Sub-phonemic featural dimensions mediate consonantal co-occurrence biases in a cross-linguistically consistent manner

Bruno Ferenc Šegedin (bruno_ferenc_segedin@brown.edu)
Brown University, Department of Cognitive Psychological Sciences; Porgram in Linguistics

Abstract

Studies of segmental co-occurrence constraints have consistently produced evidence consistent with a cross-linguistic anti-similarity bias for consonants (e.g. Frisch et al. (2004), Pozdniakov & Segerer (2007), Walter (2010), Doucette et al. (2024)). The current study tests whether, in spite of this universal similarity avoidance bias, there are cross-linguistically consistent featural harmony biases the world's lexicons. In particular, we test for the presence of a nasal consonant harmony bias, given the fact that categorical nasal consonant harmony is attested in multiple language families and that nasal harmony is phonetically-motivated. A Bayesian negative binomial model of 91 typologically diverse languages' type frequencies for two-consonant words shows evidence of a weak but reliable cross-linguistic bias in favor of nasal harmony, as well as a comparable bias in favor of voicing harmony. The findings also show patterns consistent with a similarity-avoidance bias, most notably a strong cross-linguistic bias against coronal harmony. Taken together, these findings support the notion that similarity-based co-occurrence constraints may be feature dependent in cross-linguistically consistent ways, and more generally that featural dimensions are relevant for understanding the role of segmental redundancy in lexicons.

Keywords: the mental lexicon; consonant harmony; nasal harmony; communicative biases; corpus analysis; bayesian modeling

Introduction

Segmental redundancy in lexicons

The mapping between lexical forms and their respective meanings is known to be arbitrary. An extensive body of research has suggested that in spite of this fundamental arbitrariness, the organization of lexicons is nevertheless optimized for efficient communication (e.g. Piantadosi et al. (2012), Dautriche, Mahowald, Gibson, & Piantadosi (2017), Dautriche, Mahowald, Gibson, Christophe, & Piantadosi (2017), Mahowald et al. (2018) Trott & Bergen (2020), Trott & Bergen (2020)). These studies attempt infer the communicative pressures that shape lexicons by measuring the extent of redundancy (or predictability) in distributions of segments across lexical items. For example, some have argued that lexicons exhibit a pressure against redundancy and in favor of "dispersion", suggesting that lexical items overlap less in their phonological composition than they would given phonotactic constraints alone- and that languages thus "avoid" phenomena like homophony that introduce redundancy in the lexicon (e.g. (Mahowald et al., 2018; Trott & Bergen, 2020, 2022)). Others have suggested that redundancy and lexical ambiguity are a desirable property of lexicons and come about from a production-based the pressure to keep utterances short to retain a compact and easily-accessible inventory of sounds or sub-lexical structures (e.g. Piantadosi et al. (2012), Mahowald et al. (2018)).

Irrespective of ostensible communicative pressures, all human lexicons are shaped by phonetic and phonotactic constraints that inherently cause redundancy and predictability in the distribution of sounds across lexical forms (e.g. Wilson (2006), Hayes & Wilson (2008)). For example, in English, the sound /p/ never precedes /s/ syllable-initially, making a word like [psit] "grammatically" impermissible. Such phonotactic restrictions enhance the predictability of segmental transitions by reducing the space of possible word forms. The aforementioned studies attempting to characterize the effect of communicative pressures on lexicons typically control for the contribution of phonotactic restrictions by calculating phonotactic baselines and comparing real lexical distributions to these baselines (e.g. Piantadosi et al. (2012), Trott & Bergen (2020)). An assumption underlying this approach is that while phonotactic restrictions constrain the range of licit word forms, they are not directly implicated in communicative optimization of lexicons.

A parallel body of work uses lexical statistics to investigate the very substantive phonetic and phonotactic biases that studies of communicative optimization control for (e.g. (Frisch et al., 2004; Hayes & Wilson, 2008; Albright & Breiss, 2024)). Beyond just categorical phonotactic constraints, much of this work uses the segmental organization of lexicons to make inferences about constraints in segmental co-occurrence patterns that manifest not as categorical rules, but as a statistical trend. These are sometimes referred to as "soft constraints" (Frisch et al. (2004)), and include non-local dependencies between particular subclasses of sounds like segmental harmony (e.g. Stanton (2021)).

