Decomposing and composing phonemes: a distributional approach to phonotactics

Basilio Calderone

The Distributional Hypothesis (DH) claims that words that occur in the same distributional contexts tend to have similar meanings. In particular, recent research on Distributional Semantics focuses on modeling the correlation between distributional similarity and similarity of meaning. Starting from this assumption, we try to transfer the DH into the domain of phonology in order to test the possible correlation between distributional similarity and phonotactic/phonological similarity. The underlying idea of our approach is to model the behavior of phonemes inductively by appealing solely to the phonotactic statistical regularities of the language and general distributional considerations.

Methodologically, our analysis exploits vector-based representations of phonemes based on their distributional properties in large phonologically-transcribed corpora. By manipulating these high-dimensional vectors, we can characterize some (basic) phonological issues such as the vowel/consonant distinction, the identification of natural phonological classes and more complex aspects, such as the composition of consonant clusters.