Investigating the universality of consonant and vowel co-occurrence restrictions

Amanda Doucette
McGill University
amanda.doucette@mail.mcgill.ca

Timothy J. O'Donnell
McGill University,
Canada CIFAR AI Chair, Mila
timothy.odonnell@mcgill.ca

Morgan Sonderegger
McGill University
morgan.sonderegger@mcgill.ca

Heather Goad
McGill University
heather.goad@mcgill.ca

Abstract Certain phonotactic constraints on the co-occurrence of segments appear to be much more common across the world's languages than others. In many languages, similar consonant co-occurrence is restricted through Obligatory Contour Principle (OCP) effects, while there are some exceptions for identical consonants. In vowels, the opposite pattern appears to hold: many languages have vowel harmony processes, where vowels within a domain are required to share some feature. Languages that encourage similar consonant co-occurrence or restrict similar vowel co-occurrence appear to be exceedingly uncommon. However, evidence of this pattern so far only comes from studies of individual languages or families, or of only consonants or vowels. We investigate patterns of co-occurrence in vowels and consonants in 107 Northern Eurasian languages across 21 families using Bayesian negative binomial regression to explicitly model the effects of aggregate similarity and segment identity on co-occurrence counts (the results of which can be interpreted similarly to observed/expected ratios). We find that the effect of similarity is remarkably consistent across languages: similar consonant co-occurrence is disfavored, while aggregate similarity has no effect on vowel co-occurrence. Identical segment co-occurrence effects are much more variable across languages, with a tendency towards disfavoring identical consonants, and favoring identical vowels. We also find larger effects in consonants than in vowels, suggesting that consonant co-occurrence is more strongly constrained than vowel co-occurrence. We also find that there is no evidence for or against any correlations between vowel and consonant co-occurrence, suggesting that more data is needed to evaluate this possibility.

Keywords: phonology; phonotactics; harmony; Bayesian modeling; typology; co-occurrence; similarity

1 Introduction

It is well known that the possible words in a language are governed by strong *phonotactic constraints* privileging the co-occurence of certain combinations of segments, while discouraging the co-occurrence of others. For example, in Tuvan, all vowels in a word are alike in backness ([idegel] 'hope' vs. *[adegel], Rose & Walker 2011). This is called *harmony*, a phenomenon where segments sharing some feature are more likely to co-occur. In other cases, the co-occurrence of segments sharing a feature is restricted: In Sanskrit roots, only one aspirated consonant is permitted ([b^{h}id] 'to split' vs. *[b^{h}id^{h}], Ito & Mester 1998). While a great deal of work has been devoted to characterizing *formal*

aspects of phonotactics – that is, the nature of representations needed to describe (Chomsky & Halle 1968; Clements & Hume 1995) or learn (Hayes & Wilson 2008; Heinz 2010; Saffran 2003) co-occurrence constraints – it is less well understood why some particular classes of sounds appear to be *a priori* more or less likely to co-occur. Some work has attributed the distribution of phonotactic constraints to properties of memory (Gafos 2021; Endress et al. 2009), language processing (Frisch 2004), acoustic or articulatory properties of sounds (Ohala 1994), and signaling domain boundaries (Trubetzkoy 1939; Kaye 1989). However, *which* phonotactic constraints are actually more likely is not well understood – no systematic, large scale study has quantified the distribution of phonotactic patterns over a large set of languages.

A puzzling example arises in a comparison of vowels and consonants. The literature on harmony effects reveals a striking pattern: While there are hundreds of examples of vowel harmony, consonant harmony appears to be much less common (e.g., Hansson 2010). In consonants, the opposite pattern appears to be more frequent – many languages have a *restriction* against co-occurring similar consonants (McCarthy 1986; Frisch et al. 2004; Gordon 2016).

While a tendency toward similar vowel co-occurrence and similar consonant avoidance appears to hold true across languages, this does not seem to apply to *identical* segments. In Arabic, for example, there is a restriction against consonants sharing place of articulation co-occurring in a root, but no restriction if they are identical (McCarthy 1986; Frisch et al. 1997; Coetzee & Pater 2008).

The generalizations of published results would seem to indicate that vowels and consonants interact within words in opposite ways, and that identical segments are treated differently from merely similar segments. However, evidence of this remains mostly confined to studies of individual languages or families. To understand the extent of this difference between consonants and vowels, we need to examine more subtle patterns both within and across languages. While many of the phonotactic constraints identified in the literature apply categorically with few exceptions, it has long been known that gradient phonotactic patterns exist as well (Greenberg & Jenkins 1964; Ohala & Ohala 1986; Luce & Pisoni 1998; Frisch et al. 2000; 2001; Hammond 2004). There is a significant body of work describing and modeling gradient phonotactic patterns (Hammond 2004; Hayes & Wilson 2008; Coetzee & Pater 2008; Anttila 2008; Futrell et al. 2017), including several which examine the effect of similarity on consonant co-occurrence (Frisch et al. 2004; Pozdniakov & Segerer 2007; Mayer et al. 2010; Graff 2012) and vowel co-occurrence (Walter 2010; Archangeli et al. 2012). However, no cross-linguistic study has investigated vowel and consonant co-occurrence effects simultaneously, or explored differences between the effects of similarity and identity.

In this paper, we investigate vowel and consonant co-occurrence effects in 107 languages across 21 families, using hierarchical Bayesian negative binomial regression. Vowel and consonant co-occurrences are jointly modeled, allowing us to investigate *relationships* between vowel and consonant effects – a factor which has not been considered in previous work. This allows us to test for both differences in effect sizes and potential correlations between vowel and consonant effects. To explore differences between identical and similar (but not identical) segments, we model identity and similarity as independent effects. Bayes factors (Jeffreys 1961; Kass & Raftery 1995) are used to distinguish between existence and non-existence of effects, and to quantify uncertainty about the existence of an effect. Finally, we include random effects of language and language family to examine how these effects vary across languages. We address the following questions: Do all languages have similar co-occurrence restrictions? Do restrictions vary by language or by

language family? Is there any relationship between restrictions on consonant and vowel co-occurrence?

We find that for consonants, similar pairs tend to be avoided within words, as expected. We find that identical pairs of consonants are avoided as well. These effects do not qualitatively differ across languages, although there is more variance in consonant identity effects than similarity effects. For vowels, we do not find so clear a result – we find a consistent tendency for identical vowels to co-occur, but no strong tendency for similar (non-identical) vowels to co-occur. Across languages, vowel effects differ qualitatively: In some languages, similar vowels are more likely to co-occur, in others the opposite is true, and in some there is a null effect. We also find that consonant co-occurrence effects are consistently larger than vowel co-occurrence effects. Finally, we find that there is inconclusive evidence for or against any correlation between vowel and consonant co-occurrence effects, suggesting there are not enough languages in this sample to assess potential correlations between consonant and vowel co-occurrence effects.

2 Background

2.1 Similar vowel co-occurrence effects

One of the most well-documented vowel co-occurrence effects is vowel harmony, a phonological process in which similar vowels in a particular domain (generally a word) are more likely to co-occur than dissimilar vowels. Vowel harmony processes have been documented in many languages, across many language families (Archangeli & Pulleyblank 2007; Rose & Walker 2011; Gordon 2016; Ritter & van der Hulst 2024).

While vowel harmony is typically described as applying categorically in the relevant contexts (with few exceptions), cases of *gradient* harmony, where the phonological process applies probabilistically, have been identified using statistical methods (Archangeli et al. 2012: for Bantu languages). By examining co-occurrence probability as a function of vowel similarity, gradient vowel harmony patterns can be identified in languages without documented vowel harmony processes. Walter (2010) identifies a gradient vowel harmony pattern in Croatian, and speculates that the tendency toward similar vowel co-occurrence may be universal – that similar vowels are more likely to co-occur than dissimilar vowels across *all* languages. Our paper explicitly tests this possibility by examining the effect of similarity on vowel co-occurrences across a large sample of languages.

While harmony appears to be the most widely studied vowel co-occurrence effect, cases of *anti-harmony* – where dissimilar vowels are more likely to co-occur – have also been documented, often as exceptions to otherwise regular harmony processes. For example, Russian loanwords in Tuvan appear to be exceptions to backness harmony ([ɨraketa] 'rocket', Harrison 1999). In other words, although similar vowel co-occurrences appear to be much more common across languages, similar vowel avoidance is also possible.

2.2 Similar consonant co-occurrence effects

In contrast to vowel co-occurrence effects, similar consonant co-occurrence appears to be avoided. Similar consonant avoidance restrictions have been documented in many languages. The Obligatory Contour Principle (OCP) is a widely discussed co-occurrence restriction that prohibits adjacent identical elements from co-occurring (Leben 1973; McCarthy 1986). While the OCP was originally formulated as a categorical restriction, gradient consonant co-occurrence restrictions have also been identified in several lan-

guages. Frisch et al. (2004) found that the more natural classes that are shared between a pair of consonants, the less likely that pair is to occur in an Arabic root. Pozdniakov & Segerer (2007), Coetzee & Pater (2008), Mayer et al. (2010), and Graff (2012) identified gradient co-occurrence restrictions in many languages (mainly on place of articulation features), suggesting that this may be a universal restriction against similar consonant co-occurrence. The models described in this paper will test for this, generalizing from similarity based only on place of articulation features to similarity across all features.

However, there are also languages where similar consonants are likely to co-occur. Consonant harmony has been identified in some languages (Hansson 2010; Rose & Walker 2011), but it appears to be restricted to sets of segments defined by particular features (coronals, sibilants, etc.) or by intersections of features, in most cases.

2.3 Identical segment co-occurrence effects

Further complicating this picture of co-occurrence effects, there is evidence that in some languages, completely identical pairs of segments can behave differently from merely similar segments.[1] We define *identity* as total identity: agreement for all features. In contrast, we define *similarity* in terms of the shared features between a pair of non-identical segments. For example, [t] and [t] are considered identical because they share all features, while [t] and [d] are considered similar because they differ in the [voice] feature.

MacEachern (1999) found that in several languages with strong co-occurrence restrictions on similar consonants, identical consonants could still co-occur. In Peruvian Aymara, words with two ejective consonants are prohibited in general, but are allowed if they are identical.[2] Pozdniakov & Segerer (2007) and Coetzee & Pater (2008) found that identical consonant pairs occurred more frequently than expected in many languages. In Arabic roots, consonants with the same place of articulation cannot co-occur in general, but identical consonants can co-occur in specific circumstances (Greenberg 1950; McCarthy 1986).

There is less evidence for the effect of identity on vowel co-occurrence. In Croatian and Spanish, Walter (2010) found that identical vowels pairs are underattested. In Croatian, similar but not identical vowel pairs are overattested, but in Spanish they are underattested. There does not appear to be evidence of vowel identity effects differing from similarity effects in other languages.

Finally, there is evidence that humans have a specialized cognitive mechanism for processing identical elements (Endress et al. 2009). This evidence, along with evidence of identical segments behaving unusually in several languages, suggests that identity does not function simply as a stronger form of similarity, and should be considered separately. For these reasons, identity and similarity were included as independent effects *a priori* in our models.

2.4 Relationships between vowel and consonant effects

Co-occurrence effects on vowels and consonants have been investigated independently of each other, so these analyses fail to include the possibility that they are related in some way. There are several ways in which vowel and consonant co-occurrence effects could be related. First, consonant or vowel effects could be systematically larger than

[1] We exclude geminate consonants (e.g. [tt]), as we are only examining co-occurrence at a distance.
[2] [k'ink'u] 'clay' is a well-formed word in Peruvian Aymara, while *[t'ink'u] is not.

the other. Differences in relative effect sizes could explain discrepancies between how often vowel and consonant co-occurrence patterns are identified in languages.

For example, it could be possible that similar vowel co-occurrence restrictions are near-universal, while similar consonant co-occurrence effects are fairly common but not universal, or that identical consonants co-occur freely, while identical vowel co-occurrence is restricted. Another possibility is that both vowel and consonant co-occurrence effects are near-universal, but one is significantly smaller and harder to detect. We investigate these possibilities by comparing vowel and consonant effect sizes in our models.

Second, the size of consonant and vowel co-occurrence effects could be correlated across languages. The structure of the Bayesian regression models we use will allow us to estimate correlations between vowel and consonant effects across languages. We have no strong expectations about the existence or direction of a correlation. There could be no correlation, meaning that vowel and consonant effects are completely independent. Another possibility is that languages will have strong co-occurrence effects in both vowels and consonants, perhaps because learning "language x has harmony" is easier than "language x has harmony in vowels, but not in consonants." Finally, languages could have strong co-occurrence effects in *either* consonants or vowels (but not both) due to information theoretic restrictions on the lexicon. If vowel co-occurrences are strongly restricted in a language, consonants would need to be relatively unrestricted to allow for sufficient coding space in the lexicon. As far as we know, no work has been done investigating these possibilities.

3 Methods

3.1 Data

We use data from NorthEuraLex (Dellert et al. 2020), a set of comparable lexicons for 107 Northern Eurasian languages, across 21 families. Language families included in NorthEuraLex are listed in Table 1, along with the languages and subfamilies included in them. For each language, there are IPA transcriptions of 1,016 basic concepts (where a "concept" is a whole word, including inflectional morphology). Because it is typically assumed that co-occurrence restrictions are influenced more by word *types* rather than *tokens*, we do not include any further information about concept frequency (Frisch et al. 2004; Wilson & Obdeyn 2009). Transcriptions for some languages were automatically generated, so some manual data cleaning was necessary to make transcriptions consistent across all languages. For example, different Unicode symbols are used to represent the same IPA symbols in some cases, and some transcriptions have erroneous IPA diacritics (e.g. occurring independently of a base segment).[3]

We chose NorthEuraLex over other parallel corpora because it contains a much larger set of languages than others. Although there are a relatively small number of words per language, we are interested in cross-linguistic generalizations about co-occurrences, rather than properties of individual languages. Although inflectional morphology is included in the data for some languages, the majority of concepts are lemmas without inflection (Pimentel et al. 2020). Thus, our results are unlikely to be affected by repeated instances of inflectional morphemes.

