Discovering New Vowel Harmony Patterns Using a Pairwise Statistical Model

Nathan Sanders and K. David Harrison, Swarthmore College
{nsander1,dharris2}@swarthmore.edu

1 Introduction

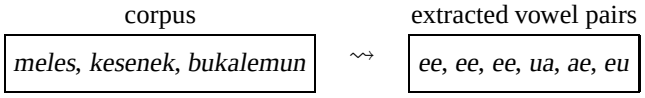
The goal of this project is to define a basic statistical measure of vowel harmony over an arbitrary corpus, such that this measure can be used to meaningfully compare the relative harmony between any languages, corpora, or phonological features. For example, we might want to know whether Finnish is more harmonic for backness than Hungarian is, or whether Tuvan is more harmonic for backness than for roundness, or whether Turkish is more harmonic for backness in literary writing than in academic writing. With such a measure of vowel harmony applied to the appropriate temporally spaced corpora, we could even determine the quantitative trajectory of a language's harmony over time: when and how fast it increased or decreased.

2 Methodology

Vowel harmony is typically taken to be a mostly categorical phenomenon: a given language either has harmony, or it does not; a given vowel in a given environment either harmonizes, or it does not. However, a more gradient, statistical measure of harmony could reveal more fine-grained information and trends that may prove useful in modern phonological analyses that are sensitive to non-categorical patterns. Thus, we propose a measure of vowel harmony that uses a very low-level domain of harmony: tier-adjacent vowel pairs. This is a much smaller domain than the word, which is typically used for computing vowel harmony (e.g., as with Harrison et al.'s (2002–2004) Vowel Harmony Calculator, henceforth VHCalc).

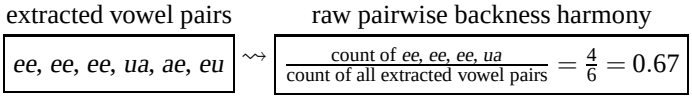
The algorithm we use is given in (1)–(5) below. A corpus is any list of words, using any notation that uniquely specifies every phonemic vowel contrast, such as the IPA or transparent orthographies like those used for Finnish and Turkish. For simplicity, diphthongs are taken to be sequences of vowel phonemes, vowel features such as backness and roundness are taken to be binary (central vowel are classified as back), and the possibility of vowel neutrality is ignored. For measurements of height harmony, vowels were split into high and non-high (i.e., mid plus low). These choices are for demonstration purposes only; this algorithm can easily be modified to accommodate multi-valued features, different categorizations (low versus non-low, tense versus lax, etc.), and neutral vowels (by classifying them as part of multiple categories).

(1) Extract all tier-adjacent vowel pairs for the entire corpus. For example, if the corpus consists of the three words *meles*, *kesenek*, and *bukalemun*, the algorithm would extract the six vowel pairs *ee*, *ee*, *ee*, *ua*, *ae*, and *eu*:



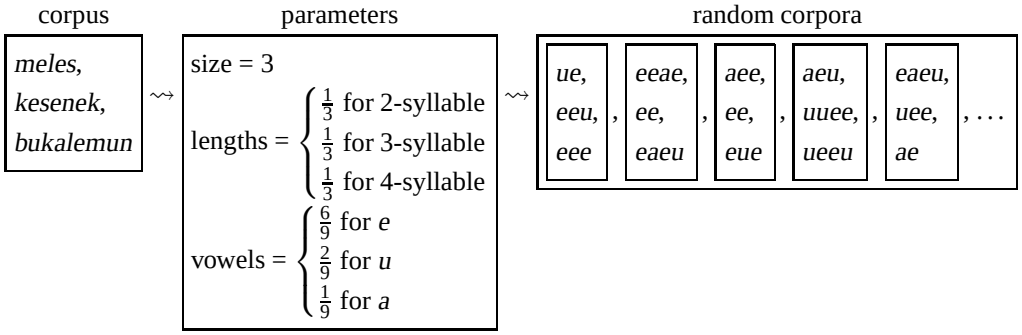
Note the repetition of *ee*, which counts separately for each occurence in the corpus.

(2) Compute the raw pairwise harmony for a given feature by dividing the total number of vowel pairs that are harmonic for that feature by the total number of vowel pairs in the corpus. In the current example, the raw backness harmony would be about 0.67, since there are four vowel pairs that are harmonic for backness (*ee*, *ee*, *ei*, *ua*) out of six total vowel pairs:



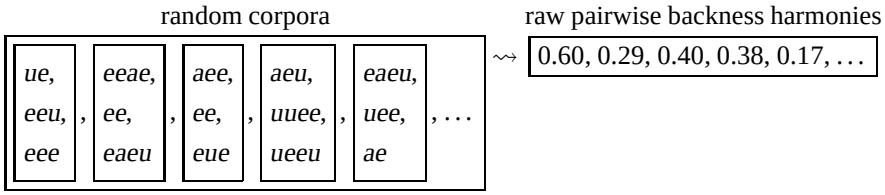
Note that *ee* and *ua* each count as harmonic for backness, because the vowels within each pair both have the same backness, even though one pair is front and the other is back. That is, harmony (or lack of harmony) is only measured within a vowel pair, not between different vowel pairs.

