not occurring in height harmony domain. Right: tokens that occur in the height harmony domain.

In an effort to keep our research reproducible, we used a publically available corpus - a recording of the Kera New Testament. Speakers of the Kera language confirm that this contains natural sounding speech produced by fluent readers (Pearce, personal communication). The narrator contributed the most speech. Figure 2 shows a formant plot of this speaker for part of the corpus. Vowel tokens /i/ and /u/ that have a short duration (less than 50ms) are shown. The left graph shows tokens not occurring in the vowel height harmony domain, these show a normal pattern of reduction. The right graph shows tokens that do occur in the vowel height harmony domain. These show a different distribution and tend to preserve their F1 value. All our results are derived from automatic measurements; the only data provided was the recordings and the transcription. The results are in agreement with Pearce (2012).

This confirms the hypothesis for Kera, and allows us to have confidence in the procedure to test other languages. We also expect that our corpus phonetics workflow will be helpful to other phonologists who wish to test larger data collections.

References

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

Pearce, M. D. (2012). Effects of harmony on reduction in Kera. *Linguistic Variation*, *12*(2), 292–320.

Yuan, J., Lai, W., Cieri, C., & Liberman, M. (2018). Using Forced Alignment for Phonetics Research. *Chinese Language Resources and Processing: Text, Speech and Language Technology. Springer*.