Similarity Avoidance for Consonants

Similarity-avoidance, one empirically well-supported cross-linguistic "soft" constraint, can be characterized as the generalization that non-adjacent consonants are less likely to co-occur with other consonants that are phonologically similar. Frisch et al. (2004) finds evidence that in Arabic, the count of roots with similar consonants is smaller than that predicted by a baseline where sounds are allowed to freely co-occur given their position-specific frequencies. This result has been replicated across a wide range of typologically diverse languages (e.g. Walter (2010), Pozdniakov & Segerer (2007), Doucette et al. (2024)). Most recently, in a corpus analysis of 107 Northern Eurasian languages, Doucette et al. (2024) find a qualitative and cross-linguistically consistent difference between consonant and vowel co-occurrence restrictions. Specifically, consonant co-occurrence patterns show evidence of the aforementioned anti-similarity bias in a

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

relatively consistent manner across languages, whereas vowels show no effect of similarity avoidance and also a preference for vowel identity.

The fact that subclasses of sounds mediate co-occurrence patterns in cross-linguistically consistent ways likely reflects the fact that sounds are not merely abstract and equivalent symbolic units, but rather vary in their articulatory and perceptual attributes. Yet, it remains understudied whether lexicons universally rely on particular subclasses of sounds over others to introduce or alleviate redundancy in the lexicon. Some evidence suggests that communicative factors constrain which sounds undergo changes in particular languages. For example, Wedel et al. (2013) finds that languages are less likely to undergo the loss of a contrast for sounds responsible for maintaining many lexical distinctions, and Cohen Priva (2017) finds that voiceless stops with lower average informativity are more likely to undergo phonetic lenition (weakening) than voiceless stops that happen to be informative in that language. While such findings demonstrate that language-specific distributions of particular sound classes determine their proclivity to change, it is under-explored whether certain subclasses of sounds are uniquely likely to exhibit particular distributional properties to begin with.

Consonant Harmony

The current study aims to make headway on this front by testing whether, in spite of the well-supported anti-similarity bias for consonants, there is a universal preference in favor of similarity along particular featural dimensions. The existence of biases in favor of similarity at the featural level are consistent with the existence of consonantal harmony systems that force consonants within a word to align along featural dimensions (e.g. Rose & Walker (2011)). Consonantal harmony systems occur in a wide range of typological contexts and involve a wide range of features. The most common kind of consonant harmony is sibilant harmony, for which a pair of sibilants in the same word must match in place of articulation (e.g. Berkson (2013)). The focus of the current study is nasal consonant harmony. Nasal sounds (like /m/ and /n/ in English) can most simply be described as those for which the velum is lowered during articulation, allowing airflow through the nasal cavity (e.g. Kurowski & Blumstein (1987)). In nasal consonant harmony, all consonants in a given domain like a word are forced to align in their nasality specification, regardless of the specification of intervening vowels (Walker (2011); see Rose & Walker (2011) for a more exhaustive overview of vowel and consonant harmony systems). An example of a categorical and morphologically-productive nasal consonant harmony system can be found in the Bantu language Yaka, where the suffix [*-ini*] is attached to a stem whose last consonant is nasal, and the corresponding oral suffix [*-idi*] is attached when the last consonant of the stem is oral (Hyman (1995)). We choose to focus on nasal harmony because, unlike sibilant harmony which only affects a narrow class of sibilant sounds, nasality is a suprasegmental feature that can co-occur with a wide variety of place and manner features. As such, a broad range of consonants can in principle participate in nasal harmony, allowing its effects to be observed across a wider span of lexical items. In addition, a potential nasal harmony bias is well-motivated both typologically and phonetically. Typologically, while most common in Bantu languages like Yaka, nasal consonant harmony also occurs in other language families, such as Mixtec and Tupi languages (e.g. Piggott (1992)). Beyond nasal harmony systems that strictly affect consonants at a distance, local nasal harmony systems between consonants and vowels are also widely attested (e.g. Walker (2011)). Phonetically, nasal coarticulation is a well-studied phenomenon in which the articulatory gesture of nasality, and its acoustic effects, spill over into adjacent and underlyingly non-nasal segments (e.g. Zellou & Tamminga (2014), Kurowski & Blumstein (1987)). Thus, because nasal harmony is common and phonetically-motivated, we predict that there should be a cross-linguistic bias for nasal harmony in consonantal co-occurrence patterns. If such a bias holds across languages, it would constitute one potential exception to the well-supported generalization that consonant co-occurrence patterns are subject to an anti-similarity bias (e.g. Pozdniakov & Segerer (2007), Doucette et al. (2024)).