While NorthEuraLex represents 21 language families, two are particularly overrepresented – Indo-European with 37 languages, and Uralic with 26 languages. To account for this disparity in language family representation, we instead classify these languages

[3] Scripts used for cleaning data can be found in the OSF project for this paper, at https://osf.io/sgu4w/.

Family	Languages (Subfamilies)
Uralic	Estonian, Finnish, Livonian, North Karelian, Olonets Karelian, Veps (Finnic); Inari Saami, Kildin Saami, Lule Saami, Northern Saami, Skolt Saami, Southern Saami (Saami); Forest Enets, Nganasan, Northern Selkup, Tundra Nenets (Samoyedic); Komi-Permyak, Komi-Zyrian, Udmurt (Permian); Hill Mari, Meadow Mari (Mari); Erzya, Moksha (Mordvin); Hungarian (Hungarian); Northern Khanty (Khantyic); Northern Mansi (Mansic)
Indo-European	Belarusian, Bulgarian, Croatian, Czech, Latvian, Lithuanian, Polish, Russian, Slovak, Slovenian, Ukrainian (Balto-Slavic); Danish, Dutch, English, German, Icelandic, Norwegian Bokmål, Swedish (Germanic); Catalan, French, Italian, Latin, Portuguese, Romanian, Spanish (Italic); Bengali, Hindi, Northern Kurdish, Northern Pashto, Ossetian, Western Farsi (Indo-Iranian); Breton, Irish, Welsh (Celtic); Standard Albanian (Albanian); Armenian (Armenic); Modern Greek (Graeco-Phrygian)
Turkic	Bashkir, Kazakh, Tatar (Kipchak); North Azerbaijani, Turkish (West Oghuz); Chuvash (Bolgar); Sakha (North Siberian Turkic); Southern Uzbek (Turkestan Turkic)
Nakh-Daghestanian	Avar, Tsez (Avar-Andic-Tsezic); Dargwa (Dargwic); Lak (Lak); Lezgian (Lezgic); Chechen (Nakh)
Dravidian	Kannada, Malayalam, Tamil, Telugu (South Dravidian)
Eskimo-Aleut	Aleut (Aleut); Central Siberian Yupik, Kalaallisut (Eskimo)
Mongolic	Kalmyk, Khalkha Mongolian, Russia Buriat (Eastern Mongolic)
Tungusic	Nanai (Central Tungusic); Manchu (Manchu-Jurchen); Evenki (Northern Tungusic)
Abkhaz-Adyge	Abkhaz (Abkhaz-Abaza); Adyghe (Circassian)
Afro-Asiatic	Standard Arabic, Modern Hebrew (Semitic)
Chukotko-Kamchatkan	Chukchi (Chukotian); Itelmen (Itelmen)
Yukaghir	Southern Yukaghir (Kolymic); Northern Yukaghir (Northern Yukaghir)
Ainu	Hokkaido Ainu (Hokkaido-Kuril Ainu)
Basque	Basque (Basque)
Burushaski	Burushaski (Burushaski)
Japonic	Japanese (Japanesic)
Kartvelian	Georgian (Georgian-Zan)
Koreanic	Korean (Korean)
Nivkh	Nivkh (Nivkh)
Sino-Tibetan	Mandarin Chinese (Sinitic)
Yeniseian	Ket (Northern Yeniseian)

Table 1: Languages, language families, and subfamilies, as categorized by NorthEuraLex (Dellert et al. 2020).

by their *subfamilies*. This change allows for a more even distribution of language family sizes.[4]

Counting co-occurrences requires defining what counts as a co-occurrence in the lexicon. This has been done in various ways in previous work, e.g. just using parts of words (Frisch et al. 2004) or words of a certain shape (Graff 2012) to deal with statistical non-independence. We instead choose to maximize the amount of data per language, by making the following choices. Co-occurrences were counted as pairs of vowels separated by one consonant or consonant cluster, and pairs of consonants separated by one vowel. Because many languages in NorthEuraLex do not allow vowel or consonant clusters, only non-adjacent pairs were counted to ensure results are comparable across languages. For example, the word [kɑnsənənt] would result in the consonant pairs [kn], [sn], and [nn], and the vowel pairs [ɑə], and [əə].

3.2 Similarity metrics

To ensure that results are not dependent on a particular similarity metric, we fit models using two different feature-based similarity metrics – *Feature Similarity*, and *Natural Class Similarity*. Both of these similarity metrics are calculated using the phonological features of a pair of segments. Phonological features describe the articulatory and acoustic dimensions along which speech sounds vary, abstracting away from raw acoustic measurements. For all languages, we used the same set of binary features from PanPhon (Mortensen et al. 2016), which are listed in Appendix A.

Feature Similarity (Equation 1), first used by Pierrehumbert (1993), is the simplest possible measure: a normalized intersection size between the phonological features of two segments. Although the PanPhon feature set contains 24 features, not all of these are relevant in every language. For example, if a language has a consonant inventory of [p t k], the feature [strident] is not relevant because there are no [s] or [ʃ]-like consonants. The features [coronal], [labial], and [back] would be included in the feature set for this language, because they are contrastive for the set of three obstruents. In Equation 1, $|Feats|$ refers to the number of relevant features in a given language.

$$FeatSim(x, y) = \frac{|Feats(x) \cap Feats(y)|}{|Feats|} \tag{1}$$

Although Feature Similarity has a range of [0, 1] (from sharing no features to being completely identical), in practice there is a minimum similarity value in real languages – no pair of consonants or vowels in an inventory share zero features. All pairs of vowels will share a large number of features, as the features that are only relevant to contrasting consonants will be shared among all vowels. This results in a skewed distribution – under the Feature Similarity metric, vowels are inherently more similar than consonants. This can be seen in Figure 1.

The *Natural Class Similarity* metric, originally defined by Frisch (1996); Frisch et al. (1997; 2004) and widely used in other studies of gradient phonotactics (MacEachern 1999; Wilson & Obdeyn 2009; Walter 2010), defines *natural classes* as sets of segments that share one or more features. In this similarity metric, features that are only partially contrastive in an inventory contribute less to the similarity score than fully contrastive features (in Feature Similarity, all features contribute equally).[5] Natural Class Similarity

[4] We thank an anonymous reviewer for suggesting this.

[5] For example, in a language with [s], the feature [strident] is only contrastive within the set of consonants, but not in vowels.

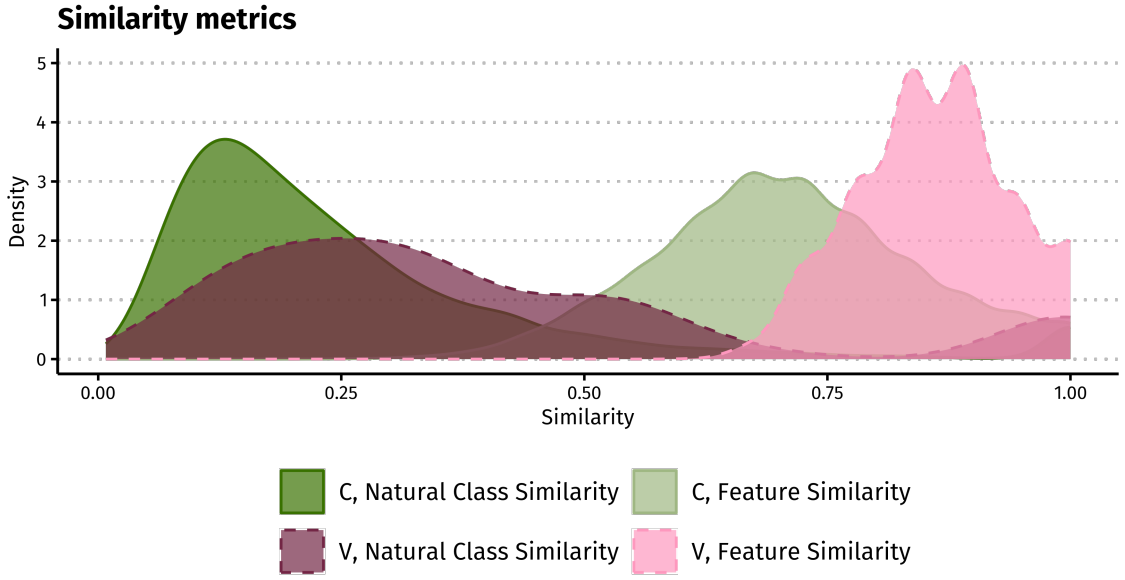

Figure 1: Density plot of similarity metrics across all consonant and vowel pairs.

is calculated according to Equation 2, where $NC(x)$ gives the set of natural classes for segment x.

$$NCSim(x,y) = \frac{|NC(x) \cap NC(y)|}{|NC(x) \cup NC(y)|} \tag{2}$$

Natural Class Similarity also has a range of [0, 1], but in this case it is possible for a language to have a pair of segments with similarity zero.[6] Unlike in Feature Similarity, vowels are not inherently more similar under the Natural Class Similarity metric, as shown in Figure 1. Features that are shared between all vowels are "ignored" in the Natural Class Similarity calculation, because they do not allow for the formation of a distinct natural class. Although these two metrics are clearly correlated, we fit models for both of them in part because of their different treatment of vowel similarity.

3.3 Models

Previous work has used observed/expected (O/E) ratios or logistic regression to model co-occurrences (Frisch et al. 2004; Graff 2012). An O/E ratio for a particular pair of segments x and y is calculated according to Equation 3, where N is the total number of observed pairs, $x+$ is the number of times x occurs as the first segment in a pair, and $+y$ is the number of times y occurs as the second segment in a pair.

$$\frac{O_{xy}}{E_{xy}} = \frac{O_{xy}}{N \cdot \frac{O_{x+}}{N} \cdot \frac{O_{+y}}{N}} \tag{3}$$

While this ratio has the benefit of being easily interpretable (values greater than 1.0 correspond to over-attestation, while values under 1.0 correspond to under-attestation), it has several disadvantages. First, it assumes that the O/E ratio is directly correlated with

[6] For example, consider the similarity of [p] and [m] in a language with a consonant inventory of [p b m]. $NCSim(p,m) = \frac{|\{\{p,b\},\{p\}\} \cap \{\{b,m\},\{m\}\}|}{|\{\{p,b\},\{p\}\} \cup \{\{b,m\},\{m\}\}|} = \frac{|\{\}|}{|\{\{p,b\},\{b,m\},\{p\},\{m\}\}|} = \frac{0}{4} = 0$

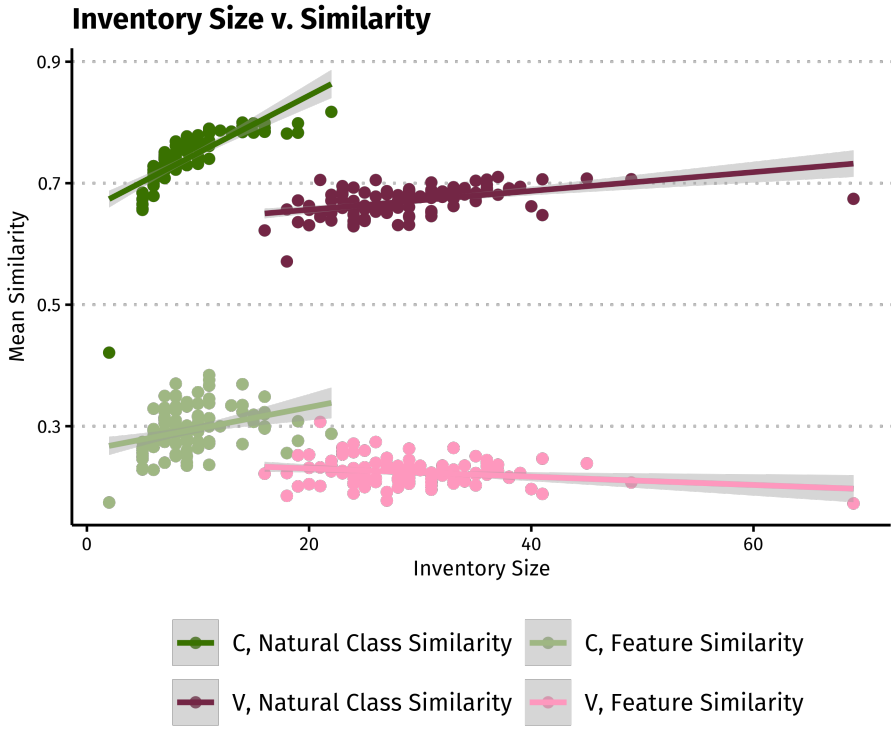

Figure 2: Consonant and vowel inventory size plotted against mean similarity per language, with linear smooths (lines) and 95% confidence intervals (shading).

the variables of interest (in our case, similarity and identity). As pointed out by Wilson & Obdeyn (2009), there is no theoretical justification for why we should choose O/E over a different value, like O − E. It is also not possible to account for control predictors in an O/E ratio: It simply accounts for the positional frequencies of individual segments. For example, we may want to account for varying inventory sizes across languages. Because similarity values depend on the number of distinctive features or natural classes in a language, they are expected to be somewhat dependent on inventory size: A larger inventory requires more distinctive features. This is shown in Figure 2. Thus, in order to compare results across languages, we need a method that allows for control predictors.

To compensate for the shortcomings of O/E ratios, we use multivariate Bayesian regression to model the effects of similarity and identity on consonant and vowel co-occurrence. This type of model offers several advantages over previous models. This method follows Graff (2012) and Wilson & Obdeyn (2009) in using regression modeling, which allows for control predictors. We are able to directly model co-occurrence counts, rather than the probability of a co-occurrence being attested in a language. We are also able to estimate effects across a large set of languages, and include random effects for language and language family (as a basic control for typological relatedness). By using multivariate regression, we can simultaneously model consonant and vowel co-occurrence, and estimate correlations between random effects. Through Bayesian regression, we obtain posterior distributions for each parameter in the model, rather than point estimates and confidence intervals. From these posteriors, we can conduct Bayesian hypothesis tests, allowing us to reject *or* accept null hypotheses, as described in §4.

