Quantitative Phonological Comparisons of Corpora Using Cross-Language Categories

Edward Jahn

George Mason University, Fairfax VA U (Graduate Student)

ejahn3141@gmail.com

This paper describes algorithms and data structures to assign measurements to language corpora, in a way that allows direct comparison across languages. Viewing the measurements as points in a multi-dimensional space, the distance between them is a measure of how different the corresponding corpora are from each other. Points that cluster together are typologically similar in a quantifiable way. This method is applied to phoneme frequency distributions. By partially overcoming the difficulty of comparing phonemes that are phonetically different, it enables comparisons across a wide range of languages.

Each phoneme in a corpus is mapped to a set of categories that represent phonetic characteristics (voice, manner, place, etc) that exist in all the languages to be compared. The categories form a vector $\boldsymbol{c}$*, defined so that each* $\boldsymbol{c}_i$ *is either 0 or 1. The frequency* f *of each phoneme is multiplied across* $\boldsymbol{c}$ *to give a vector* $\boldsymbol{p}$*, each member of which is equal either to* f *or to 0. The vectors* $\boldsymbol{p}$ *of all the phonemes are added to give the measure* $\boldsymbol{m}$ *of the corpus. The Euclidian distances between the measures of multiple corpora are then computed. Descriptions of the algorithms and data structures are provided, such that they can be implemented in different kinds of computer software.*

Tests were done on 35 languages from a wide range of geographic areas and genetic classifications. Phoneme frequency distribution data were partly obtained from the literature, and partly computed from texts collected from the Internet. The categories used represented articulatory gestures. Within this sample, distances tend to be smaller between languages that are closely related genetically, and also between languages that belong to known Sprachbünde; but genetic and areal effects account for only a small part of the variation in distance. These results are consistent with reasonable assumptions about phoneme frequencies. Typological information may be gathered from analysis of the remaining variation, but a much larger sample will be needed before firm conclusions are possible.

This method is capable of being generalized to any language characteristics that can (a) be represented by a frequency distribution, and (b) be classified into cross-language categories. It can handle phoneme sequences as well as individual phonemes. Phonological categories that can be used include distinctive features and acoustic parameters. The method could also potentially be used for cross-language studies of morphology or syntax.