Study

The current study tests whether there is evidence for a universal nasal consonant harmony bias. Given the properties of nasality described above, we predict a cross-linguistic bias in favor of nasal consonant harmony- in spite of a well-documented anti-similarity bias for consonant co-occurrence. Alternatively, it may be the case that languages also exhibit a dissimilarity preference along the dimension of nasality, which would be in keeping with with the notion that the general anti-similarity bias is not systematically sensitive to consonants' featural dimensions. It may also be the case that the effect of nasality on co-occurrence restrictions is not consistent across languages, and thus that whatever properties differentiate nasal consonants from other segments do not affect consonant co-occurrence in a way that is robust to language-specific factors.

Data

To test for the presence of a universal nasal harmony bias, we use phonologically-transcribed data from 91 languages'[1] con-

[1]The 91 languages included in our study are the following: Akawaio, Apalaí, Wayana (Carib); Albanian, Aragonese, Armenian, Asturian, Belarusan, Bulgarian, Czech, Greek, Macedonian, Nepali, Romanian, Russian Buriat, Slovak, Spanish, Tajik, Ukrainian, Upper Sorbian, Yiddish (Indo-European); Amanab, Amele, Angor, Ankave, Bargam, Benabena, Borong, Daga, Guhu-Samane, Iduna, Komba, Kunimaipa, Mauwake, Mountain Koiali, Nabak, Rawa, Somba-Siawari, South Tairora (Trans-New Guinea); Bugis, Kiribati, Malagasy, Mamasa, Manam, Mapos Buang, Muna, Nehan, Samoan, Sinaugoro, Tongan (Austronesian); Azerbaijani, Bashkir, Crimean Tatar, Tatar, Turkish, Uyghur, Uzbek (Turkic); Kannada, Malayalam, Telugu (Dravidian); Erzya, Hungarian, Mari (Uralic); Jola-Fogny, Kagulu, Namaande, Nigerian Fulfulde, Wolof (Niger-Congo); Apurinã, Asháninka, Yine (Arawakan); Huarijío, Tarahumara (Uto-Aztecan); Francisco León Zoque (Mixe-Zoque); Georgian (Kartvelian); Inuktitut (Eskimo-Aleut); Shilha (Afro-

sonant co-occurrence patterns. Languages are selected from a pool of 201 languages in the XPF corpus (Cohen Priva et al. (2021)), which consists of phonologically-transcribed word lists derived from language-specific orthography-to-phonetic transcription rules. The wordlists are originally from from the Crúbadán (Scannell[1] (2007)), OpenSubtitles (Tiedemann (2016)), and LDC (Bills et al. (2016)) corpora. From the original pool of 201 languages, we omit languages that fit any of the following exclusion criteria: (1) more than 2% of untranslated words, (2) fewer than 2500 translatable tokens, (3) 200 or fewer disyllabic words. We restrict our analysis to words with only two consonants where the consonants are separated by at least one vowel. We label the segments in these wordlists using segment-to-feature correspondences from the PHOIBLE corpus (Moran & McCloy (2019)).

Modeling Approach

Following prior work investigating segmental co-occurrence patterns in lexicons, we test whether counts of particular consonant sequences are over- or under-attested relative to what the count of these sequences would be if sounds were allowed to combine freely. Our approach follows the general logic of prior studies that compute observed/expected (OE) measures (e.g. Frisch et al. (2004), Walter (2010), Stanton (2021)), except that we employ log-linear modeling which enables the measurement of the over- or under-attestation of certain sequences while also controlling for other predictors in the baseline (Wilson & Obdeyn (2009), Breiss & Albright (2022), Albright & Breiss (2024), Doucette et al. (2024)). We use a Bayesian negative binomial model to predict type-frequency of consonant pairs based on a handful of phonological predictors predictors (see equation 1). A negative binomial model is a version of a Poisson model (used for modeling co-occurrence constraints in Albright & Breiss (2024)) except that it also estimates a free parameter ϕ for the dispersion of the data, and thus does not assume that the distributions of counts narrowly follow a Poisson distribution. The model formula is presented in equation (1), and each predictor is explained below.