Although multivariate Bayesian regression allows for greater flexibility in modeling co-occurrence, the results obtained are not as directly interpretable as O/E ratios. Our model

estimates the number of times a particular segment pair occurs – an estimate of the O_{xy} in Equation 3. $\frac{O_{x+}}{N}$ and $\frac{O_{+y}}{N}$ are the frequencies of x and y in a sample of N pairs, which can be obtained by counting segments in the NorthEuraLex data. These frequencies are included as *offsets* in the models, allowing us to interpret the model predictions as the degree of co-occurrence beyond what is expected from individual segment frequencies. Therefore, we can transform any model prediction into a predicted O/E ratio, accounting for the model's control predictors and random effect estimates, according to Equation 3. This predicted O/E ratio can be calculated for every value of similarity and identity, for each language in our sample, and will be used to aid in interpretation of results in §4.

Observed pair counts for consonant ($O_{xy/C}$) and vowel ($O_{xy/V}$) co-occurrences are each modeled with a negative binomial distribution, as shown in Equations 4 and 5.[7] Negative binomial regression is a generalization of Poisson regression (Winter & Bürkner 2021) used to model count data, which relaxes the latter's assumption of equal mean and variance by including an additional shape parameter (ϕ in Equations 4 and 5), allowing the data to be *overdispersed*.

$$O_{xy/C} \sim \text{NegBinom}(\mu_C, \phi_C) \tag{4}$$

$$O_{xy/V} \sim \text{NegBinom}(\mu_V, \phi_V) \tag{5}$$

The expected values of these distributions are the mean parameters μ_C and μ_V:

$$\mathbb{E}[O_{xy/C}] = \mu_C \tag{6}$$

$$\mathbb{E}[O_{xy/V}] = \mu_V \tag{7}$$

The variances of the distributions are functions of the means μ_C and μ_V and the shape parameters ϕ_C and ϕ_V:

$$\text{Var}[O_{xy/C}] = \mu_C + \frac{\mu_C^2}{\phi_C} \tag{8}$$

$$\text{Var}[O_{xy/V}] = \mu_V + \frac{\mu_V^2}{\phi_V} \tag{9}$$

The means of the distributions are modeled as linear functions of predictors. Main effect predictors, represented by β parameters in equations, include similarity (sim), identity (ident), log consonant inventory size ($\ln(|\text{C}_{\text{inv}}|)$), log vowel inventory size ($\ln(|\text{V}_{\text{inv}}|)$), and an intercept.[8] For example, $\beta_{\text{intercept}|C}$ is the intercept for the consonant co-occurrence model. Random effects for similarity and identity are included by language family (γ parameters, i.e. $\gamma_{\text{sim}|C}$, the by-family random effect of similarity in the consonant co-occurrence model), and by language within family (α parameters, i.e. $\alpha_{\text{ident}|V}$, the by-language random effect of identity in the vowel co-occurrence model). As an individual language only belongs to one family, these random effects are nested (see Sonderegger 2023: §10.2.2 for more details). The similarity effect parameters are multiplied by $(1 - \text{ident}(x, y))$, which has the effect of setting them to zero when x and y are identical.

[7] The model described by these equations is also shown in *brms* formula format in Appendix B

[8] Models without these inventory size predictors were also fit, and gave results qualitatively identical to the models including inventory size. For simplicity, only the models including inventory size are reported.

Thus, when x and y are identical the similarity effect parameters do not contribute to μ, and when x and y are not identical, the identity effect parameters do not contribute to μ. The log frequencies of each segment in the pair are included as offsets in the models (1 is added to observed counts to avoid ln(0) in the case where a segment does not occur in a position):

$$\begin{aligned}\ln(\mu_C) =& \beta_{\text{intercept}|C} + \gamma_{\text{intercept}|C} + \alpha_{\text{intercept}|C} \\ &+ (1 - \text{ident}(x,y)) \cdot \text{sim}(x,y) \cdot (\beta_{\text{sim}|C} + \gamma_{\text{sim}|C} + \alpha_{\text{sim}|C}) \\ &+ \text{ident}(x,y) \cdot (\beta_{\text{ident}|C} + \gamma_{\text{ident}|C} + \alpha_{\text{ident}|C}) \\ &+ \beta_{\text{C.inv}|C} \ln(|\text{C}_{\text{inv}}|) + \beta_{\text{V.inv}|C} \ln(|\text{V}_{\text{inv}}|) \\ &+ \underbrace{\ln\left(\frac{O_{x+}+1}{N} \cdot \frac{O_{+y}+1}{N}\right)}_{\text{segment frequency offset}}\end{aligned} \tag{10}$$

$$\begin{aligned}\ln(\mu_V) =& \beta_{\text{intercept}|V} + \gamma_{\text{intercept}|V} + \alpha_{\text{intercept}|V} \\ &+ (1 - \text{ident}(x,y)) \cdot \text{sim}(x,y) \cdot (\beta_{\text{sim}|V} + \gamma_{\text{sim}|V} + \alpha_{\text{sim}|V}) \\ &+ \text{ident}(x,y) \cdot (\beta_{\text{ident}|V} + \gamma_{\text{ident}|V} + \alpha_{\text{ident}|V}) \\ &+ \beta_{\text{C.inv}|V} \ln(|\text{C}_{\text{inv}}|) + \beta_{\text{V.inv}|V} \ln(|\text{V}_{\text{inv}}|) \\ &+ \underbrace{\ln\left(\frac{O_{x+}+1}{N} \cdot \frac{O_{+y}+1}{N}\right)}_{\text{segment frequency offset}}\end{aligned} \tag{11}$$

Exploratory data analysis showed that languages differ in the variance of pair counts. Therefore, the shape parameters ϕ_C and ϕ_V are allowed to vary by language family and by language within family:

$$\ln(\phi_C) = \gamma_{\phi|C} + \alpha_{\phi|C} \tag{12}$$

$$\ln(\phi_V) = \gamma_{\phi|V} + \alpha_{\phi|V} \tag{13}$$

In Bayesian regression, prior distributions need to be specified for each parameter in the model. Although priors can be informed by domain knowledge, any prior will be overwhelmed by a large quantity of data. We use *weakly informative* priors, meaning that they impose weak constraints on the likely values of parameters (Nicenboim & Vasishth 2016; Vasishth et al. 2018). For example, for all main effect parameters, the following prior is used:

$$\beta \sim \mathcal{N}(0, 4) \tag{14}$$

This distribution is plotted in Figure 3. Although all values are possible, it puts slightly more weight on values near zero with 95% falling between −7.84 and 7.84. A change of 7.84 in log space corresponds to a change of approximately 2540 in raw pair counts. All pair counts in NorthEuraLex are under 1000, making this a very conservative prior.

The separate models for consonant and vowel co-occurrence are tied together through their random effect priors. Random effects by language have a multivariate normal prior:

$$\begin{bmatrix} \alpha_{\text{intercept}|C} \\ \alpha_{\text{intercept}|V} \\ \alpha_{\text{sim}|C} \\ \alpha_{\text{sim}|V} \\ \alpha_{\text{ident}|C} \\ \alpha_{\text{ident}|V} \end{bmatrix} \sim \text{MVNorm}(0, \Sigma_\alpha) \tag{15}$$

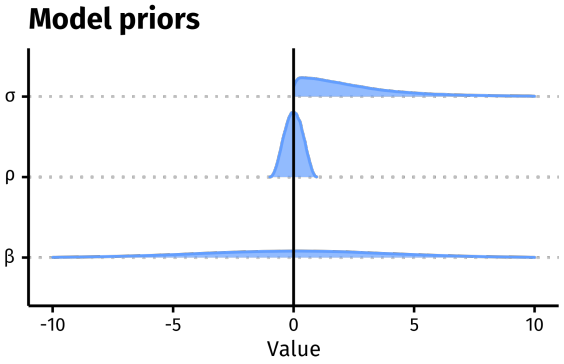

Figure 3: Prior distribution densities for standard deviation (σ), correlation (ρ), and fixed effect (β) model parameters.

The covariance matrix Σ_α is shared across the consonant and vowel models, and is parameterized in a standard way (Equation 16: see McElreath 2020 for more details). After fitting the model, correlations between consonant and vowel random effects can be extracted from this matrix. Σ_α is defined as:

$$\Sigma_\alpha = S_\alpha R S_\alpha \tag{16}$$

Where S_α is a diagonal matrix:

$$S_\alpha = \begin{bmatrix} \sigma(\alpha_{\text{intercept}|C}) & 0 & \cdots & 0 \\ 0 & \sigma(\alpha_{\text{intercept}|V}) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \sigma(\alpha_{\text{ident}|C}) & 0 \\ 0 & \cdots & 0 & \sigma(\alpha_{\text{ident}|V}) \end{bmatrix} \tag{17}$$

And R is a correlation matrix with an LKJ hyperprior (Lewandowski et al. 2009):

$$R \sim \text{LKJcorr}(1.5) \tag{18}$$

The LKJ prior penalizes strong (non-zero) correlations, and is plotted in Figure 3. We have no *a priori* evidence of any relationship between consonant and vowel effects, so we use a weakly informative prior that penalizes strong correlations. Language family random effects (γ parameters) are parameterized in exactly the same way as by-language random effects.

The prior for standard deviation parameters, plotted in Figure 3, is:

$$\sigma \sim \text{HalfStudentT}(3, 0, 2.5) \tag{19}$$

And finally, we specify a prior for the shape parameters ϕ_C and ϕ_V as the default weakly informative prior used by *brms* (Bürkner 2017):

$$\phi_C, \phi_V \sim \Gamma(0.01, 0.01) \tag{20}$$

Two models, one using Natural Class Similarity (referred to as the **NC** model) to define $\text{sim}(x, y)$ in Equations 10 and 11, and one using Feature Similarity (the **Feat** model) to define $\text{sim}(x, y)$ in Equations 10 and 11, were implemented in *brms* (Bürkner 2017), an R (R Core Team 2020) interface to the Stan probabilistic programming language (Stan Development Team 2019; Gabry et al. 2023). Models were fit using four Markov chains with 5000 warmup samples and 20000 post-warmup samples each. $\hat{R}$ and ESS values suggested model convergence.

4 Results

For each of the two models (**NC** and **Feat**), we report fixed effect regression coefficients in Tables 2 and 3. Standard deviations of random effect parameters are reported in Tables 4 and 5, and correlation parameters are reported in Tables 6 and 7. Because similarity and identity values for vowels and consonants in both the **NC** and **Feat** models are normalized to the same $[0, 1]$ scale, the magnitudes of these coefficients are directly comparable.

For each model parameter, we report an estimate, a 95% credible interval (CredI), a probability of direction (p_d), and a Savage-Dickey density ratio (Bayes Factor, henceforth BF).[9] The parameter estimate is the value of the parameter with the highest posterior probability. Posterior distributions for all parameters, including their priors, are plotted in Figures 5 and 6. 95% credible intervals are the interval centered around the parameter estimate that contains 95% of the posterior probability density, similar (but not identical) to a confidence interval. Probability of direction is the proportion of posterior probability density that falls above or below zero, depending on whether the parameter estimate is positive or negative. A value approaching 1.0 can be interpreted as "completely confident in the direction of the effect," and a value of 0.5 would mean "completely uncertain about the direction of the effect." As standard deviation parameters are positive by definition, p_d is not reported in Tables 4 and 5.

A Bayes Factor is a ratio of the marginal likelihoods of two statistical models. Essentially, it tells us the strength of evidence provided by our data for one model over the other. When the prior probabilities of the two models are equal, the Bayes Factor reduces to the ratio of the models' posterior probabilities. This allows us to conduct a *Bayesian hypothesis test*. Suppose we have data D, a model M, and a null hypothesis that a particular model parameter θ is equal to θ_0. Using Markov chain Monte Carlo (MCMC) sampling, we can estimate the posterior of a model M_0 with parameter $\theta = \theta_0$, and the posteror of the full model M. The Bayes Factor, representing the amount of evidence for the null hypothesis model M_0 over the full model M, is:

$$BF = \frac{p(D \mid M_0)}{p(D \mid M)} \tag{21}$$

However, this ratio can also be approximated without estimating the full posterior for the null hypothesis model. Instead, we can simply divide the posterior density for θ by the prior density for θ at the value of the null hypothesis. This is known as the Savage-Dickey density ratio (Dickey & Lientz 1970), and is essentially a ratio of the probability of a particular parameter value *after* observing data to the probability of that parameter value *before* observing data:

$$BF = \frac{p(D \mid M_0)}{p(D \mid M)} \approx \frac{p(\theta = \theta_0 \mid D, M)}{p(\theta = \theta_0 \mid M)} \tag{22}$$

A Bayes Factor less than 1.0 can be interpreted as evidence for the null hypothesis, and a Bayes Factor greater than 1.0 can be interpreted as evidence against the null hypothesis. A typical scale for interpreting Bayes Factors suggests that a value of less than $10^{-1/2}$ (approximately 1/3) or greater than 3 constitutes "substantial" evidence (Jeffreys 1961). A Bayes Factor close to 1.0 does not provide substantial evidence for either model under consideration. For more in-depth coverage of Bayesian hypothesis testing, see Wagenmakers et al. (2010).

[9] For convenience, parameter names defined in §3.3 (i.e. $\beta_{\text{sim}|C}$) are used to mean MCMC estimates of the parameter (i.e. $\hat{\beta}_{\text{sim}|C}$ or $\mathbb{E}[\beta_{\text{sim}|C}]$).

