Are Abstract Phonetic Categories Encoded in Probabilistic Phonotactics?

Eleonora Albano & Antonio Pessotti

State University of Campinas

albano@unicamp.br, antoniopessotti@gmail.com

Induction of phonetic categories is challenged by evidence of lack of acoustic phonetic invariance in the speech signal, even in relatively concrete place or manner of articulation classes (e.g., coronals, dorsals, etc., or fricatives, rhotics, etc.). Invariance is still more unlikely in highly abstract categories, such as the so-called major classes: obstruents, sonorants, liquids, and glides. However, abstract phonetic categories may act as: (1) targets of phonological processes, e.g., liquid devoicing in English, as in 'clash' and 'crash'; (2) triggers of phonological processes, e. g, pre-sonorant voicing in West Flemish, as in /zɛs ja:r/ [zɛz ja:r]/ 'six years'. Hence, they are cognitively relevant for speakers/hearers. The ensuing language cognition question is: *can abstract phonetic categories be bootstrapped from information in the signal other than invariance*?

Phonotactic bootstrapping has been proposed for word segmentation, which also challenges induction. Infants are sensitive to phonotactic probabilities (Jusczyk, Luce, & Charles-Luce, 1994). Phonotactically-based computer simulations have been successful in extracting words from running unsegmented text (Adriaans, 2011). However, there has been no attempt to date to explore encoding of abstract phonetic categories by probabilistic phonotactics.

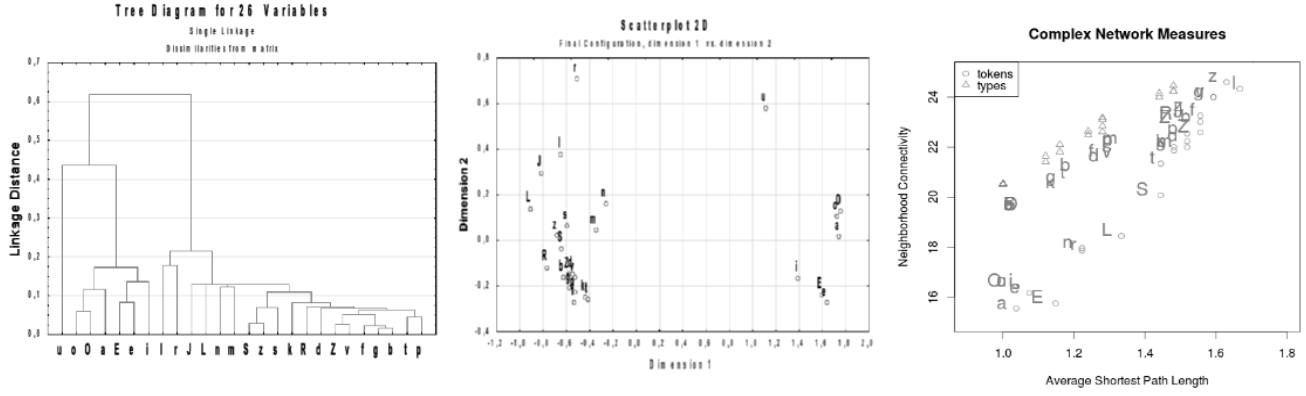
A first aim of this paper is to explore phonotactic probabilities as one of the possible bootstrapping mechanisms for abstract phonetic categories. Another aim is to compare, to this end, the two sources of phonotactic probabilities, viz.: token frequencies in corpora and type frequencies in lexicons. A third aim is to compare multivariate exploratory statistics with complex network modeling as a means of grouping phonemes into abstract phonetic categories.

The phonemic analysis consists of 19 Cs and 7 Vs. The consonant inventory is as follows: (a) labials: /p, b, f, v, m/; (b) coronals: /t, d, s, z, n, l, r/; (c) post-alveolars: /ʃ, ʒ, ʧ, ʤ, /ʃ, ʒ, ʧ, ʤ, j/; (d) dorsals: /k, ɡ, ɾ/. No archiphonemes are posited: coda allophones are assigned to /s, l, r, n/. Glides are counted as vowels. The vowel inventory is as follows: (a) front vowels: /i, e, ɛ/; (b) back vowels: /u, o, ɔ, a/. Nasal Vs are assigned to /Vn/. Reduced post-stressed Vs are assigned to /i, a, u/. Stress is assigned phonemic status, but ignored in cooccurrence counts.

Multidimensional scaling (MDS), cluster analysis (CA) and two complex network measures (neighborhood connectivity and shortest average path length) were calculated using 1-Spearman's R as a distance measure. Spearman's R was derived from confusion matrices containing the co-occurrence frequencies of all phoneme pairs.

A graphic overview of the results is given by Figure 1.

Figure 1: MDS, CA and two complex network measures (neighborhood connectivity and shortest average path length).



MDS and CA tend to consistently produce some very basic manner of articulation groupings, such as C and V, or aproximants vs. obstruents plus nasals. Other subgroups may also emerge, but they tend to be incomplete, e.g., sibilants. Complex network measures, in turn, not only replicate the C/V split but also point to some place of articulation groupings, such as coronals. Moreover, types and tokens yield very similar grouping patterns, regardless of statistical technique.

It should be stressed that no classificatory information other than the phonemic analysis was fed to the statistical software. The observed groupings are entirely based on distance measures directly derived from phonotactic probabilities. Consequently, these results, however exploratory, suggest that the idea of extracting phonetic classification from co-occurrence frequencies is worth pursuing.

References

Adriaans, F. W. 2011. *The Induction of Phonotactics for Speech Segmentation: Converging evidence from computational and human learners.* Unpublished doctoral dissertation, Utrecht Institute of Linguistics.

Costa, L. da F.; Oliveira, O. N.; Travieso, G., Rodrigues; F. A., Villas Boas, P. R.; Antiqueira, L.; Viana, M. P.; Rocha, L. E. C. 2011. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3), 329-412.

Jusczyk, P. W.; Luce, P. A.; & Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630-645.

LAEL http://www2.lael.pucsp.br/corpora/.

***