$$\begin{aligned} \text{count} \sim\ & 1 + \text{nasal} + \text{coronal} \\ & + \text{sonorant} + \text{voicing} \\ & + \text{identity} + \text{log.c1.count} + \text{log.c2.count} \\ & + (1 + \text{nasal} + \text{coronal} + \text{sonorant} \\ & \quad + \text{voicing} + \text{identity} \mid\mid \text{lang}) \end{aligned} \tag{1}$$

Nasal harmony is captured via a binary predictor (*'nasal'*), whose value is 1 if both consonants in a sequence align in their nasality specification- if they are both nasal or both non-nasal, and 0 otherwise. To assess the robustness of a potential nasality bias, we include two predictors for featural dimensions that may be confounded with nasality: voicing (*'voice'*) and sonority (*'sonorant'*). We also include a predictor for a feature that is likely to capture an independent source of variation to nasality: coronal harmony (*'coronal'*). Coronals are the class of sounds like /t/, /s/ and /n/, for which the tongue top is the articulator. Like '*nasal*', all binary predictors have a label of 1 for a sequence if the specifications for for the consonants' feature are either both positive or both negative, and 0 if one consonant has a positive value for a feature and the other a negative feature. Lastly, we include a predictor for segmental identity, whose value is 1 for sequences of identical consonants and 0 for sequences of nonidentical consonants. For each binary predictor, we also include random slopes across the 91 languages to account for language-specific variability in the effects of each predictor on counts of consonantal sequences. To control for position-specific frequencies of each sound within each language, we use variables *'log.c1.count*, and *'log.c2.count*, which are the log of the proportions of each sound within each position for a particular language. A positive coefficient estimated by the negative binomial model on any of the binary predictors is equivalent to an OE ratio above 1 reported in prior work (e.g. Frisch et al. (2004)); it means that that configuration of sounds is overrepresented relative to what would be expected at chance, and when exponentiated, the estimate can more directly be compared to an OE value. From an information-theoretic standpoint, a positive effect would mean that a particular specification of nasality on c1 predicts the same specification on c2 and vice versa, while a negative value would mean that a particular specification on c1 predicts the opposite specification on c2.

We use weakly informative priors across all effects except for both of the position-specific frequency predictors *log.c1.count* and *log.c2.count*. For both of these, we use an informative prior $\beta \sim \mathit{Normal}(1,3)$. This reflects the prior expectation that the position specific frequency of c1 or c2 should be positively correlated with the frequency of the pair c1_c2. For all binary fixed effects, we use a normal distribution centered at 0 with a standard deviation of 3 for all fixed effects: $\beta \sim \mathit{Normal}(0,3)$. This prior reflects no prior expectation about the direction of any of the possible co-occurrence constraints. For the standard deviations of the random effects (*language* and *family*), we used an exponential distribution with a rate parameter of 0.5: $\sigma \sim \mathit{Exponential}(0.5)$. This is a conservative prior that allows for variability in random effects only if strongly supported by the data. For the dispersion parameter ϕ, we used a log-normal distribution with a mean of 0 and a standard deviation of 0.5: $\phi \sim \mathit{Log\text{-}Normal}(0,0.5)$.

Results

Fig. 1 shows the posterior distributions for the main effects of the five binary predictors, and table 1 shows the point estimates (posterior means), and credible intervals for all binary predictors, as well as for the intercept and the position specific frequencies. The data support our prediction of a nasal harmony bias: there is a evidence of a weak but reliable bias in favor of nasal harmony, as well as a similar bias in fa-

Asiatic); Aymara (Aymaran); Warlpiri (Australian); Bora (Witotoan); Chayahuita (Cahuapanan); Colorado (Barbacoan); Bislama, Chavacano (Creole); Basque, Candoshi-Shapra, Korean (Language Isolates).