Several plots are also presented to aid in the interpretation of results. Because the model contains separate parameters for main effects, random effects by language family, and random effects by language within family, these parameters need to be added together to obtain the total effect of consonant or vowel similarity or identity for an individual language. For example, the total effect for consonant similarity would be $\beta_{\text{sim}|C} + \gamma_{\text{sim}|C} + \alpha_{\text{sim}|C}$. Figure 4 shows the posterior densities of these "total effects". Consonant effects are plotted on the x-axes, and vowel effects on the y-axes to show how these effects are related.

Figures 7 and 8 show predicted O/E ratios for each language as a function of similarity, calculated from model predictions according to Equation 3. These predictions were obtained by calculating the value of $|\text{C}_{\text{inv}}|$, $|\text{V}_{\text{inv}}|$, and the mean of $\frac{O_{x+}}{N} \cdot \frac{O_{+y}}{N}$ across all pairs of segments for each language, and generating samples from the models for similarity values between the minimum similarity by language (0.0 for **NC**, varying for **Feat**) and 0.995, and samples for similarity 0.0 and identity 1.0. For ease of interpretation, several languages are highlighted in these figures: Turkish, which has vowel harmony; Basque, which has consonant harmony (Hualde 1991; Hansson 2010); Arabic, where similar consonant co-occurrence is restricted (Frisch et al. 2004); Spanish, where similar and identical vowel pairs are underattested (Walter 2010); and English, for reference.

4.1 Vowel effects

4.1.1 Vowel similarity effects

Both the Natural Class Similarity (**NC**) and Feature Similarity (**Feat**) models do not show a clearly positive or negative effect for vowel similarity, as shown in Tables 2 and 3 and Figure 5 (**NC**: $\beta_{\text{sim}|V} = 0.05$, 95% CredI $= [-0.04, 0.14]$, BF $= 2.07 \times 10^{-2}$; **Feat**: $\beta_{\text{sim}|V} = 0.05$, 95% CredI $= [-0.18, 0.28]$, BF $= 3.20 \times 10^{-2}$). In both models, 95% credible intervals contain zero. In addition, both models have small Bayes factors (BF $<$ 0.1), suggesting that we can be fairly confident that there is in fact zero vowel similarity effect.

This is further supported by the finding that vowel similarity effects do not vary across either families (**NC**: $\sigma(\gamma_{\text{sim}|V}) = 0.06$, 95% CredI $= [0.00, 0.18]$, BF $= 2.80 \times 10^{-2}$; **Feat**: $\sigma(\gamma_{\text{sim}|V}) = 0.16$, 95% CredI $= [0.01, 0.35]$, BF $= 1.03 \times 10^{-1}$) or languages within families (**NC**: $\sigma(\alpha_{\text{sim}|V}) = 0.06$, 95% CredI $= [0.00, 0.17]$, BF $= 3.15 \times 10^{-2}$; **Feat**: $\sigma(\alpha_{\text{sim}|V}) = 0.10$, 95% CredI $= [0.00, 0.24]$, BF $= 5.03 \times 10^{-2}$), as shown in Tables 4 and 5 and Figure 5. In all cases, the estimated effects have small Bayes factors. In Figure 4, we can see that the total vowel similarity effect for languages (calculated by adding the main effect $\beta_{\text{sim}|V}$, the by-family random effect $\gamma_{\text{sim}|V}$, and the by-language random effect $\alpha_{\text{sim}|V}$) is centered around zero with relatively small variance. This zero or near-zero total effect has a miniscule impact on predicted O/E ratios, as can be seen in Figures 7 and 8. As vowel similarity increases, the estimated O/E ratio remains at 1.0, meaning that vowel pairs occur at the expected rate, regardless of similarity. This result is particularly striking, given the prevalence of vowel harmony in Northern Eurasian languages.[10] For example, Turkish, a language with extensive vowel harmony, appears on Figures 7 and 8 with a near-zero slope. Feature-specific vowel harmony has a small effect on overall local vowel co-occurrence, and does not imply a general tendency for similar vowel co-occurrence:

[10] As pointed out by a reviewer, 40/107 languages belong to families that are well-known for having vowel harmony: Uralic, Turkic, Mongolian and Tungusic.

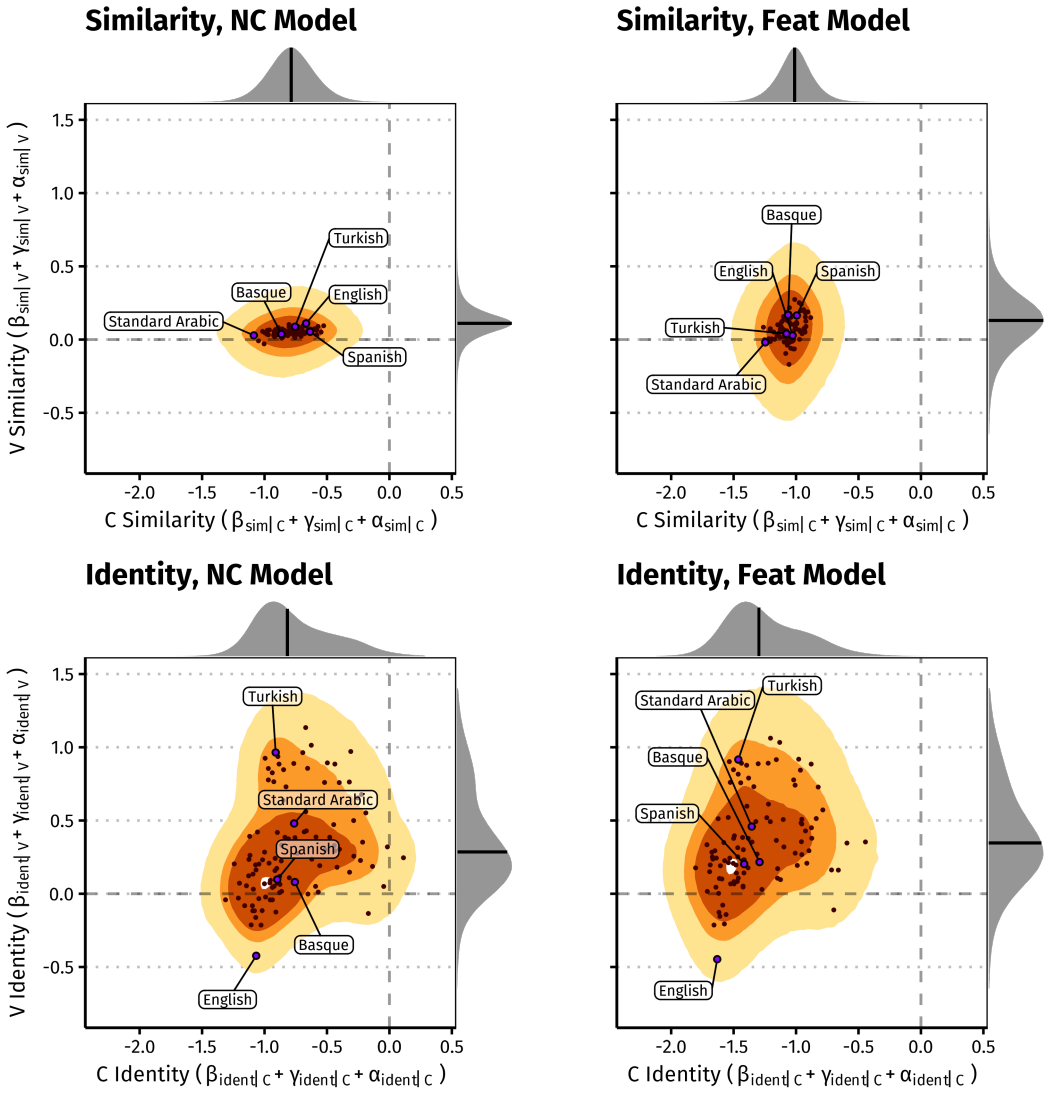

Figure 4: Estimated total similarity and identity effects across languages. Ellipses show 95%, 75%, and 50% credible intervals for joint distribution of total vowel and consonant effects. Density plots on axes show posterior distributions of total vowel or consonant effects. Points display mean posterior estimates for each of the 107 languages included in NorthEuraLex. Languages mentioned in text are highlighted.

Despite having restrictions on co-occurrence, vowels in harmony languages maintain contrast.

4.1.2 Vowel identity effects

Despite there being no vowel similarity effect, we observe a consistently positive vowel identity effect in both models, as shown in Tables 2 and 3 and Figure 5. Identical pairs of vowels are more likely to co-occur than non-identical pairs (**NC**: $\beta_{\text{ident}|V} = 0.34$, 95% CredI $= [0.20, 0.49]$, BF $= 1.10 \times 10^2$; **Feat**: $\beta_{\text{ident}|V} = 0.37$, 95% CredI $= [0.14, 0.60]$, BF $= 3.58$). This vowel identity effect has large variance across language families (**NC**: $\sigma(\gamma_{\text{ident}|V}) = 0.37$, 95% CredI $= [0.25, 0.52]$, BF $= 5.16 \times 10^{14}$; **Feat**: $\sigma(\gamma_{\text{ident}|V}) = 0.35$,

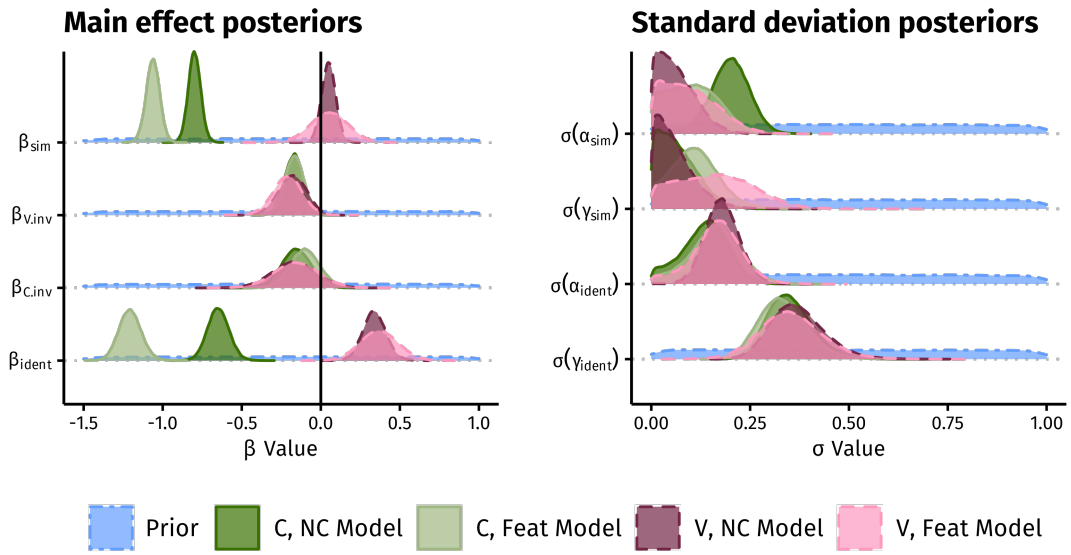

Figure 5: Posterior densities of main effect (β) and standard devation (σ) parameters, with priors for reference. For example, β_{sim} for C (in **NC** and **Feat** models) refers to $\beta_{\text{sim}|C}$ in Equation 10.

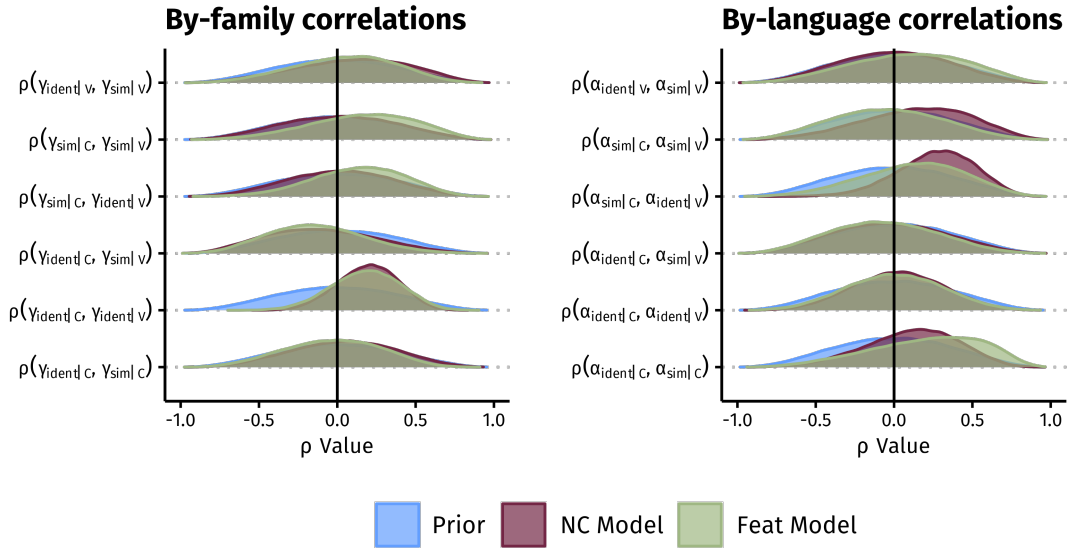

Figure 6: Posterior densities of correlation (ρ) parameters, with priors for reference.

Parameter	Estimate	95% CredI	p_d	Bayes Factor
$\beta_{\text{intercept}\|C}$	8.51	[7.83, 9.19]	1.00	NA
$\beta_{\text{intercept}\|V}$	8.08	[7.08, 9.10]	1.00	NA
$\beta_{\text{sim}\|C}$	−0.80	[−0.88, −0.72]	1.00	5.20×10^{16}
$\beta_{\text{sim}\|V}$	0.05	[−0.04, 0.14]	0.87	2.07×10^{-2}
$\beta_{\text{ident}\|C}$	−0.65	[−0.78, −0.51]	1.00	∞
$\beta_{\text{ident}\|V}$	0.34	[0.20, 0.49]	1.00	1.10×10^{2}
$\beta_{\text{C.inv}\|C}$	−0.15	[−0.34, 0.03]	0.95	9.45×10^{-2}
$\beta_{\text{C.inv}\|V}$	−0.17	[−0.44, 0.10]	0.89	7.59×10^{-2}
$\beta_{\text{V.inv}\|C}$	−0.17	[−0.29, −0.06]	1.00	9.66×10^{-1}
$\beta_{\text{V.inv}\|V}$	−0.18	[−0.36, −0.01]	0.98	1.80×10^{-1}

Table 2: NC Model fixed effect estimates. Each row corresponds to a parameter in Equations 10 and 11.

