Natural phonological classes as phonotactic matrices

Bernard Laks[1] and Basilio Calderone[2]

[1]CNRS & Université Paris Ouest Nanterre La Défense
[2]CNRS & Université de Toulouse- Le Mirail

How is it possible to infer linguistic structures (such as phonemes, morphemes, syllables) starting from the phonotactics of a given language? In other words, how could some theoretical linguistic notions find a genuine data-based counterpart and be inductively modeled in terms of phonotactic constraints and statistical regularities without any external (linguistic) supervision?

The goal of this research is to extract, in an unsupervised fashion, phonologically-motivated features exclusively out of the distributional information, i.e. co-occurrence matrices with variable 'phonotactic windows'.

The approach is based on the assumption that invariant regularities in a language (such as phonotactic regularities) are meaningful enough to derive universal generalizations concerning basic phonological distinctions, without an a priori characterization (in computational terms) and interpretation (in linguistic terms) of what vowels and consonants are.
Evidences come from English, Finnish, Italian, and French.