An algorithm for learning phonological classes from distributional data

Connor Mayer
UCLA

How much information about phonological classes can be found in the distribution of sounds in a language? Classic work suggests that there is a universal tendency for learners to group sounds based on their phonetic properties (e.g. Chomsky & Halle, 1968), but the presence of phonetically disparate classes and variable phonological classification of the same sound across languages (e.g. /l/ as [+/- continuant]) has led some researchers to propose that distributional learning plays a greater role than previously thought (e.g. Mielke, 2008).

Following work such as Goldsmith and Xanthos (2009), this project describes an algorithm that learns phonological classes from only the distribution of the sounds in a language: that is, from the contexts in which they do and do not occur. This is not to suggest that other information sources, such as phonetics and allophony, do not play an important role in characterizing phonological classes. Rather, it is an attempt to see how far learning can proceed when restricted to just one of the many available sources of information.

Because it is not clear a priori what classes will be recoverable from the distribution of sounds in a natural language, I first test the algorithm on a simple CV toy language, *Parupa*, which has strict phonotactic constraints that refer to the class structure illustrated in Figure 1. Constraints include word-level vowel backness harmony, restrictions on word-initial consonants, and restrictions on which consonants and vowels may co-occur in a syllable.

The algorithm has two stages. First, sounds in a corpus are projected into a vector space based on their distributional properties: specifically, the counts of the trigram contexts in which they occur. These counts are weighted using *positive pointwise mutual information*, an information theoretic measure that emphasizes contexts in which sounds occur more frequently than chance. A visualization of the sounds in Parupa is shown in Figure 2.

Next, principal component analysis (PCA) is carried out on the vector embedding of the sounds. PCA projects a matrix onto a set of new, uncorrelated dimensions called principal components (PC). Clusters are then retrieved by doing optimal k-means clustering on some of the PCs, and recursively applying the same process to the discovered classes. Because PCs are associated with the amount of variance they capture, the algorithm can be parameterized to only attempt to cluster based on the more robust patterns in the data.

This algorithm successfully retrieves all of the classes in Parupa, showing that, to the extent phonological classes are apparent in the distribution, it can detect both overlapping and nested classes. When applied to Samoan, English, and French corpora, it successfully differentiates consonants and vowels, and is able to identify some more detailed aspects of phonological structure, although the three languages differ in which phonological classes are most salient. In French, for example, the class of consonants is successfully partitioned into non-approximants, liquids, rounded glides, and unrounded glides.

These results are encouraging given the paucity of the input data, and suggest that further refined models of distributional learning may be able to effectively learn at least some phonological classes. I will discuss limitations of the current implementation and possibilities for future research on distributional learning.

References

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
Goldsmith, J., & Xanthos, A. (2009). Learning phonological categories. *Language*, *85*, 4-38.
Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.

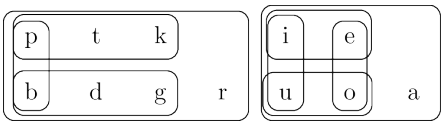

Figure 1: The phonological classes of the toy language Parupa

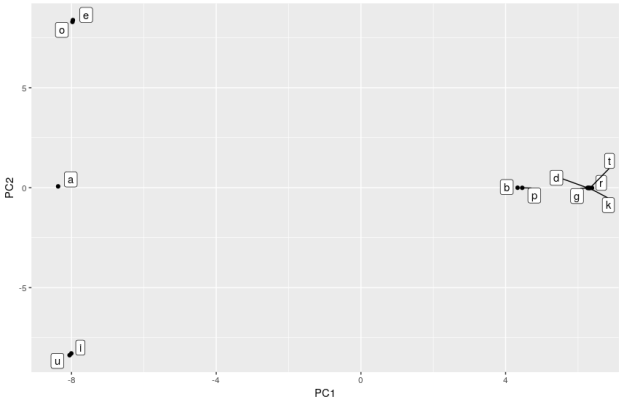

Figure 2: PCA visualizations of the vector embeddings of Parupa

.

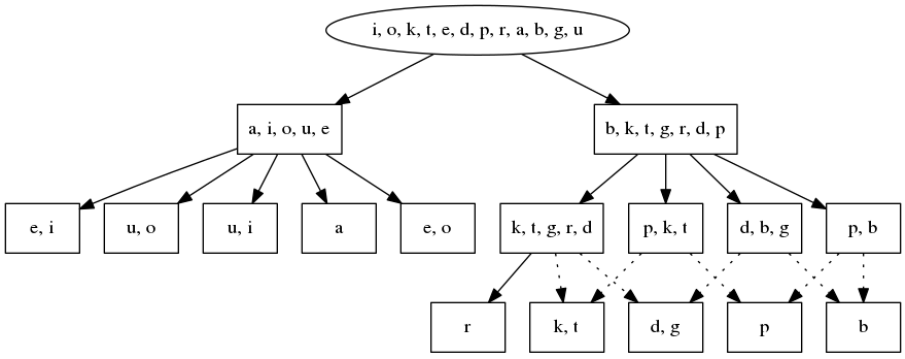

Figure 3: The class structure learned from Parupa. The learned classes not shown in Figure 1 are either complements or intersections of the expected classes.