Parameter	Estimate	95% CredI	p_d	Bayes Factor
$\beta_{\text{intercept}\|C}$	8.87	[8.17, 9.57]	1.00	NA
$\beta_{\text{intercept}\|V}$	8.06	[7.01, 9.12]	1.00	NA
$\beta_{\text{sim}\|C}$	−1.06	[−1.15, −0.98]	1.00	1.46×10^{15}
$\beta_{\text{sim}\|V}$	0.05	[−0.18, 0.28]	0.68	3.20×10^{-2}
$\beta_{\text{ident}\|C}$	−1.21	[−1.34, −1.06]	1.00	∞
$\beta_{\text{ident}\|V}$	0.37	[0.14, 0.60]	1.00	3.58
$\beta_{\text{C.inv}\|C}$	−0.10	[−0.29, 0.09]	0.86	4.25×10^{-2}
$\beta_{\text{C.inv}\|V}$	−0.15	[−0.43, 0.13]	0.86	6.63×10^{-2}
$\beta_{\text{V.inv}\|C}$	−0.17	[−0.29, −0.05]	1.00	7.09×10^{-1}
$\beta_{\text{V.inv}\|V}$	−0.21	[−0.39, −0.03]	0.99	3.14×10^{-1}

Table 3: Feat Model fixed effect estimates. Each row corresponds to a parameter in Equations 10 and 11.

95% CredI $= [0.20, 0.52]$, BF $= 5.17 \times 10^{4}$). There is also evidence of variation across languages within families in the **NC** model, and weaker evidence for the **Feat** model (**NC**: $\sigma(\alpha_{\text{ident}|V}) = 0.18$, 95% CredI $= [0.10, 0.27]$, BF $= 2.23 \times 10^{1}$; **Feat**: $\sigma(\alpha_{\text{ident}|V}) = 0.17$, 95% CredI $= [0.03, 0.29]$, BF $= 2.97 \times 10^{-1}$), as shown in Tables 4 and 5 and Figure 5. By examining the posterior distribution of the difference between parameters, we can see that in both models, standard deviations for the vowel identity effect are larger across language families than languages within families (**NC**: $\sigma(\gamma_{\text{ident}|V}) - \sigma(\alpha_{\text{ident}|V}) = 0.19$, 95% CredI $= [0.06, 0.33]$, BF $= 1.05 \times 10^{2}$; **Feat**: $\sigma(\gamma_{\text{ident}|V}) - \sigma(\alpha_{\text{ident}|V}) = 0.19$, 95% CredI $= [0.02, 0.36]$, BF $= 2.75 \times 10^{1}$).[11]

The impact of this large variance in vowel identity effects can be seen in Figure 4. Although the main effect ($\beta_{\text{ident}|V}$) is positive and the majority of languages are predicted to have a positive vowel identity effect, the by-family and by-language variance allows for languages with negative vowel identity effects. This is reflected in Figures 7 and 8, where a majority of languages are predicted to have an O/E ratio greater than 1.0, but some have an O/E of less than 1.0.

[11] This method of examining the posterior distribution of a quantity of interest is used repeatedly to report differences in degrees of variability or differences in effect size.

4.1.3 Relationships between vowel similarity and identity effects

Identity and similarity effects in vowels appear to follow drastically different distributions across languages. In both models, identity effects are larger in magnitude than similarity effects (**NC**: $|\beta_{\text{ident}|V}| - |\beta_{\text{sim}|V}| = 0.28$, 95% CredI $= [0.16, 0.41]$, BF $= 7.50 \times 10^3$; **Feat**: $|\beta_{\text{ident}|V}| - |\beta_{\text{sim}|V}| = 0.27$, 95% CredI $= [0.05, 0.42]$, BF $= 3.27 \times 10^1$), as can be seen in the y-axes of Figure 5. Vowel identity effects are not equivalent to linearly extrapolating similarity effects to 1.0. This can be seen in Figures 7 and 8, where the slope of O/E ratio over similarity is not equal to the slope between similarity 0.995 and identity.

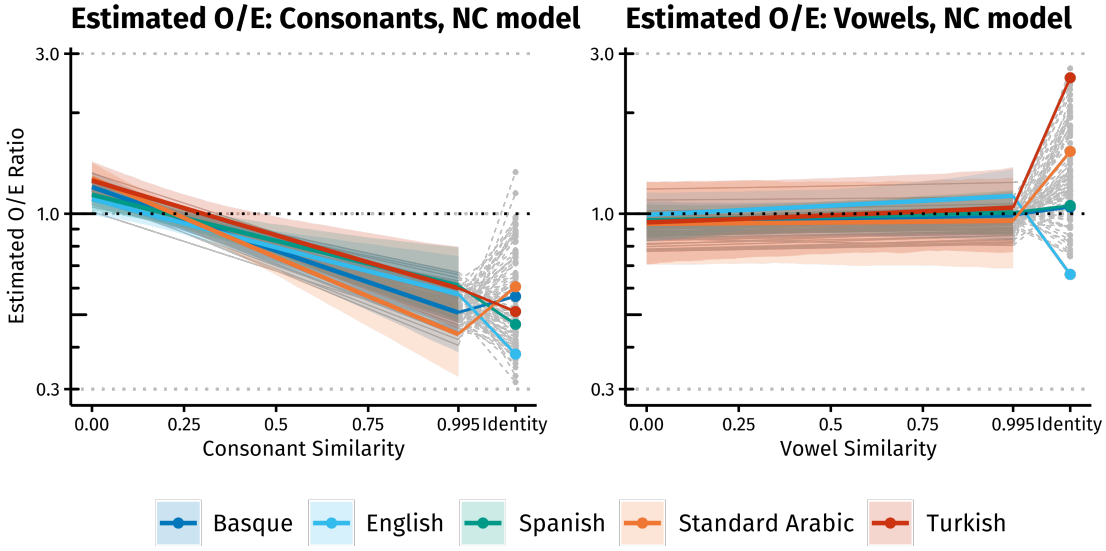

Figure 7: Estimated O/E ratios computed from NC Model for similarity values between 0.00 and 0.995, and identity. Each line represents a language from NorthEuraLex, with mentioned languages highlighted. This visualization shows both the similarity and identity effects for each language, as the left (0-0.005) and right (0.995-Identity) segments.

Where all languages in our sample have no detectable vowel similarity effect, identical vowel co-occurrence effects vary in both magnitude and direction. This is confirmed by comparing standard deviations for vowel identity effects to standard deviations for vowel similarity effects. Identity effects vary more than similarity effects across both language families (**NC**: $\sigma(\gamma_{\text{ident}|V}) - \sigma(\gamma_{\text{sim}|V}) = 0.31$, 95% CredI $= [0.18, 0.44]$, BF $= 4.00 \times 10^3$; **Feat**: $\sigma(\gamma_{\text{ident}|V}) - \sigma(\gamma_{\text{sim}|V}) = 0.19$, 95% CredI $= [-0.02, 0.39]$, BF $= 1.39 \times 10^1$) and languages within families, although less clearly in the **Feat** model (**NC**: $\sigma(\alpha_{\text{ident}|V}) - \sigma(\alpha_{\text{sim}|V}) = 0.11$, 95% CredI $= [0.00, 0.21]$, BF $= 1.97 \times 10^1$; **Feat**: $\sigma(\alpha_{\text{ident}|V}) - \sigma(\alpha_{\text{sim}|V}) = 0.07$, 95% CredI $= [-0.12, 0.20]$, BF $= 3.53$). This can be seen in the standard deviation posteriors shown in Figure 5.

Finally, as shown in Tables 6 and 7 and Figure 6, there are no clear non-zero correlations between vowel similarity and vowel identity effects, either across families (**NC**: $\rho(\gamma_{\text{sim}|V}, \gamma_{\text{ident}|V}) = 0.09$, 95% CredI $= [-0.61, 0.72]$, BF $= 1.04$; **Feat**: $\rho(\gamma_{\text{sim}|V}, \gamma_{\text{ident}|V}) = 0.05$, 95% CredI $= [-0.60, 0.63]$, BF $= 0.90$) or across languages (**NC**: $\rho(\alpha_{\text{sim}|V}, \alpha_{\text{ident}|V}) = 0.00$, 95% CredI $= [-0.65, 0.64]$, BF $= 0.93$; **Feat**: $\rho(\alpha_{\text{sim}|V}, \alpha_{\text{ident}|V}) = 0.09$, 95% CredI $= [-0.61, 0.73]$, BF $= 1.03$). While the correlations estimated by the models are near zero, the Bayes factors for both are close to 1.0, suggesting that we do not have enough

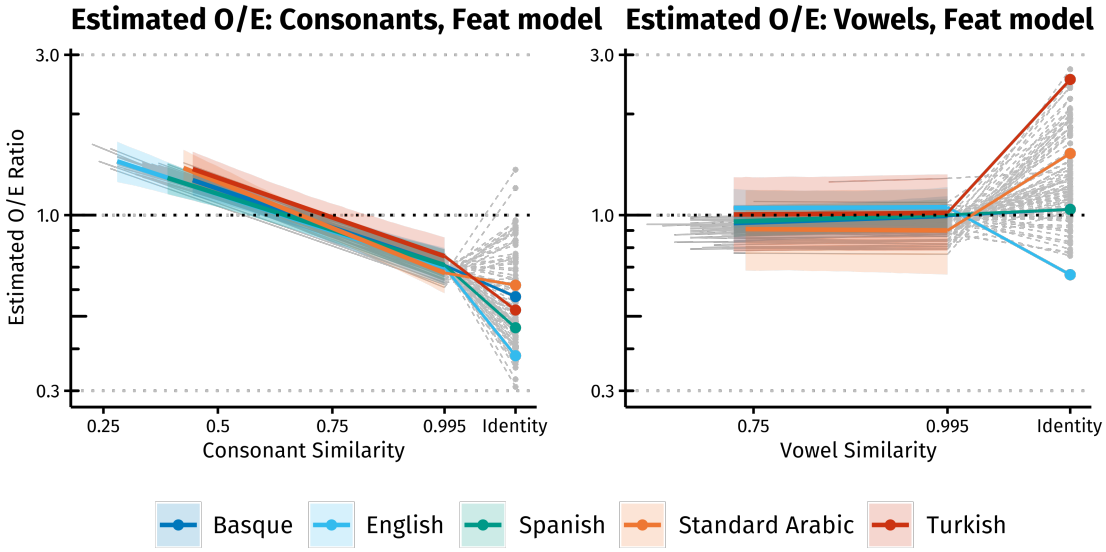

Figure 8: Estimated O/E ratios computed from Feat Model for similarity values between 0.00 and 0.995, and identity. Each line represents a language from NorthEuraLex, with mentioned languages highlighted. This visualization shows both the similarity and identity effects for each language, as the left (0-0.005) and right (0.995-Identity) segments.

evidence to support the null hypothesis that vowel similarity and identity effects are not correlated. In Figure 6, we can see that the posteriors for these correlations are very similar to the prior distribution. This suggests that more data is needed to make any conclusions about these correlations.

4.2 Consonant effects

4.2.1 Consonant similarity effects

Both models predict a negative consonant similarity effect, as shown in Tables 2 and 3 and Figure 5 (**NC**: $\beta_{\text{sim}|C} = -0.80$, 95% CredI $= [-0.88, -0.72]$, BF $= 5.20 \times 10^{16}$; **Feat**: $\beta_{\text{sim}|C} = -1.06$, 95% CredI $= [-1.15, -0.98]$, BF $= 1.46 \times 10^{15}$). This implies that more similar pairs of non-identical consonants are less likely to co-occur. Like we saw in vowel similarity effects, there is evidence that consonant similarity effects do not vary by language family (**NC**: $\sigma(\gamma_{\text{sim}|C}) = 0.07$, 95% CredI $= [0.00, 0.18]$, BF $= 3.33 \times 10^{-2}$; **Feat**: $\sigma(\gamma_{\text{sim}|C}) = 0.11$, 95% CredI $= [0.01, 0.23]$, BF $= 1.06 \times 10^{-1}$), as shown in Tables 4 and 5 and Figure 5. Variance across languages within families is less clear. In the **NC** model, there is evidence of non-zero standard deviation across languages (**NC**: $\sigma(\alpha_{\text{sim}|C}) = 0.20$, 95% CredI $= [0.09, 0.29]$, BF $= 5.54$), but in the **Feat** model we see evidence of zero variation across languages (**Feat**: $\sigma(\alpha_{\text{sim}|C}) = 0.10$, 95% CredI $= [0.01, 0.23]$, BF $= 5.96 \times 10^{-2}$). We also see evidence that families vary more than languages in the **NC** model, but not in the **Feat** model (**NC**: $\sigma(\gamma_{\text{sim}|C}) - \sigma(\alpha_{\text{sim}|C}) = -0.13$, 95% CredI $= [-0.24, 0.01]$, BF $= 6.00 \times 10^{-2}$; **Feat**: $\sigma(\gamma_{\text{sim}|C}) - \sigma(\alpha_{\text{sim}|C}) = 0.01$, 95% CredI $= [-0.15, 0.16]$, BF $= 1.12$), as shown in Figure 5.