(3) Using the original corpus's size, distribution of word lengths, and frequencies of each individual vowel, construct a large number of corpora based on those parameters, but with the vowels randomly distributed. In the current example, the corpus size is 3; each word has an equal (1/3) chance of having 2, 3, or 4 syllables; and the individual vowel frequencies are 6/9 for *e*, 2/9 for *u*, and 1/9 for *a*:

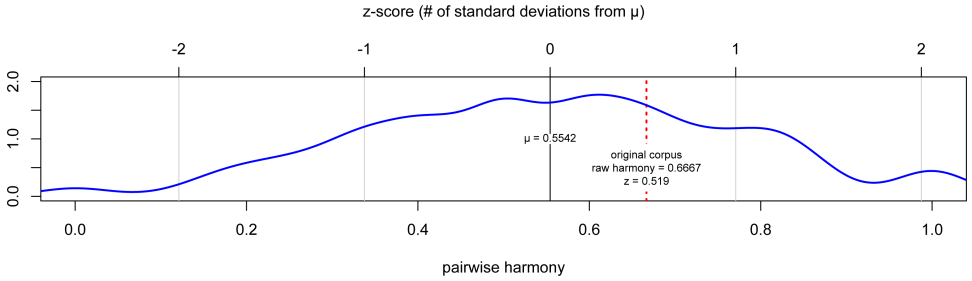


Note that the random corpora do not have consonants, since consonants are ignored for computing vowel harmony. Also note that the individual random corpora do not need to have the exact same distribution of word lengths or vowels as the original corpus, since these parameters represent probabilities for the random corpora, not firm restrictions. Corpus size, however, is held constant between the original and the random corpora.

(4) For each random corpus, compute its raw pairwise harmony for a given feature, using the same calculation as in steps (1)–(2):



(5) Finally, assume that the raw pairwise harmonies for the random corpora (plotted as the blue curve below) follow a normal distribution, and compute how far away (in standard deviations, plotted as the thin vertical grey lines) the raw pairwise harmony of the original corpus is from the mean of this distribution. In the current example, after generating 2000 random corpora, the mean raw pairwise harmony comes out to about 0.55 (plotted as the thin vertical black line), and the standard deviation comes out to about 0.22. The original corpus has a raw pairwise harmony of 0.67 (plotted as the vertical dotted red line), which is about half of a standard deviation away from the mean, which means that the original corpus has a *z*-score of about 0.5:



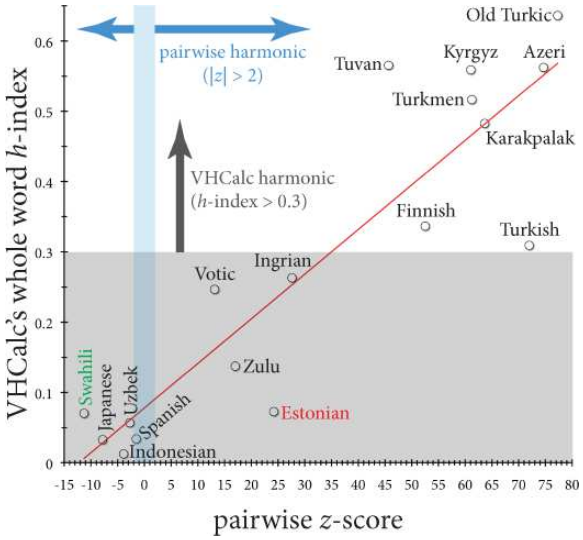
The critical values (for 95% confidence) for the *z*-score are about -2 and 2. If a language has a *z*-score below -2 or above 2, then its raw pairwise harmony is statistically significantly different from what we might expect by random chance. In the current example, such a difference would require the original corpus to have a raw pairwise harmony greater than about 0.98 (which means all nine of its vowel pairs would have to be harmonic), or less than about 0.12 (no more than one of its nine vowel pairs could be harmonic). Any other amount of pairwise harmony falls within two standard deviations (i.e., $-2 < z < 2$), and thus, could be reasonably expected to occur just by random chance.