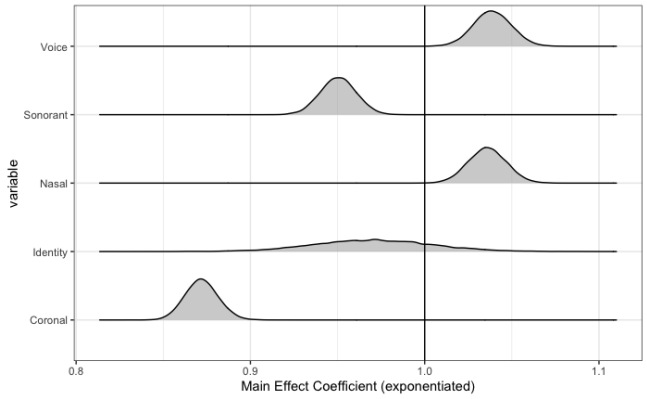

Figure 1: Exponentiated posterior distributions of estimates for each binary predictor. The under-representation of words with coronal harmony is the most extreme effect.

vor of voicing harmony. There is also evidence for a bias against sonorant harmony. Notably, there is a strong effect *against* harmony for the coronal feature: its posterior mean is $\exp(-0.14) = 0.85$, which means that consonant sequences with coronal harmony are on average 85% as frequent as the average frequency of a consonant sequence. Lastly, while the posterior distribution of identity estimates skews negatively, there is a lack of robust evidence of a bias against identity, reflected in the fact that the 95% credible interval extends as high as 0.04. The estimates for the random effects (table 2) reveal a considerable degree of between-language variability for the main effect of identity, while the featural predictors show little variability across languages, and have 0.00 within their 95% credible intervals. This suggests that estimates for the featural predictors are cross-linguistically stable and not subject to language-specific factors, at least not to the degree that they are for identity.[2]

The large degree of between-language variability for identity compared to the featural predictors is consistent with the findings of Doucette et al. (2024). Taken together, these patterns support the notion that the degree of co-occurrence constraints is mediated by featural dimension, and that the effects of particular featural dimensions are largely stable across languages.

Discussion

The results of the current corpus analysis are consistent with prior work that finds evidence of a universal an anti-identity co-occurrence restriction for consonants (Doucette et al. (2024)). Most notably, the strongest effect was a preference against coronal harmony. While we did not find a reliable anti-identity bias as might be predicted by a similarity-avoidance bias, our results are consistent with the large degree of between language variability for identity biases reported by Doucette et al. (2024). Our results are also consistent with our prediction of a universal nasal harmony bias, even though we observe a weak effect. To our knowledge, this is the first study to report empirical evidence of a cross-linguistic bias in favor of nasal consonant harmony. We also find evidence of an equally strong bias in favor of voicing harmony, which we did not expect a priori. These results put into context prior findings of an anti-similarity bias for consonants, by demonstrating that such a bias is likely to be more sensitive to overlap along certain feature dimensions than others.

	Estimate	Est. Error	CI (95%)
Intercept	8.06	0.10	[7.88, 8.26]
identity	-0.03	0.04	[-0.10, 0.04]
nasal	0.04	0.01	[0.01, 0.06]
sonorant	-0.05	0.01	[-0.07, -0.03]
voice	0.04	0.01	[0.02, 0.06]
coronal	-0.14	0.01	[-0.16, -0.12]
log.c1.count	0.65	0.01	[0.63, 0.66]
log.c2.count	0.65	0.01	[0.64, 0.67]

Table 1: Summary of parameter estimates for main effects.

	Estimate	Est.Error	CI (95%)
sd(Intercept)	0.75	0.06	[0.65, 0.87]
sd(identity)	0.25	0.03	[0.20, 0.32]
sd(sonorant)	0.02	0.01	[0.00, 0.05]
sd(voice)	0.02	0.01	[0.00, 0.05]
sd(nasal)	0.03	0.02	[0.00, 0.06]
sd(coronal)	0.04	0.02	[0.01, 0.07]

Table 2: By-language random effects summary: estimates for the standard deviation for variability of each predictor across languages. Identity has higher between-language variability than the featural harmony predictors.

Why harmony biases for some features and anti-harmony for others?

