Similarity scores within adjacent phoneme pairs show a weakly bimodal distribution

Ignas Rudaitis • Vilnius University • 8 June 2020
Submission to Academia Salensis XVII

If one picks two phonemes that stand next to each other in a word, are they likely to be similar? Dissimilar? Is some middle ground to be expected instead?

(Hulden, 2017) refers to an array of cross-linguistic evidence for the avoidance of pairs of segments that are *both* highly similar *and* placed adjacently. This phenomenon seems to be best known by the name of "obligatory contour principle" (OCP), originally proposed by (Leben, 1973) for abstract tonal features.

Thus, OCP predicts that two adjacent phonemes are more likely to be dissimilar.

In 2016, the PanPhon software package was made available by (Mortensen, et al., 2016). It made IPA symbols machine-readable, including easy decomposition into an approximate system of phonetic features. This year, (Lee, et al., 2020) introduced WikiPron, an open effort for collecting multilingual phonemic transcriptions *en masse* from Wiktionary.

Combining the two tools, the present author has attempted to quantify OCP. For each language in a sample of size 27, every pair of adjacent segments was scored for matching non-null features. In Figure 1, the horizontal (x) axis is marked for counts of matching features, and the heights of the bars indicate relative frequencies of the counts.

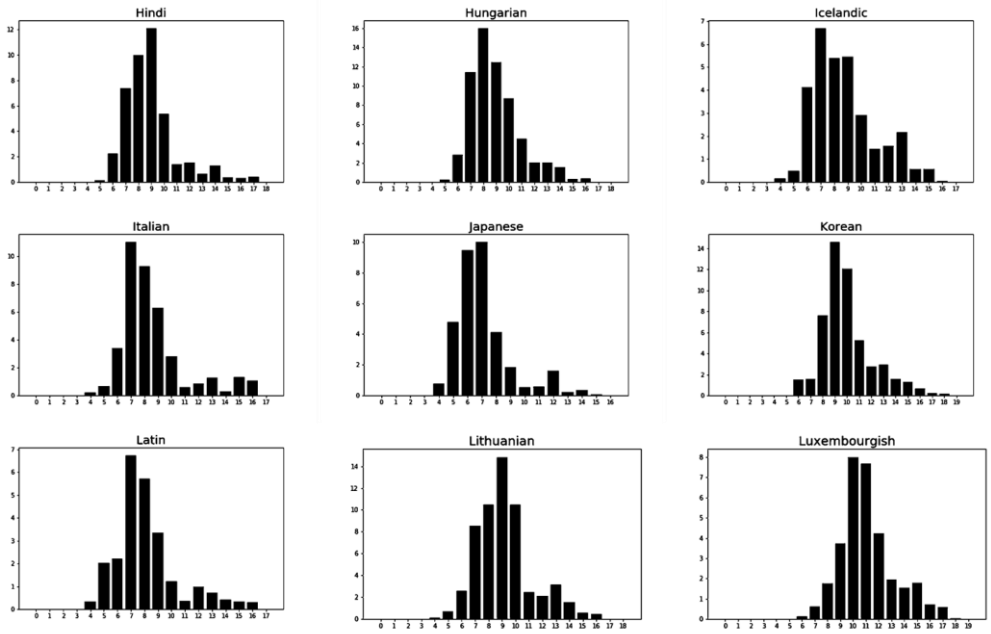

Figure 1. Relative frequencies of matching feature counts

The charts, of which 9 are shown in Figure 1, show a clear semblance to skewed normal (bell-shaped) distributions. The maximum is skewed leftwards, supporting OCP.

We bear in mind that the features that PanPhon assigns to the segments are language-agnostic and thus imperfect for this purpose, and a replacement would be necessary, should one intend to proceed with the study.

However, one cannot help but notice a very slight but almost ubiquitous "anti-OCP" in the charts. Cross-linguistically, the skewed normal distribution will very often jump up just a little when it has already dwindled to a local minimum. This way, the distribution has two peaks, or, in other words, is *bimodal*.

This is a highly preliminary finding, but if the effect were to be demonstrated more reliably, it would shed new light on assimilation and dissimilation, and perhaps some other phenomena.

Namely, the bimodal nature of the distribution would mean that while very dissimilar phonemes are the best choice for adjacent phonological time slots, the second-best choice is, in fact, very similar phonemes. It is phonemes of average similarity that are then the worst choice.

This would construe assimilation and dissimilation not as competing phenomena, but as different ways to achieve the same goal, which is to avoid the adjacency of *medium-similarity* phonemes. Given how both assimilation and dissimilation often target natural classes of phonemes (already implying *some* pre-existing similarity!), and not the entirety of the inventory, this does not seem impossible.

References

Hulden, M. (2017). A phoneme clustering algorithm based on the obligatory contour principle. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 290-300). Vancouver, Canada: ACL.

Leben, W. R. (1973). *Suprasegmental Phonology.* Massachusetts Institute of Technology.

Lee, J. L., Ashby, L. F., Garza, M., Lee-Sikka, Y., Miller, S., Wong, A., . . . Gorman, K. (2020). Massively Multilingual Pronunciation Modeling with WikiPron. *Proceedings of The 12th Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. S. (2016). PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3475-3484). ACL.