Because the *z*-score is a normalized measure of statistical deviance, it can be meaningfully compared between multiple cases. Thus, we can use it to compare the relative amount of pairwise harmony between two languages, between two features, or between two corpora. The main drawback from this method is the need to generate random corpora and evaluate their raw pairwise harmony; this can take a few hours if the original corpus is large.

3 Results

In (6)–(8), the *z*-scores obtained from the algorithm outlined in (1)–(5) are directly compared to the *h*-index of VHCalc, which computes the harmony of a corpus based on how many words are completely harmonic. The corpora used are for 17 languages with downloadable corpora from the VHCalc website: Azeri, Estonian, Finnish, Indonesian, Ingrian, Japanese, Karapalak, Kyrgyz, Old Turkic, Spanish, Swahili, Turkish, Turkmen, Tuvan, Uzbek, Votic, and Zulu.

(6) *z*-score versus *h*-index for backness (front versus non-front)



The *z*-score and *h*-index are unsurprisingly correlated very strongly ($r \approx 0.9$; $r = 1.0$ is perfect correlation): a corpus with a large number of harmonic words will obviously have a large number of harmonic vowel pairs. The regression line between these measures of harmony for backness is plotted in (6) as the upward sloping diagonal red line.

The threshold for harmony for VHCalc's whole-word *h*-index is 0.3; an *h*-index below that level indicates a language with very little whole-word harmony. The grey region in the lower portion of (6) indicates where the *h*-index is below 0.3. The thresholds for harmony for our pairwise *z*-score are -2.0 and 2.0; a *z*-score within that range indicates a language with no statistically significant pairwise harmony pattern. The vertical light blue region in (6) shows where $-2 < z < 2$.

There are two notable results from this comparison. First, there are cases of **"hidden harmony"**, where a language is unharmonic according to VHCalc's *h*-index, but still has a large amount of pairwise vowel harmony according to our *z*-score. Estonian (shown in red in (6)) has an *h*-index of about 0.07 (much less than the 0.3 threshold), but a *z*-score of about 24 (well above the 2.0 threshold). This is likely due to historical vowel harmony that Estonian lost over time, leaving behind harmonic residue in the form of a strong statistical tendency for pairwise vowel harmony.

Second, there are cases of **"anti-harmony"**, where a language has a *z*-score below -2. Swahili (shown in green in (6)) has a *z*-score of about -11, which is clearly statistically significant. This level of anti-harmony indicates that vowels within a word tend to alternate for the harmonic feature, more so than would be expected by random chance. This is a statistically important result: if a coin is flipped, we expect to see about 50% heads and 50% tails, but we don't expect to see the heads and tails perfectly interleaved, with each head followed by a tail. Randomness is streaky, and there should a large number of times when multiple coin flips in a row have the same outcome. Any significant deviation towards perfect alternation is indicative of some external factor influencing the result.

Thus, for anti-harmony, something must be driving the language towards an alternating pattern. For Swahili, it's not clear what this might be, but for moderately anti-harmonic Uzbek ($z \approx -2.7$), there is an explanation. Like Estonian, Uzbek used to have historical vowel harmony, but lost it over time. But not only did it lose vowel harmony, it also underwent a collapse of the vowel system, in which the front vowels *y* and *ø* backed, merging with *u* and *o*. The result is that a a word that used to be fully front, but had a number of instances of *y* or *ø*, would now be pronounced with a large amount of pairwise disharmony. For example, the historical string of vowels *eyie* would now be pronounced as *euie*, drastically lowering the raw pairwise harmony from 1.0 (fully harmonic) to 0.33. This reduction in pairwise harmony would happen for every polysyllabic word that originally contained *y* or *ø*, so it's clear why the combination of historical harmony for a given feature and a vowel merger across that harmonic feature could result in statistical anti-harmony.

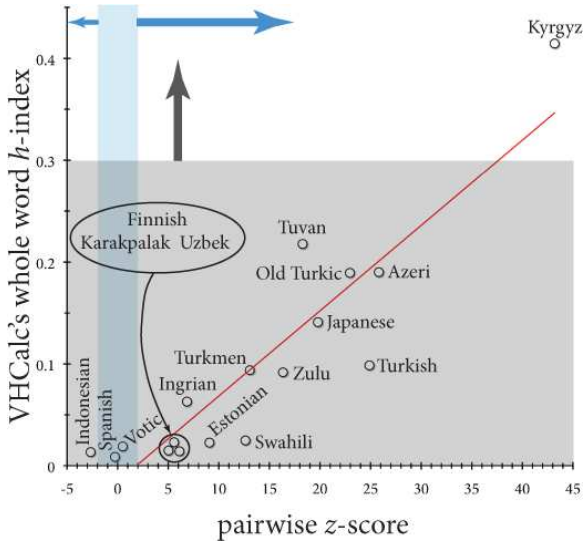
From both of these cases, we can see that the pairwise *z*-score could be used as a preliminary diagnostic for historical harmony. This could be valuable in trying to figure out the genetic relationship between languages, and perhaps even the relative timing of certain historical processes. One intriguing possibility is that comparing the *z*-scores of many harmonic languages across a range of time periods, we may be able to uncover some statistical predictors for the emergence and/or death of vowel harmony within a language's timeline.

