# Vowel dispersion in the lexicon: A corpus-based study on 25 languages

**Benjamin Storme, Kowsar Amiri & Alba Hermida Rodríguez (Leiden University)**

**1. Introduction.** Following Liljencrants and Lindblom's (1972) seminal paper on the Dispersion Theory, a number of works have shown that vowel inventories across the world's languages tend to favor acoustically dispersed vowels (e.g., Schwartz 1997, Flemming 2002, Becker-Kristal 2010, Cotterel & Eisner 2018). However there has been little research on whether acoustic dispersion also shapes the frequency of sounds *within* languages. Preliminary evidence suggests that it does: peripheral vowels (such as [a, i, u]) are on average more frequent than central vowels (such as [ʌ, y, ɯ]) in the lexicons of languages (Gordon 2016: Chapter 3.6).

This paper further investigates this question by fitting dispersion-based models to the frequency distribution of vowels in the lexicons of 25 languages. The results show that models with a bias towards acoustic dispersion generally provide a better fit to the data than models that do not include this bias, in line with the hypothesis that the pressure for acoustic distinctiveness shapes not only the sound inventories of languages but also their lexicons (e.g., Martin 2007).

**2. Methods.** 25 languages were sampled from the DoReCo speech corpus (Seifart, Paschen & Stave 2022), based on word token count (priority was given to languages with larger corpora) and on the availability of the data online.

Speech corpora from these 25 languages were used to get estimates of both (i) the frequency of use of oral vowels in the lexicons of these languages and (ii) the acoustic realization of these vowels, corresponding to the first four formants measured at vowel midpoint, using the FastTrack plugin (Barreda 2021) in Praat (Boersma & Weenink 2024). Following Liljencrants & Lindblom (1972) and subsequent work on the Dispersion Theory, only short oral vowels were included in the analysis.

The lexical frequency of vowels was then modeled in a MaxEnt grammar (Goldwater & Johnson 2003, Hayes & Wilson 2008) including a constraint favoring more distinct vowel contrasts. The specific implementation of this constraint closely follows Schwartz et al.'s (1997) proposal, with vowel contrasts being penalized according to the square of their (weighted) Euclidean distance in the Bark-transformed F1 $\times$ F2' space, where F2' is a linear function of F2, F3 and F4. Following Schwartz et al. (1997), we allowed languages to vary in the way F1 and F2' are weighted when calculating the Euclidean distance between vowels. Finally, we used the marginalization method proposed by Storme (2023) to infer the weight of the dispersion constraint from the lexical frequency of individual vowels in MaxEnt.

The dispersion-based model was then compared to two alternative models: (i) an unbiased model without any pressure towards dispersion and (ii) a model with a bias favoring vowels with formants in close proximity, as in Quantal Theory (Stevens 1972, 1989). The unbiased model just predicts that vowels should be equally probable in the lexicon. The constraint favoring quantal vowels was operationalized as in Schwartz et al (1997), allowing the distance between F1 and F2 to be weighted higher than the distance between the other formants (F2-F3, F3-F4) in the overall formant-proximity measure.

All models were implemented as Bayesian multinomial regressions, using the JAGS package in R. Models were compared using the Deviation Information Criterion (DIC).

**3. Results.** Dispersion-based models were generally found to be characterized by a smaller DIC than the other two models, indicating a better fit to the frequency distribution of vowels in the lexicon. This was the case in 24 out of 25 languages. Figure 1 shows the predicted vs attested lexical frequency of vowels from the corpus under the dispersion-based model. The model provides a reasonable fit to the data ($R^2 = 0.63$). As expected, peripheral vowels, and in particular [a], tend to be more frequent than non-peripheral vowels in the lexicons of languages. Figure 2 shows the same data, but this time aggregated across languages.

