Discovering new vowel harmony patterns using a pairwise statistical model

Nathan Sanders and K. David Harrison (Swarthmore College)

20th Manchester Phonology Meeting, 24–26 May 2012

Goals

Define a statistical measurement of vowel harmony over a corpus, in such a way that it can be meaningfully compared across corpora, languages, and harmonic features.

Methods

(1) Extract tier-adjacent vowel pairs from the corpus. For example, if the corpus contains the word *krematoryum*, the algorithm extracts *ea*, *ao*, and *ou*. For a given feature, like backness, compute the overall harmony percentage: the number of harmonic pairs divided by the total number of pairs. In this case, 2 out of 3 pairs are harmonic (*ao*, *ou*), so the percentage is 67%.

(2) Using the total size of the corpus, the distributions of the vowels, and the distribution of word lengths, randomly generate a large number of corpora, and calculate the harmony percentage for each.

(3) Compare the harmony of the original corpus to the distribution of harmony in the random corpora. The number of standard deviations the real corpus is from the mean of the random corpora is its *z*-score for harmony.

Benefits

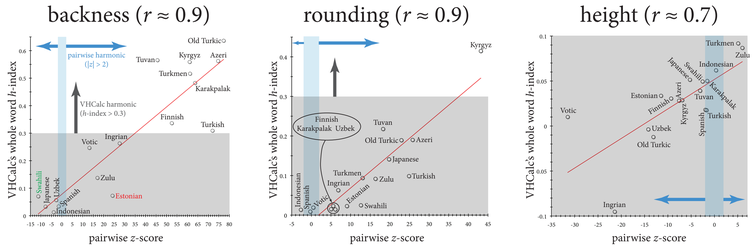
The *z*-score is a normalized measure of statistical deviance, so it can be meaningfully compared from one case to any other, like the relative harmony between languages.

Measures of whole-word harmony, like the *h*-index of Harrison et al.'s (2002–2004) Vowel Harmony Calculator (VHCalc), do not distinquish between different levels of disharmony, as in *krematoryum* versus *eskavatör* (both are categorically disharmonic), which contribute differently to the calculation of the *z*-score (2 harmonic pairs vs. 1).

Comparison with VHCalc

Despite the different underlying mathematics, the pairwise *z*-score and VHCalc's whole-word *h*-index are strongly correlated, however there are still important differences:

backness ($r \approx 0.9$) rounding ($r \approx 0.9$) height ($r \approx 0.7$)



White regions are harmonic for both VHCalc and *z*-score (*h*-index > 0.3, |*z*-score| > 2). Gray regions are unharmonic for VHCalc, and blue regions are unharmonic for *z*-score.

New Harmony Patterns

First, there are bizarre cases of "**anti-harmony**", where a language has a statistically significant negative *z*-score. For example, Swahili's pairwise backness harmony has a *z*-score of –11, well beyond the –2 needed for statistical significance. It's not clear what anti-harmony might be…

More importantly, many languages do not have harmony at the word level, but do have a large amount of pairwise harmony, and thus, have "**hidden harmony**". For example, Estonian's *h*-index for backness harmony is 0.07, but it is very harmonic for vowel pairs ($z = 24$). In this case, the discrepancy is due to historical harmony that has left remnants in the lexicon. Thus, hidden harmony could be a diagnostic tool for reconstructing historical harmony.

Hidden harmony may also have implications for the learnability of harmony or for development and/or loss of harmony over time.

References

Harrison, K.D., E. Thomforde, & M. O'Keefe. 2002–2004. The Vowel Harmony Calculator. `http://www.swarthmore.edu/SocSci/harmony/public_html/index.html`