What are the articulatory and perceptual properties that might make featural dimensions like voicing and nasality particularly resistant to anti-similarity biases? One possibility is that the articulatory structure of nasality lends itself to manifesting as a word-level or suprasegmental property. Namely, nasality is achieved by lowering the velum to allow airflow through the nasal cavity, which can be carried out independently of other articulatory gestures that determine features like place of articulation (e.g. Kurowski & Blumstein (1987)). While we did not predict a bias in favor of voic-

[2]To assess robustness, we repeated the analysis using a less conservative prior on by-language random effects, Student-$t(\nu = 3,\ \mu = 0,\ \sigma = 3)$, to test whether the original prior suppressed between-language variation. We also excluded 10 languages due to suspected transcription issues: Akawaio, Mapudungun, Mari, Asháninka, Malayalam, Bargam, Erzya, Yine, Rawa, and Slovak. All estimates for binary predictors were identical except for minimal shifts in a few effects or credible intervals: between-language variability for identity (**0.26**; CI = [**0.21**, **0.33**]) and coronal (0.04; CI = [0.02, **0.08**]), and main effects for sonorant (**–0.06**; CI = [**–0.08**, –0.03]) and coronal (–0.14; CI = [–0.16, **–0.11**]).

ing harmony, the results are consistent with the existence of categorical laryngeal harmony systems (e.g. Hansson (2001), Rose & Walker (2011)). Additionally, like nasality, voicing is a feature that can change independently of precise articulatory targets (many voiceless sounds have voiced counterparts) and it may thus be articulatorily expedient to maintain the same laryngeal setting for all consonants across a word.

The strong bias against coronal harmony is broadly consistent with prior evidence for anti-similarity biases, and in particular proposals that place articulation at the center of these biases. For example, Pozdniakov & Segerer (2007) suggests that avoidance of similar place of articulation is a primary driver of anti-similarity biases across languages. One possibility is that coronal harmony is itself uniquely dis-preferred, independently of place of articulation. For example, just as it may be articulatorily taxing to reuse the same place of articulation in quick succession it may be articulatorily undesirable to repeatedly use a particular part of the tongue -independent of its articulatory target. Conversely, it may be the case that because coronal gestures are confined to places of articulation near the front of the mouth, our predictor for coronal harmony is confounded with a predictor that more specifically isolates place of articulation. In any case, the difference between coronal harmony from nasality and voicing is consistent with the notion that consonantal co-occurrence is restricted more along articulatory dimensions that involve specific tongue gestures, than those that involve nasal or laryngeal settings that are articulatorily orthogonal to tongue movements. However, why sonorant harmony is also weakly avoided cross-linguistically is less clear. We leave disentangling the precise dimensions along which anti-similarity biases are strongest to future work. The current results nevertheless suffice to show that featural dimensions mediate co-occurrence constraints and that they do so in cross-linguistically consistent ways.

Implications for communicative explanations of lexical organization

In studies using lexicons and their phonological compositions as a source of data, is often a (well-justified) matter of convenience to consider words as as strings of segments bearing equivalent status, and to relegate phonotactic constraints and phonologically-informed differentiation of sounds to phonotactic baselines (e.g. Trott & Bergen (2020)), or to ignore them outright. Measures of neighborhood density used in behavioral studies, for example, have proven to be sufficient for predicting behavioral patterns like reaction time, even if they treat differences between all segments as equivalent (Gahl & Strand (2016)). The empirical findings of the current study underscore the possibility that different kinds of sounds carry out qualitatively different information-theoretic roles. More generally, we suggest that languages' phonotactic patterns should be of interest to researchers focused on quantifying the effect of communicative pressures on segmental redundancy the lexicon, and it should not be assumed that the effect of phonological factors on the lexicon precede any process of communicative optimization (e.g. Trott & Bergen (2020); Piantadosi et al. (2012)).

Future work should ultimately explore more exhaustively whether there are cross-linguistically consistent subclasses of sounds that fulfill particular communicative desiderata in lexicons. Trade-offs between dispersion and redundancy may be mapped onto particular articulatory dimensions; for example, similarity preferences along certain articulatory dimensions could plausibly either offset or further exacerbate the information loss caused by a anti-similarity biases for other features. Thus, future work should quantify not just the magnitude of co-occurrence restrictions but also their effect on the communicative capacity of the lexicon.