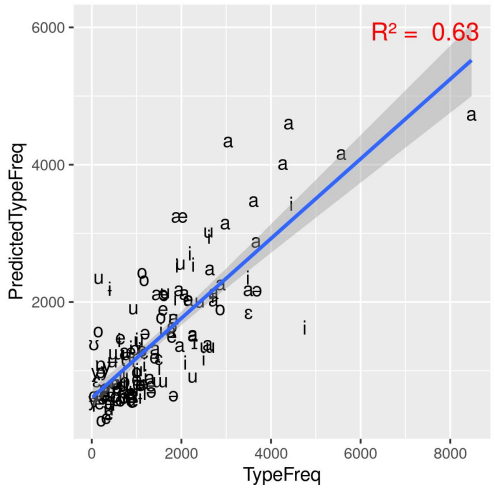


**Fig 1:** Attested vs predicted vowel count in the lexicons of languages (each point represents a vowel in one of the 25 languages in the corpus)

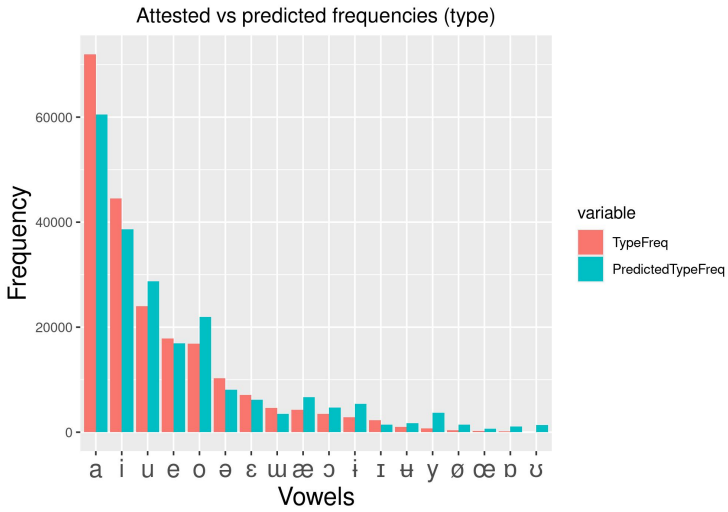


**Fig 2:** Attested vs predicted vowel count in lexicons (aggregated across languages)

**4. Discussion.** For one of the languages, Evenki (Tugunsic, Russia), the dispersion-based model was found to provide a worse fit than the two other, non-dispersion-based models. This problem results from the presence of a highly frequent schwa vowel in the language. Schwa is known to be generally problematic for the dispersion theory and its prevalence across languages has been explained as due to effort minimization, with schwa being short enough to remain sufficiently distinct from other oral vowels (Schwartz et al 1997). In line with this hypothesis, we found that, among the short vowels [a, i, u, o, ə] occurring in the Evenki corpus, schwa has the shortest duration. In future research, one might include other variables in dispersion-based models, besides the first four formants: duration, as suggested above, or phonotactics, in particular for languages where schwa resuls from vowel reduction.

The dispersion theory was also found to make interesting and unexpected predictions. In particular, the three cardinal vowels [a, i, u] were found to differ in their lexical frequency, with [a] being generally more frequent than [i] and [i] more frequent than [u]. Dispersion-based models can actually predict this ordering, provided that F1 is weighted more than F2' when calculating the Euclidean distance between vowels: in that case, [a] is the preferred vowel quality because it is distant from both [i] and [u] along F1. And [u] is the least preferred vowel quality because it stands between [i] and [a] along F1, [u] being generally slightly lower than [i] (de Boer 2011).

All in all, this paper provides quantitative evidence that acoustic dispersion plays a role beyond phonology and also shapes the way lexicons are built across languages.

**Selected references.** Gordon 2016. *Phonological Typology*. Oxford Academic. • Liljencrants & Lindblom 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48, 839–62. • Martin 2007. The evolving lexicon. UCLA PhD dissertation. • Seifart, Paschen, Stave (eds.). 2022. *Language Documentation Reference Corpus (DoReCo)* 1.2. Berlin & Lyon. • Schwartz et al 1997. The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* 25, 255-286.