Despite differences in standard deviation estimates, the total consonant similarity effects by language plotted in Figure 4 are qualitatively similar. For both models, all languages

Parameter	Estimate	95% CredI	Bayes Factor
$\gamma_{\text{intercept}\|C}$	0.13	[0.05, 0.20]	4.44×10^{-1}
$\gamma_{\text{intercept}\|V}$	0.26	[0.16, 0.37]	3.86×10^{14}
$\gamma_{\text{sim}\|C}$	0.07	[0.00, 0.18]	3.33×10^{-2}
$\gamma_{\text{sim}\|V}$	0.06	[0.00, 0.18]	2.80×10^{-2}
$\gamma_{\text{ident}\|C}$	0.35	[0.25, 0.48]	3.44×10^{14}
$\gamma_{\text{ident}\|V}$	0.37	[0.25, 0.52]	5.16×10^{14}
$\gamma_{\phi\|C}$	1.07	[0.84, 1.36]	∞
$\gamma_{\phi\|V}$	2.05	[1.52, 2.72]	2.23×10^{15}
$\alpha_{\text{intercept}\|C}$	0.18	[0.15, 0.22]	2.70×10^{15}
$\alpha_{\text{intercept}\|V}$	0.24	[0.19, 0.30]	1.87×10^{17}
$\alpha_{\text{sim}\|C}$	0.20	[0.09, 0.29]	5.54
$\alpha_{\text{sim}\|V}$	0.06	[0.00, 0.17]	3.15×10^{-2}
$\alpha_{\text{ident}\|C}$	0.13	[0.01, 0.24]	1.70×10^{-1}
$\alpha_{\text{ident}\|V}$	0.18	[0.10, 0.27]	2.23×10^{1}
$\alpha_{\phi\|C}$	0.25	[0.19, 0.31]	∞
$\alpha_{\phi\|V}$	1.08	[0.87, 1.34]	∞

Table 4: NC Model random effect standard deviation estimates. Each row represents the standard deviation of language family random effects (γ parameters), or language within family random effects (α parameters), corresponding to parameters in Equations 10 and 11.

Parameter	Estimate	95% CredI	Bayes Factor
$\gamma_{\text{intercept}\|C}$	0.14	[0.02, 0.24]	2.71×10^{-1}
$\gamma_{\text{intercept}\|V}$	0.21	[0.03, 0.37]	3.29×10^{-1}
$\gamma_{\text{sim}\|C}$	0.11	[0.01, 0.23]	1.06×10^{-1}
$\gamma_{\text{sim}\|V}$	0.16	[0.01, 0.35]	1.03×10^{-1}
$\gamma_{\text{ident}\|C}$	0.33	[0.22, 0.47]	2.67×10^{14}
$\gamma_{\text{ident}\|V}$	0.35	[0.20, 0.52]	5.17×10^{4}
$\gamma_{\phi\|C}$	1.09	[0.86, 1.40]	∞
$\gamma_{\phi\|V}$	2.03	[1.51, 2.71]	5.90×10^{15}
$\alpha_{\text{intercept}\|C}$	0.20	[0.15, 0.28]	8.21×10^{14}
$\alpha_{\text{intercept}\|V}$	0.25	[0.16, 0.35]	3.29×10^{1}
$\alpha_{\text{sim}\|C}$	0.10	[0.01, 0.23]	5.96×10^{-2}
$\alpha_{\text{sim}\|V}$	0.10	[0.00, 0.24]	5.03×10^{-2}
$\alpha_{\text{ident}\|C}$	0.16	[0.02, 0.28]	2.85×10^{-1}
$\alpha_{\text{ident}\|V}$	0.17	[0.03, 0.29]	2.97×10^{-1}
$\alpha_{\phi\|C}$	0.25	[0.19, 0.31]	1.35×10^{14}
$\alpha_{\phi\|V}$	1.08	[0.87, 1.34]	6.60×10^{15}

Table 5: Feat Model random effect standard deviation estimates. Each row represents the standard deviation of language family random effects (γ parameters), or language within family random effects (α parameters), corresponding to parameters in Equations 10 and 11.

have a clearly negative consonant similarity effect. In Figures 7 and 8, this corresponds to every language having a negative slope in O/E ratio as consonant similarity increases. More similar pairs of consonants occur less often than expected (O/E < 1.0), while less similar pairs occur either more than expected (O/E > 1.0) or about as often as expected (O/E $= 1.0$). Even languages with known consonant harmony processes, such as Basque, have a negative consonant similarity effect in this model. This may be because consonant harmony generally only targets segments that are highly similar to begin with (i.e., stridents, nasals, liquids). Co-occurrence of these similar segments, which only differ in the targeted features, is *disfavored* in languages with consonant harmony.[12] Overall, our results suggest that like vowel harmony, consonant harmony for a specific feature has a small effect on overall consonant co-occurrence, and there is still a strong restriction against similar consonants co-occurring.

4.2.2 Consonant identity effects

Consonant identity effects are also negative in both models, as shown in Tables 2 and 3 and Figure 5, meaning that identical consonant pairs are less likely to co-occur than non-identical pairs (**NC**: $\beta_{\text{ident}|C} = -0.65$, 95% CredI $= [-0.78, -0.51]$, BF $= \infty$; **Feat**: $\beta_{\text{ident}|C} = -1.21$, 95% CredI $= [-1.34, -1.06]$, BF $= \infty$). These effects vary across language families (**NC**: $\sigma(\gamma_{\text{ident}|C}) = 0.35$, 95% CredI $= [0.25, 0.48]$, BF $= 3.44 \times 10^{14}$; **Feat**: $\sigma(\gamma_{\text{ident}|C}) = 0.33$, 95% CredI $= [0.22, 0.47]$, BF $= 2.67 \times 10^{14}$), as shown in Tables 4 and 5 and Figure 5. Across languages within families, there is evidence that consonant identity effects do not vary in the **NC** model ($\sigma(\alpha_{\text{ident}|C}) = 0.13$, 95% CredI $= [0.01, 0.24]$, BF $= 1.70 \times 10^{-1}$), but only weak evidence against variation in the **Feat** model ($\sigma(\alpha_{\text{ident}|C}) = 0.16$, 95% CredI $= [0.02, 0.28]$, BF $= 2.85 \times 10^{-1}$). Like vowel identity effects, there is evidence that consonant identity effects vary more across families than across languages within families (**NC**: $\sigma(\gamma_{\text{ident}|C}) - \sigma(\alpha_{\text{ident}|C}) = 0.22$, 95% CredI $= [0.07, 0.37]$, BF $= 1.60 \times 10^{2}$; **Feat**: $\sigma(\gamma_{\text{ident}|C}) - \sigma(\alpha_{\text{ident}|C}) = 0.17$, 95% CredI $= [0.02, 0.34]$, BF $= 2.73 \times 10^{1}$).

In Figure 4, we can see the total consonant identity effect across languages. Although the main effect is negative, accounting for variance across families and languages allows for a very small number of languages to have a positive total consonant identity effect in the **NC** model. In the **Feat** model, languages with positive consonant identity effects are even less likely, with none falling within the 95% credible interval. In Figures 7 and 8, nearly all languages have an O/E ratio of less than 1.0, with the exception of two languages. While rare, it is possible for identical consonant pairs to occur more often than expected in a language.

4.2.3 Relationships between consonant similarity and identity effects

Like in vowels, identity and similarity effects follow different distributions in consonants as well. In Figures 7 and 8, we see that linear extrapolation of the consonant similarity effect does not provide the same O/E ratio as the one predicted by the consonant identity effect. Identical consonants can be over-attested in relation to highly similar vowels (shown as a positive slope between similarity 0.995 and identity) or under-attested (a negative slope between similarity 0.995 and identity). As expected from the results of McCarthy (1986), identical consonant pairs in Standard Arabic are over-attested when compared to highly similar consonant pairs.

[12] Thank you to a reviewer for pointing this out.

As shown in Figure 5, the consonant identity effect in the **NC** model is larger than the similarity effect (**NC**: $|\beta_{\text{ident}|C}| - |\beta_{\text{sim}|C}| = -0.15$, 95% CredI $= [-0.28, -0.03]$, BF $= 2.00 \times 10^{-2}$), but in the **Feat** model, the opposite is true (**Feat**: $|\beta_{\text{ident}|C}| - |\beta_{\text{sim}|C}| = 0.14$, 95% CredI $= [0.03, 0.26]$, BF $= 4.27 \times 10^{1}$). In both models, consonant identity effects vary more than similarity effects across language families (**NC**: $\sigma(\gamma_{\text{ident}|C}) - \sigma(\gamma_{\text{sim}|C}) = 0.28$, 95% CredI $= [0.16, 0.40]$, BF $= 3.53 \times 10^{3}$; **Feat**: $\sigma(\gamma_{\text{ident}|C}) - \sigma(\gamma_{\text{sim}|C}) = 0.22$, 95% CredI $= [0.09, 0.36]$, BF $= 4.83 \times 10^{2}$). Across languages within families, there is no strong evidence that standard deviation estimates do or do not differ between identity and similarity effects (**NC**: $\sigma(\alpha_{\text{ident}|C}) - \sigma(\alpha_{\text{sim}|C}) = -0.06$, 95% CredI $= [-0.19, 0.06]$, BF $= 2.50 \times 10^{-1}$; **Feat**: $\sigma(\alpha_{\text{ident}|C}) - \sigma(\alpha_{\text{sim}|C}) = 0.06$, 95% CredI $= [-0.08, 0.18]$, BF $= 3.32$).

We also see no evidence of a correlation between consonant similarity and identity effects, across families (**NC**: $\rho(\gamma_{\text{sim}|C}, \gamma_{\text{ident}|C}) = 0.04$, 95% CredI $= [-0.60, 0.64]$, BF $= 0.89$; **Feat**: $\rho(\gamma_{\text{sim}|C}, \gamma_{\text{ident}|C}) = -0.03$, 95% CredI $= [-0.63, 0.54]$, BF $= 0.84$) or languages within families (**NC**: $\rho(\alpha_{\text{sim}|C}, \alpha_{\text{ident}|C}) = 0.11$, 95% CredI $= [-0.52, 0.63]$, BF $= 0.88$; **Feat**: $\rho(\alpha_{\text{sim}|C}, \alpha_{\text{ident}|C}) = 0.20$, 95% CredI $= [-0.54, 0.77]$, BF $= 1.25$), as shown in Figure 6. Just like the previous results for correlations between vowel effects, we see Bayes factors near zero, and cannot conclude that consonant similarity and identity are or are not correlated without more data.

4.3 Relationships between vowel and consonant effects

Although vowel and consonant co-occurrence effects are in different directions, we can compare the absolute value of the effect sizes because both are measured on the same similarity scale. Consonant identity effects are on average larger than vowel identity effects in both models (**NC**: $|\beta_{\text{ident}|C}| - |\beta_{\text{ident}|V}| = 0.31$, 95% CredI $= [0.13, 0.48]$, BF $= 2.92 \times 10^{2}$; **Feat**: $|\beta_{\text{ident}|C}| - |\beta_{\text{ident}|V}| = 0.84$, 95% CredI $= [0.60, 1.07]$, BF $= \infty$). Identical consonant co-occurrence is disfavored to a greater degree than identical vowel co-occurrence is preferred. The same is true for similarity effects. In both models, consonant similarity effects are larger than vowel similarity effects (**NC**: $|\beta_{\text{sim}|C}| - |\beta_{\text{sim}|V}| = 0.74$, 95% CredI $= [0.65, 0.83]$, BF $= \infty$; **Feat**: $|\beta_{\text{sim}|C}| - |\beta_{\text{sim}|V}| = 0.96$, 95% CredI $= [0.80, 1.09]$, BF $= \infty$). This suggests that across languages, consonant co-occurrence is more strongly constrained than vowel co-occurrence.

We also see no evidence of correlations between vowel and consonant effects, as shown in Tables 6 and 7 and Figure 6: All credible intervals include zero, and all Bayes factors are between $1/3$ and 3. There is no evidence of a correlation between consonant and vowel similarity effects across language families or across languages within families. There is also no evidence of a correlation between consonant and vowel identity effects across families or languages. As with previous results for correlations, we cannot conclude anything about correlations between consonant and vowel effects without more data.

5 Discussion

5.1 Similarity effects

We have shown that across a large set of languages, similar consonant co-occurrence is strongly restricted. On average, similar pairs of consonants occur less frequently than dissimilar pairs of consonants. We also found this effect to have little to no variance across languages: In this sense, it is universal. These findings align well with previous

Parameter 1	Parameter 2	Estimate	95% CredI		p_d	Bayes Factor
$\gamma_{\text{sim}\|C}$	$\gamma_{\text{sim}\|V}$	0.01	[−0.66,	0.67]	0.51	0.98
$\gamma_{\text{sim}\|C}$	$\gamma_{\text{ident}\|C}$	0.04	[−0.60,	0.64]	0.55	0.89
$\gamma_{\text{sim}\|C}$	$\gamma_{\text{ident}\|V}$	0.02	[−0.61,	0.64]	0.52	0.89
$\gamma_{\text{sim}\|V}$	$\gamma_{\text{ident}\|C}$	−0.10	[−0.71,	0.58]	0.62	1.01
$\gamma_{\text{sim}\|V}$	$\gamma_{\text{ident}\|V}$	0.09	[−0.61,	0.72]	0.61	1.04
$\gamma_{\text{ident}\|C}$	$\gamma_{\text{ident}\|V}$	0.19	[−0.20,	0.55]	0.84	0.83
$\alpha_{\text{sim}\|C}$	$\alpha_{\text{sim}\|V}$	0.14	[−0.56,	0.74]	0.67	1.05
$\alpha_{\text{sim}\|C}$	$\alpha_{\text{ident}\|C}$	0.11	[−0.52,	0.63]	0.66	0.88
$\alpha_{\text{sim}\|C}$	$\alpha_{\text{ident}\|V}$	0.27	[−0.26,	0.70]	0.86	1.23
$\alpha_{\text{sim}\|V}$	$\alpha_{\text{ident}\|C}$	−0.03	[−0.67,	0.64]	0.54	0.95
$\alpha_{\text{sim}\|V}$	$\alpha_{\text{ident}\|V}$	0.00	[−0.65,	0.64]	0.50	0.93
$\alpha_{\text{ident}\|C}$	$\alpha_{\text{ident}\|V}$	0.02	[−0.54,	0.58]	0.53	0.75

Table 6: NC Model correlation estimates. Each row represents a correlation between random effect parameters across language family (γ parameters), or language within family (α parameters). Additional correlation estimates not relevant to results are listed in Appendix C.