Conclusion

This corpus study of 91 languages' lexicons is the first, to our knowledge, to report evidence of a cross-linguistic nasal consonant harmony bias. We speculate that the differing contrasting co-occurrence biases between gesture-based featural dimensions like +/-coronal on the one hand, and non-gesture based features like nasality and voicing on the other, make the latter more likely to be implicated in harmony biases, or resistant to the effects of a similarity-avoidance bias. The cross-linguistic consistency with which featural dimensions seem to mediate co-occurrence patterns, compared to identity, suggests that considering sub-phonemic information is relevant for building and testing theories about the effect of communicative factors on segmental redundancy in the lexicon.

Acknowledgments

I thank Uriel Cohen Priva for invaluable discussions and feedback.

References

Albright, A., & Breiss, C. (2024). A poisson model of phonological cooccurrence restrictions.

Berkson, K. H. (2013). Optionality and locality: Evidence from navajo sibilant harmony. *Laboratory Phonology*, *4*(2), 287–337.

Bills, A., David, A., Dubinski, E., Fiscus, J., Hammond, S., Gann, K., et al. (2016). Iarpa babel georgian language pack iarpa-babel404b-v1. 0a. *Web Download. Philadelphia: Linguistic Data Consortium*.

Breiss, C., & Albright, A. (2022). Cumulative markedness effects and (non-) linearity in phonotactics.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, 569–597.

Cohen Priva, U., Strand, E., Yang, S., Mizgerd, W., Creighton, A., Bai, J., ... Wiepert, D. (2021). The cross-linguistic phonological frequencies (xpf) corpus.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science*, *41*(8), 2149–2169.

Doucette, A., O'Donnell, T. J., Sonderegger, M., & Goad, H. (2024). Investigating the universality of consonant and vowel co-occurrence restrictions. *Glossa: a journal of general linguistics*, *9*(1). doi: 10.16995/glossa.9373

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the ocp. *Natural language & linguistic theory*, *22*(1), 179–228.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178.

Hansson, G. (2001). The phonologization of production constraints: Evidence from consonant harmony. In *Chicago linguistic society* (Vol. 37, p. 187).

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

Hyman, L. M. (1995). Nasal consonant harmony at a distance the case of yaka. *Studies in African Linguistics*, *24*(1), 6–30.

Kurowski, K., & Blumstein, S. E. (1987). Acoustic properties for place of articulation in nasal consonants. *The Journal of the Acoustical Society of America*, *81*(6), 1917–1927.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive science*, *42*(8), 3116–3134.

Moran, S., & McCloy, D. (2019). *Phoible 2.0. jena: Max planck institute for the science of human history. retrieved april 28, 2023.*

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Piggott, G. L. (1992). Variability in feature dependency: the case of nasality. *Natural Language & Linguistic Theory*, *10*(1), 33–77.

Pozdniakov, K., & Segerer, G. (2007). Similar place avoidance: A statistical universal.

Rose, S., & Walker, R. (2011). Harmony systems. *The handbook of phonological theory*, 240–290.

Scannell[1], K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *Building and exploring web corpora (wac3-2007): Proceedings of the 3rd web as corpus workshop, incorporating cleaneval* (Vol. 4, p. 5).

Stanton, J. (2021). An identity preference in ngbaka vowels. In *Proceedings of the annual meetings on phonology.*

Tiedemann, J. (2016). Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 3518–3522).

Trott, S., & Bergen, B. (2020). Why do human languages have homophones? *Cognition*, *205*, 104449.

Trott, S., & Bergen, B. (2022). Languages are efficient, but for whom? *Cognition*, *225*, 105094.

Walker, R. (2011). Nasal harmony. *The Blackwell companion to phonology*, 1–28.

Walter, M. A. (2010). Harmony versus the ocp: Vowel and consonant cooccurrence in the lexicon. *Laboratory Phonology*, *1*(2), 395–413.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *128*(2), 179–186.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, *30*(5), 945–982.

Wilson, C., & Obdeyn, M. (2009). Simplifying subsidiary theory: statistical evidence from arabic, muna, shona, and wargamay. *Ms, Johns Hopkins University*.

Zellou, G., & Tamminga, M. (2014). Nasal coarticulation changes over time in philadelphia english. *Journal of Phonetics*, *47*, 18–35.