Phonological Corpus Tools

Kathleen Currie Hall, Blake Allen, Michael Fry, Scott Mackie, Michael McAuliffe University of British Columbia

In this presentation, we introduce Phonological Corpus Tools (PCT), a free, open-source tool for doing phonological analysis on transcribed corpora. There is an ever-increasing interest in exploring the roles of frequency and usage in understanding phonological phenomena (e.g., Bybee 2001, Ernestus 2011, Frisch 2011), and corpora of language give us a way of making generalizations across wide swaths of such usage, exploring patterns in under-documented languages, and creating balanced stimuli in experiments. Many corpora and existing corpus-analysis software tools, however, are focused on dialogue- and sentence-level analysis, and/or the computational skills needed to efficiently handle large corpora can be daunting to learn. PCT is designed with the phonologist in mind and has an easy-to-use graphical user interface that requires no programming knowledge. It specifically includes the following capabilities:

- Summary descriptions of a corpus, including type and token frequency of individual segments in user-defined environments;
- Ability to analyse a corpus with any set of phonological features (e.g., Hayes 2009; Mielke 2008). Phonological tiers may be extracted based on features.
- Calculation of phonotactic probability of a word, given the other words in the corpus (cf. Vitevitch & Luce 2004);
- Calculation of functional load of individual pairs of sounds, defined at either the segment or feature level (cf. Surendran & Niyogi 2003; Wedel, Kaplan, & Jackson 2012);
- Calculation of the extent to which any pair of sounds is predictably distributed given a set of environments that they can occur in (cf. Hall 2009, Hall & Hall 2013);
- Calculation of the Kullback-Leibler divergence between the distributions of two sounds (cf. Kullback & Leibler 1951; Peperkamp et al. 2006)
- Calculation of the extent to which pairs of words are similar to each other and calculation of neighbourhood density (cf. Frisch et al. 2004, Khorsi 2012; Greenberg & Jenkins 1964; Luce & Pisoni 1998; Yao 2011);
- Calculation of the mutual information between pairs of segments in the corpus (cf. Brent 1999; Goldsmith & Riggle 2012);
- Calculation of the acoustic similarity between .wav files, based on alignment of MFCCs (cf. Mielke 2012) or amplitude envelopes (cf. Lewandowski 2012).

The software can make use of pre-existing freely available corpora (e.g., the IPHOD corpus (Vaden et al. 2009)), or a user may upload his or her own corpus from a variety of formats. First, lexical lists with transcription and frequency information can be directly uploaded. Second, raw running text (orthographically and/or phonetically transcribed) can be uploaded and turned into lexical lists in columnar format for subsequent analysis. Running text can also be extracted from Praat TextGrids (Boersma & Weenink 2014), and raw sound files may be uploaded for analyses of phonetic similarity. All analyses are presented on screen and saved to plain .txt files, and the software is fully documented.

PCT works on any platform that supports Python (Windows, Mac, Linux). As part of the presentation, we will provide the motivations for building the software, information on accessing it, and a brief demonstration of its utility. More information (including all references) can be found at http://kchall.github.io/CorpusTools/.