Parameter 1	Parameter 2	Estimate	95% CredI		p_d	Bayes Factor
$\gamma_{\text{sim}\|C}$	$\gamma_{\text{sim}\|V}$	0.16	[−0.53,	0.74]	0.68	1.06
$\gamma_{\text{sim}\|C}$	$\gamma_{\text{ident}\|C}$	−0.03	[−0.63,	0.54]	0.53	0.84
$\gamma_{\text{sim}\|C}$	$\gamma_{\text{ident}\|V}$	0.13	[−0.48,	0.67]	0.68	0.90
$\gamma_{\text{sim}\|V}$	$\gamma_{\text{ident}\|C}$	−0.13	[−0.69,	0.50]	0.68	0.92
$\gamma_{\text{sim}\|V}$	$\gamma_{\text{ident}\|V}$	0.05	[−0.60,	0.63]	0.57	0.90
$\gamma_{\text{ident}\|C}$	$\gamma_{\text{ident}\|V}$	0.18	[−0.26,	0.59]	0.79	0.82
$\alpha_{\text{sim}\|C}$	$\alpha_{\text{sim}\|V}$	−0.04	[−0.68,	0.63]	0.54	0.95
$\alpha_{\text{sim}\|C}$	$\alpha_{\text{ident}\|C}$	0.20	[−0.54,	0.77]	0.72	1.25
$\alpha_{\text{sim}\|C}$	$\alpha_{\text{ident}\|V}$	0.12	[−0.56,	0.70]	0.65	0.99
$\alpha_{\text{sim}\|V}$	$\alpha_{\text{ident}\|C}$	−0.06	[−0.69,	0.59]	0.58	0.94
$\alpha_{\text{sim}\|V}$	$\alpha_{\text{ident}\|V}$	0.09	[−0.61,	0.73]	0.60	1.03
$\alpha_{\text{ident}\|C}$	$\alpha_{\text{ident}\|V}$	0.00	[−0.58,	0.59]	0.50	0.80

Table 7: Feat Model correlation estimates. Each row represents a correlation between random effect parameters across language family (γ parameters), or language within family (α parameters). Additional correlation estimates not relevant to results are listed in Appendix D.

work showing that OCP effects occur frequently, while consonant harmony is rare and feature-specific (Frisch et al. 2004; Pozdniakov & Segerer 2007; Hansson 2010; Graff 2012).

It may be the case that these effects are the result of feature-specific co-occurrence restrictions such as OCP-Place, but they are large enough to be identified with two different aggregate similarity metrics. This suggests that it may be accurate to characterize the OCP as a universal gradient constraint on similar consonant co-occurrence, rather than a categorical constraint on any particular feature. However, confirming this would require testing whether a model of individual feature co-occurrence provides a significantly better fit to the data than aggregate similarity models. We leave this for future work.

In vowels, we do not see such a robust effect of similarity on co-occurrence – in fact, we see no effect at all. Pairs of non-identical vowels occur about as often as expected, regardless of similarity. We also find little to no variance in vowel similarity effects across languages, suggesting that vowel co-occurrence is universally *not* restricted by similarity. Although this result does not appear to align with the prevalence of vowel harmony across languages, this discrepancy may follow from the fact that harmony is best characterized as operating over individual features, not aggregate similarity. Alternatively, this could be due to the properties of the NorthEuraLex dataset. Most entries are lemmas, which are often equivalent to roots. In many languages, vowel harmony only emerges in inflected forms, while roots are not harmonic. Future work with a larger dataset and a model of individual feature co-occurrence would allow us to assess the contribution of individual features, and perhaps allow for the detection of vowel harmony effects.

5.2 Identity effects

The same universals do not appear to apply to identical consonants and vowels. In both consonants and vowels, we see significantly more variation in identity effects than in similarity effects across languages. Identical consonant co-occurrence is dispreferred, but varies greatly across languages. The effect is strongly negative enough that in the O/E plots in Figures 7 and 8, identical consonants occur less often than expected in nearly all languages, but a small number have a zero or positive effect. While languages like Peruvian Aymara (MacEachern 1999) and Arabic (McCarthy 1986) allow identical consonant co-occurrence but prohibit similar consonant co-occurrence, they are fairly uncommon and perhaps have been noted in the literature because of their surprising nature. The majority of languages do not follow this pattern.

The effect of identity on vowel co-occurrence also differs from the effect of similarity. We find a positive vowel identity effect, suggesting that languages tend to repeat identical vowels within words while having no preferences for similar vowels. Like consonant identity effects, vowel identity effects vary enough to allow for languages with a zero or negative effect, like in Spanish and Croatian (Walter 2010). This can be seen in Figures 7 and 8, where most languages have pairs of identical vowels occurring more often than expected, and some less often than expected. In particular, we see languages with vowel harmony processes, like Turkish, with an O/E ratio greater than one for identical vowels even though the O/E ratio is near one for similar vowels. While vowel harmony is typically thought to arise from phonologized coarticulation (Ohala 1994), this suggests that it may have a source in the lexicon, as well. Perhaps identical vowel pairs in the lexicon are more salient to language users, and this is phonologized as a more general rule that identical *features* must co-occur, resulting in vowel harmony.

5.3 Relationships between vowel and consonant effects

We find that consonant co-occurrence is more strongly restricted than vowel co-occurrence across languages. While vowels can co-occur relatively freely, there are much stronger constraints on similar and identical consonant co-occurrence. Nespor et al. (2003) hypothesize that there is a division of labor between vowels and consonants: Consonants are responsible for making distinctions between lexical items, while vowels signal rhythmic class and syntactic structure.

Our results appear to go against this hypothesis. While restricting similar and identical consonant co-occurrence promotes distinctness in consonants within a word, the total coding space available for the lexicon is reduced. Any phonotactic constraint reduces the number of possible words available to a language, and thus the number of possible lexical items that can be distinguished. Vowel co-occurrence, on the other hand, is relatively unrestricted. While Nespor et al. (2003) claim that vowels tend to neutralize (i.e. become more similar) within words, we find little evidence of this. In our analysis, most languages have a tendency towards identical vowel co-occurrence, but there are plenty that have the opposite tendency. Nespor et al.'s (2003) claim that consonants are crosslinguistically more numerous than vowels which makes them more informative is still consistent with our results. However, their claim that the specialized role of consonants goes beyond their numerical superiority is not.

Our finding that consonant identity effects are larger than vowel identity effects may also help explain why identical consonant effects have been found more frequently than identical vowel effects. Consonant identity effects have been found in many languages through examination of OCP effects (McCarthy 1986; Coetzee & Pater 2008; Pozdniakov & Segerer 2007). Vowel identity effects are rarely mentioned (Walter 2010). This may not be because vowel identity effects are uncommon, but rather because they are smaller and more difficult to detect.

Beyond differences in effect size, we found no further relationships between vowel and consonant effects. All correlations estimated by the Bayesian models were near zero, but with large credible intervals. Bayes factors for these parameters were all approximately 1.0, suggesting that we don't have enough evidence to determine if there is or is not a correlation. If we found Bayes factors significantly smaller than 1.0, we would be able to conclude that there is no correlation, but this is not the case. Instead, we have found that either more data or a different statistical model is needed to make conclusions about the existence of a correlation between vowel and consonant co-occurrence effects.

5.4 Natural class similarity and feature similarity

To determine the robustness of our results, models were fit using two different similarity metrics, each with slightly different properties. The choice of similarity metric appears to have only minor effects on qualitative results. For all main effects, the conclusions drawn from both models are the same. The only difference between the two is found in Bayes factors for language-level random effects in three model parameters. For consonant similarity ($\alpha_{\text{sim}|C}$) and vowel identity ($\alpha_{\text{ident}|V}$), both models estimate a positive standard deviation across languages within families, but this result is only confirmed by large Bayes factors in the **NC** model. In the **Feat** model, the Bayes factors suggest that there is no variance in these effects. Similarly, for consonant identity ($\alpha_{\text{ident}|C}$), a small Bayes factor in the **NC** model suggests no variation across languages within families, but the Bayes factor in the **Feat** model shows uncertainty about the existence of this variance.

Many language families in NorthEuraLex contain a very small number of languages: one or two in many cases. Because of this, our results for these parameters may be due to the size of our dataset, rather than any real difference between the two similarity metrics. The consistency in parameter estimates using both metrics suggests that perhaps one is not "better" than the other when investigating co-occurrence across a large number of languages: both feature similarity and natural class similarity appear to be equally valid.

5.5 Future work

While these results present a clear picture of general trends in co-occurrence effects, they have several limitations which could be addressed in future work. First, the two similarity metrics used treat all phonological features equally. It is likely that certain features contribute more to similarity than others (such as place of articulation features), or that individual features have different effects on co-occurrence restrictions. Modeling the co-occurrence of individual features would allow us to examine whether or not the co-occurrence effects identified are a result of specific feature co-occurrences, or an aggregate measure of similarity.

Furthermore, we only examine co-occurrence of *pairs* of segments. Harmony and OCP processes are known to operate in the domain of an entire word or root. While co-occurrence restrictions that affect an entire word can be identified by examining local co-occurrences, our results are not informative about any long-distance patterns. Modeling pairs across longer distances, or modeling longer sequences of segments, could help account for this possibility.

We also acknowledge that while NorthEuraLex contains a fairly large set of languages, it is a sample of languages from specific geographic areas, and does not represent a truly random sample of the world's languages. In future work, gathering a similar dataset of a larger variety of languages would allow for more reliable results. A larger set of languages could also allow us to estimate the correlation parameters in our model. Unfortunately, this dataset does not currently exist, although recent work on grapheme-to-phoneme transcription suggests that it may be within reach (McCarthy et al. 2023).

Finally, we note that individual languages or families are not statistically independent. There are further dependencies between languages that are not captured by these controls, such as geographic and historical relationships. While controlling for language family provides a simple baseline, some language families are obviously more related than others. Including some measure of geographic or typological distance in the models would better account for dependencies between languages, but we leave this for future work.

5.6 Conclusion

In summary, we have shown that there is something fundamentally different about the way consonants and vowels interact within words. Across languages and families, similar and identical consonant co-occurrence is restricted. In vowels, similarity does not appear to constrain co-occurrence, but there is a preference for identical vowels to co-occur in most languages. These co-occurrence effects are consistent across a large sample of languages and language families, showing the benefit of a large-scale cross-linguistic study to shed new light on the longstanding question of what forces shape the structure of lexicons.

Data availability

All code is available on OSF: https://osf.io/sgu4w/.

Funding information

Amanda Doucette was supported by funding from the Fonds de recherche du Québec - Société et culture (FRQSC). Morgan Sonderegger acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (RGPIN-2023-04873). Timothy J. O'Donnell also gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada and the Canada CIFAR AI Chairs Program. Heather Goad acknowledges the support of the Social Sciences and Humanities Research Council of Canada (435-2022-0770) and the Fonds de recherche du Québec - Société et culture (2021-SE3-283351). This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada.

Acknowledgements

We thank the Montreal Computational & Quantitative Linguistics Lab and the audience at LabPhon 18 for helpful feedback. We also thank Glossa associate editor Björn Köhnlein and two anonymous reviewers for their feedback on an earlier draft.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

All authors: Writing – review and editing, conceptualization, methodology; AD: Writing – original draft, software, visualization; AD and MS: Formal analysis; TO, MS, and HG: Supervision.

References

Anttila, Arto. 2008. Gradient phonotactics and the complexity hypothesis. *Natural Language & Linguistic Theory* 26(4). 695–729. https://doi.org/10.1007/s11049-008-9052-2

Archangeli, Diana & Mielke, Jeff & Pulleyblank, Douglas. 2012. Greater than noise: Frequency effects in Bantu height harmony. In Botma, Bert & Noske, Roland (eds.), *Phonological explorations: Empirical, theoretical and diachronic issues*, 191–222. De Gruyter. https://doi.org/10.1515/9783110295177.191

Archangeli, Diana & Pulleyblank, Douglas. 2007. Harmony. In de Lacy, Paul (ed.), *The Cambridge handbook of phonology* (Cambridge Handbooks in Language and Linguistics), 353–378. Cambridge University Press. https://doi.org/10.1017/CBO9780511486371.016

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1—28. https://doi.org/10.18637/jss.v080.i01. Version 2.20.4

Chomsky, Noam & Halle, Morris. 1968. *The sound pattern of English* (Studies in Language). Harper & Row.

Clements, George N. & Hume, Elizabeth V. 1995. The internal organization of speech sounds. In Goldsmith, John (ed.), *The handbook of phonological theory*, 245–306. Blackwell.

Coetzee, Andries W. & Pater, Joe. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory* 26. 289–337. https://doi.org/10.1007/s11049-008-9039-z

Dellert, Johannes & Daneyko, Thora & Münch, Alla & Ladygina, Alina & Buch, Armin & Clarius, Natalie & Grigorjew, Ilja & Balabel, Mohamed & Boga, Hizniye Isabella & Baysarova, Zalina & Mühlenbernd, Roland & Johannes, Wahle & Jäger, Gerhard. 2020. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* 54(1). 273–301. https://doi.org/10.1007/s10579-019-09480-6. Version 0.9

Dickey, James M. & Lientz, B. P. 1970. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics* 41(1). 214–226. https://doi.org/10.1214/aoms/1177697203

Endress, Ansgar D. & Nespor, Marina & Mehler, Jacques. 2009. Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences* 13(8). 348–353. https://doi.org/https://doi.org/10.1016/j.tics.2009.05.005

Frisch, Stefan A. 1996. *Similarity and frequency in phonology*: Northwestern University dissertation.