Furthermore, given that speakers are known to be sensitive to statistical trends in their language (frequency effects, etc.), it's also reasonable that the *z*-score may be able to make some predictions about certain aspects of acquisition, priming, etc., related to vowel harmony.

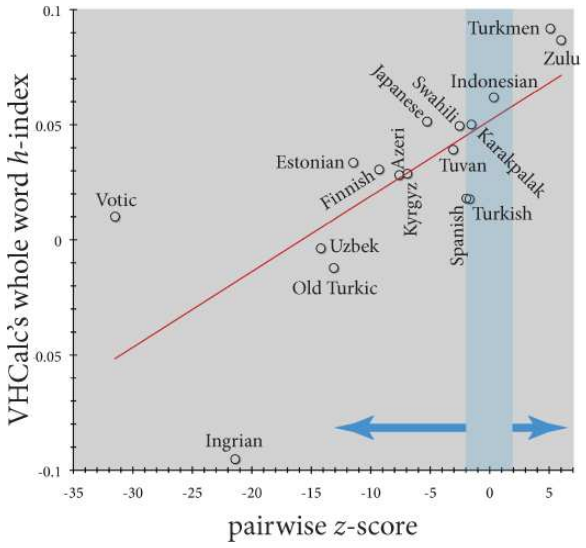
However, the potential uses for this measure of vowel harmony are still very much hypothetical and untested. Further comparisons across large numbers of corpora need to be done.

In (7) and (8), the *z*-score and *h*-index are compared for rounding harmony and height harmony, as in (6); again, there is a strong correlation ($r \approx 0.9$ and $r \approx 0.7$, respectively):

(7) *z*-score versus *h*-index for rounding (round versus non-round)



(8) *z*-score versus *h*-index for height (high versus non-high)



4 Limitations

This methodology yields a very basic measure of pairwise vowel harmony, so it has some inherent limitations that need to be overcome in an expanded model. Most importantly, position within a word is not taken into account. So, while *i* might have an overall frequency of 20% in a corpus, it may occur in the first syllable more frequently than in other syllables, and this discrepancy may grow stronger as word length increases. If these restrictions are independent of vowel harmony effects, then they could easily skew the calculation of the *z*-score. A more refined algorithm would thus need to compute the distributional frequencies of individual vowels relativized to position.

Another drawback is that our algorithm relies on text corpora and does not take into account variability in actual articulation. For example, various words in a corpus may have *u* in them, but some of these may be regularly pronounced without lip rounding. Even worse, there could even be statistically significant changes in the articulation of rounding that cannot be seen in a simple categorical division. An instance of a rounded *u* within a word containing many unrounded vowels may tend to be articulated with somewhat less rounding than within a word containing many round vowels, but both could still be categorically classified as rounded.

We also do not take into account any sort of morphological structure. While this may be a desirable feature when thinking about child acquisition (early acquisition is ordinarily blind to the internal morphemic structure of words), it may not be the best way to measure vowel harmony. However, it's tricky enough defining what counts as a word boundary, so it may be impractical to try to modify this work to take into morpheme boundaries (especially for arbitrary languages, whose morphology may not be well-studied enough to be incorporated into the harmony measure).

More work also needs to be done to determine the relative effects of type versus token frequency. The corpora used for this work are a mixture of both dictionary word lists and full texts with repeated words. Ideally, both type and token harmony would be computed, and then combined in some way to give an integrated harmony measurement.

A related problem is that certain kinds of texts may be biased. Many of the corpora used here are Bible texts, which means a lot of repetitions of foreign names such as *Jesus*, biasing the results strongly in favor of particular frequent foreign words in the text that may not be representative of the language as a whole. This can also be a problem in the other direction, with word lists being biased by morphemes that may be far more productive in the lexicon than in speech (cf. English *re-* and *un-*). Both kinds of biases may be able to be circumvented by some kind of appropriate scaling or cut-off for excessively large frequencies.

However, as a baseline measure of vowel harmony, the current algorithm seems to work well, and yields at least two interesting harmony patterns (hidden harmony and anti-harmony) that warrant further exploration.

References

Harrison, K. David, Emily Thomforde, and Michael O'Keefe. 2002–2004. The vowel harmony calculator. `http://www.swarthmore.edu/SocSci/harmony/public_html/index.html`.