Frisch, Stefan A. 2004. Language processing and segmental OCP effects. In Hayes, Bruce & Kirchner, Robert & Steriade, Donca (eds.), *Phonetically based phonology*, chap. 11, 346—371. Cambridge University Press.

Frisch, Stefan A. & Broe, Michael & Pierrehumbert, Janet. 1997. Similarity and phonotactics in Arabic. *Rutgers Optimality Archive* .

Frisch, Stefan A. & Large, Nathan R. & Pisoni, David B. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42(4). 481–496. https://doi.org/https://doi.org/10.1006/jmla.1999.2692

Frisch, Stefan A. & Large, Nathan R. & Zawaydeh, Bushra Adnan & Pisoni, David B. 2001. Emergent phonotactic generalizations in English and Arabic. In Bybee, Joan L. & Hopper, Paul J. (eds.), *Frequency and the emergence of linguistic structure*, 159–180. John Benjamins.

Frisch, Stefan A. & Pierrehumbert, Janet B. & Broe, Michael B. 2004. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22. 179–228. https://doi.org/10.1023/B:NALA.0000005557.78535.3c

Futrell, Richard & Albright, Adam & Graff, Peter & O'Donnell, Timothy J. 2017. A generative model of phonotactics. *Transactions of the Association for Computational Linguistics* 5. 73–86. https://doi.org/10.1162/tacl_a_00047

Gabry, Jonah & Češnovar, Rok & Johnson, Andrew. 2023. *cmdstanr: R interface to 'CmdStan'*. Version 0.7.1.

Gafos, Adamantios. 2021. Consonant harmony, disharmony, memory and time scales. *Society for Computation in Linguistics* 4(1). 188–205. https://doi.org/10.7275/vhf9-qh54

Gordon, Matthew K. 2016. *Phonological typology*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199669004.001.0001

Graff, Peter. 2012. *Communicative efficiency in the lexicon*: Massachusetts Institute of Technology dissertation.
Greenberg, Joseph H. 1950. The patterning of root morphemes in semitic. *WORD* 6(2). 162–181. https://doi.org/10.1080/00437956.1950.11659378
Greenberg, Joseph H. & Jenkins, James J. 1964. Studies in the psychological correlates of the sound system of American English. *WORD* 20(2). 157–177. https://doi.org/10.1080/00437956.1964.11659816
Hammond, Michael. 2004. Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies* 4(2). 1—24.
Hansson, Gunnar Ólafur. 2010. *Consonant harmony: Long-distance interaction in phonology*, vol. 145 (University of California Publications in Linguistics). University of California Press.
Harrison, K. David. 1999. Vowel harmony and disharmony in Tuvan and Tofa. In *Proceedings of the Nanzan GLOW*.
Hayes, Bruce & Wilson, Colin. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440. https://doi.org/10.1162/ling.2008.39.3.379
Heinz, Jeffrey. 2010. Learning long-distance phonotactics. *Linguistic Inquiry* 41(4). 623–661. https://doi.org/10.1162/LING_a_00015
Hualde, José Ignacio. 1991. *Basque phonology* (Theoretical Linguistics). Routledge. https://doi.org/10.4324/9780203168004
Ito, Junko & Mester, Armin. 1998. Markedness and word structure: OCP effects in Japanese. *Ms., University of California, Santa Cruz* .
Jeffreys, Harold. 1961. *Theory of probability*. Oxford University Press. https://doi.org/10.1093/oso/9780198503682.001.0001
Kass, Robert E. & Raftery, Adrian E. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795. https://doi.org/10.1080/01621459.1995.10476572
Kaye, Jonathan. 1989. *Phonology: A cognitive view* (Tutorial Essays in Cognitive Science). Routledge.
Leben, William Ronald. 1973. *Suprasegmental phonology*: Massachusetts Institute of Technology dissertation.
Lewandowski, Daniel & Kurowicka, Dorota & Joe, Harry. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100(9). 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008
Luce, Paul A. & Pisoni, David B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19(1). 1–36.
MacEachern, Margaret R. 1999. *Laryngeal cooccurrence restrictions*. Routledge. https://doi.org/10.4324/9780203823811
Mayer, Thomas & Rohrdantz, Christian & Plank, Frans & Bak, Peter & Butt, Miriam & Keim, Daniel A. 2010. Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. In *Proceedings of the 2010 workshop on NLP and linguistics: Finding the common ground*. 70–78. Association for Computational Linguistics.
McCarthy, Arya D. & Lee, Jackson L. & DeLucia, Alexandra & Bartley, Travis & Agarwal, Milind & Ashby, Lucas F.E. & Del Signore, Luca & Gibson, Cameron & Raff, Reuben & Wu, Winston. 2023. The SIGMORPHON 2022 shared task on cross-lingual and low-resource grapheme-to-phoneme conversion. In Nicolai, Garrett & Chodroff, Eleanor & Mailhot, Frederic & Çöltekin, Çağrı (eds.), *Proceedings of the 20th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology*.

230–238. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.sigmorphon-1.27
McCarthy, John J. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17(2). 207–263.
McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC 2nd edn.
Mortensen, David R. & Littell, Patrick & Bharadwaj, Akash & Goyal, Kartik & Dyer, Chris & Levin, Lori. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*. 3475–3484.
Nespor, Marina & Peña, Marcela & Mehler, Jacques. 2003. On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio* 2(2). 203–230. https://doi.org/10.1418/10879
Nicenboim, Bruno & Vasishth, Shravan. 2016. Statistical methods for linguistic research: Foundational ideas — part II. *Language and Linguistics Compass* 10(11). 591–613. https://doi.org/10.1111/lnc3.12207
Ohala, John J. 1994. Towards a universal, phonetically-based, theory of vowel harmony. In *Proceedings of the 3rd international conference on spoken language processing (ICSLP 1994)*. 491–494. https://doi.org/10.21437/ICSLP.1994-113
Ohala, John J. & Ohala, Manjari. 1986. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In Ohala, John J. & Jaeger, Jeri J. (eds.), *Experimental phonology*, chap. 13, 239–252. Academic Press.
Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots 23(2). 367–381.
Pimentel, Tiago & Roark, Brian & Cotterell, Ryan. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics* 8. 1–18. https://doi.org/10.1162/tacl_a_00296
Pozdniakov, Konstantin & Segerer, Guillaume. 2007. Similar place avoidance: A statistical universal. *Linguistic Typology* 11(2). 307–348. https://doi.org/doi:10.1515/LINGTY.2007.025
R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/. Version 4.3.1.
Ritter, Nancy A. & van der Hulst, Harry (eds.). 2024. *The Oxford handbook of vowel harmony (in press)*. Oxford University Press.
Rose, Sharon & Walker, Rachel. 2011. Harmony systems. In *The handbook of phonological theory*, chap. 8, 240–290. John Wiley & Sons. https://doi.org/10.1002/9781444343069.ch8
Saffran, Jenny R. 2003. Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science* 12(4). 110–114. https://doi.org/10.1111/1467-8721.01243
Sonderegger, Morgan. 2023. *Regression modeling for linguistic data*. MIT Press.
Stan Development Team. 2019. Stan modeling language users guide and reference manual, version 2.29. http://mc-stan.org/.
Trubetzkoy, Nikolai. 1939. *Grundzüge der Phonologie [Principles of phonology]*. Translated by Baltaxe, Christiane A. M. 1969. University of California Press.
Vasishth, Shravan & Nicenboim, Bruno & Beckman, Mary E. & Li, Fangfang & Kong, Eun Jong. 2018. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71. 147–161. https://doi.org/10.1016/j.wocn.2018.07.008
Wagenmakers, Eric-Jan & Lodewyckx, Tom & Kuriyal, Himanshu & Grasman, Raoul. 2010. Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey

method. *Cognitive Psychology* 60(3). 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Walter, Mary Ann. 2010. Harmony versus the OCP: Vowel and consonant cooccurrence in the lexicon. *Laboratory Phonology* 1(2). 395–413. https://doi.org/10.1515/labphon.2010.020

Wilson, Colin & Obdeyn, Marieke. 2009. Simplifying subsidiary theory: Statistical evidence from Arabic, Muna, Shona, and Wargamay. *Ms., Johns Hopkins University* .

Winter, Bodo & Bürkner, Paul-Christian. 2021. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass* 15(11). e12439. https://doi.org/10.1111/lnc3.12439

Appendix A: PanPhon Feature Set

All features are coded as +, −, or 0. The following features are defined for each IPA character:

- syl: syllabic
- son: sonorant
- cons: consonantal
- cont: continuant
- delrel: delayed release
- lat: lateral
- nas: nasal
- strid: strident
- voi: voice
- sg: spread glottis
- cg: constricted glottis
- ant: anterior
- cor: coronal
- distr: distributed
- ab: labial
- hi: high (vowel/consonant, not tone)
- lo: low (vowel/consonant, not tone)
- back: back
- round: round
- velaric: velaric airstream mechanism (click)
- tense: tense
- long: long

Appendix B: *brms* Formula

Listing 1: *brms* model corresponding to Equations 10 and 11.

```
bf(pair_count | subset(isCons) ~
        offset(s1_freq_smooth + s2_freq_smooth) +
        log_c_inv + log_v_inv + sim + ident +
        (ident + sim | p | family) +
        (ident + sim | q | family:language),
    shape ~ 0 + (1 | family) + (1 | family:language),
```

```
    cmc = FALSE,
    family = negbinomial) +
bf(pair_count_v | subset(isVow) ~
        offset(s1_freq_smooth + s2_freq_smooth) +
        log_c_inv + log_v_inv + sim + ident +
        (ident + sim | p | family) +
        (ident + sim | q | family:language),
    shape ~ 0 + (1 | family) + (1 | family:language),
    cmc = FALSE,
    family = negbinomial)
```

Appendix C: Additional correlation estimates, NC model

Parameter 1	Parameter 2	Estimate	95% CredI	p_d	Bayes Factor
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{intercept}\|V}$	0.49	[−0.04, 0.81]	0.97	5.33
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{sim}\|C}$	−0.06	[−0.68, 0.61]	0.58	0.99
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{sim}\|V}$	−0.07	[−0.70, 0.60]	0.59	0.96
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{ident}\|C}$	−0.32	[−0.69, 0.14]	0.92	1.78
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{ident}\|V}$	0.06	[−0.39, 0.49]	0.61	0.63
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{sim}\|C}$	0.12	[−0.55, 0.72]	0.64	0.96
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{sim}\|V}$	−0.16	[−0.75, 0.55]	0.68	1.13
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{ident}\|C}$	−0.09	[−0.48, 0.32]	0.67	0.60
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{ident}\|V}$	−0.24	[−0.60, 0.17]	0.88	1.04
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{intercept}\|V}$	0.84	[0.71, 0.93]	1.00	7.63×10^{15}
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{sim}\|C}$	−0.30	[−0.61, 0.08]	0.94	1.94
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{sim}\|V}$	−0.02	[−0.64, 0.61]	0.53	0.87
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{ident}\|C}$	−0.16	[−0.58, 0.35]	0.76	0.81
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{ident}\|V}$	0.19	[−0.19, 0.54]	0.84	0.78
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{sim}\|C}$	−0.26	[−0.61, 0.12]	0.91	1.23
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{sim}\|V}$	−0.19	[−0.74, 0.52]	0.73	1.17
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{ident}\|C}$	−0.01	[−0.49, 0.46]	0.52	0.60
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{ident}\|V}$	0.26	[−0.13, 0.63]	0.91	1.22

NC Model correlation estimates excluded from Table 6. Each row represents a correlation between random effect parameters across language family (γ parameters), or language within family (α parameters).

Appendix D: Additional correlation estimates, Feat model

Parameter 1	Parameter 2	Estimate	95% CredI	p_d	Bayes Factor
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{intercept}\|V}$	0.34	[−0.36, 0.80]	0.87	1.91
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{sim}\|C}$	−0.30	[−0.82, 0.48]	0.80	1.57
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{sim}\|V}$	0.07	[−0.59, 0.67]	0.59	0.95
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{ident}\|C}$	−0.29	[−0.70, 0.29]	0.87	1.56
$\gamma_{\text{intercept}\|C}$	$\gamma_{\text{ident}\|V}$	0.06	[−0.46, 0.55]	0.59	0.71
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{sim}\|C}$	0.14	[−0.52, 0.72]	0.67	0.98
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{sim}\|V}$	−0.03	[−0.65, 0.64]	0.53	0.97
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{ident}\|C}$	0.00	[−0.52, 0.56]	0.51	0.72
$\gamma_{\text{intercept}\|V}$	$\gamma_{\text{ident}\|V}$	−0.11	[−0.60, 0.52]	0.67	0.87
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{intercept}\|V}$	0.72	[0.39, 0.91]	1.00	1.09×10^2
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{sim}\|C}$	−0.39	[−0.81, 0.39]	0.87	2.37
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{sim}\|V}$	0.12	[−0.55, 0.75]	0.64	0.98
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{ident}\|C}$	−0.25	[−0.67, 0.29]	0.84	1.15
$\alpha_{\text{intercept}\|C}$	$\alpha_{\text{ident}\|V}$	0.22	[−0.31, 0.69]	0.81	1.00
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{sim}\|C}$	−0.11	[−0.65, 0.51]	0.65	0.86
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{sim}\|V}$	−0.19	[−0.78, 0.56]	0.70	1.26
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{ident}\|C}$	0.02	[−0.47, 0.50]	0.53	0.64
$\alpha_{\text{intercept}\|V}$	$\alpha_{\text{ident}\|V}$	0.28	[−0.26, 0.73]	0.87	1.32

Feat Model correlation estimates excluded from Table 7. Each row represents a correlation between random effect parameters across language family (γ parameters), or language within family (α parameters).