Can a pressure against homophones explain phonological neighborhoods?

Sean Trott (sttrott@ucsd.edu)
Department of Cognitive Science
University of California, San Diego

Benjamin Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science
University of California, San Diego

Abstract

Words in human languages cluster together in phonological neighborhoods more closely than would be expected by chance. But why? One explanation is that large neighborhoods are directly selected for, possibly because they scaffold word learning and production. But it's also possible that they emerge as a byproduct of other constraints or selection pressures operating over real lexica. We advance one such selection pressure as a candidate explanation. A pressure to avoid overloading unique wordforms with homophones may lead to clusters of words that are not identical but similar. Using simulated baselines, we test the viability of this alternative account. We find that a pressure against loading too many meanings on unique wordforms—paired with the phonotactics of a target language—produces lexica with neighborhoods that are at least as large on average as those in real lexica. This does not rule out the possibility of a pro-neighborhood pressure, but it does demonstrate the viability of a parsimonious alternative account based on a pressure against homonymy for which there is independent evidence.

Keywords: phonological neighborhoods; ambiguity; language evolution; lexicon

Introduction

Why are human languages structured the way that they are? One approach to finding evolutionary causes for contemporary structure seeks to characterize the various *selection pressures* that could plausibly account for the form and content of languages (Richie, 2016). This approach has produced a growing consensus that human lexica are shaped by a pressure for cognitive and communicative efficiency (Gibson et al., 2019; Levshina & Moran, 2021), both in terms of how they carve up semantic domains (e.g., color) (Regier, Kay, & Khetarpal, 2007; Kemp & Regier, 2012; Zaslavsky, Kemp, Regier, & Tishby, 2018; Kemp, Xu, & Regier, 2018), and in the wordforms they contain (Piantadosi, Tily, & Gibson, 2011; Mahowald, Dautriche, Gibson, & Piantadosi, 2018).

But one feature of language that has to date resisted explanation in these terms is the presence of dense *phonological neighborhoods*. Lexica are *clumpy*: they contain dense pockets of wordforms differing in only one sound (e.g., "cat", "bat", and "mat")—typically called *phonological neighbors*—while leaving vast swaths of phonological space entirely unused (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017). From the perspective of communicative efficiency, this clumpiness may be surprising; allowing wordforms to cluster in particular regions of phonological space——instead of distributing them more evenly——has been found to increase the likelihood of misperceiving one wordform for another, potentially even impairing comprehensibility (Vitevitch & Luce, 1998).[1] One explanation for the prevalence of neighborhoods comes from *phonotactics*. Each language has certain rules about which sounds can start and end a word, which sounds can occur in which sequence, and so on (Frisch, Large, & Pisoni, 2000; Bailey & Hahn, 2001; Vitevitch & Aljasser, 2021). Phonotactic rules sharply constrain the space of legal words in a language, simplifying the speaker's task of selecting and producing words. And phonotactics may also account for some of the clumpiness oberved in human languages. However, recent work (Dautriche et al., 2017) has found that phonotactics alone cannot fully account for the neighborhood density of real human lexica: across four languages (English, Dutch, German, and French), phonological neighborhoods are still larger than one would expect in a lexicon whose wordforms were determined purely by the phonotactics of that language (Dautriche et al., 2017). What accounts for this gap?

A natural explanation is that dense phonological neighborhoods are directly *selected for*, i.e., they increase cognitive or communicative efficiency in some way. Indeed, there is some evidence that dense neighborhoods may facilitate both word learning (Storkel, 2004; Storkel, Armbrüster, & Hogan, 2006; Coady & Aslin, 2003; Jones & Brandt, 2020; Fourtassi, Bian, & Frank, 2020; Jones & Brandt, 2019) and word production (Vitevitch, 2002; Vitevitch & Sommers, 2003). If this interpretation is correct, it suggests that the possible benefits of dense neighborhoods (facilitation of word learning and production) "outweigh" the challenges they may pose for comprehension (Vitevitch & Luce, 1998). Thus, under this view, neighborhoods are the result of a positive selection pressure—above and beyond the phonotactics of a language.

Another possibility, however, is that dense neighborhoods are the *byproduct* of other properties or selection processes that operate over real human lexica. The fact that neighborhoods appear to confer a benefit on lexical acquisition and production does not entail that they were selected for this function; there are numerous examples in evolutionary biology of apparently adaptive traits that emerged at least partially as a byproduct of other selection pressures (Gould & Lewontin, 1979). Below, we introduce one such candidate

[1] Note that contradictory results have been obtained in Russian and Spanish, in which dense neighborhoods may actually facilitate word perception (Vitevitch & Rodríguez, 2005; Arutiunian & Lopukhina, 2020).

pressure—a selection pressure against homophony—and describe how it could result in lexica with dense phonological neighborhoods, even without a direct selection pressure for clumpiness.

Real Lexica Select Against Over-Saturation There has been a good deal of attention recently on why ostensibly efficient communication systems would evolve to contain homophony, i.e., wordforms with distinct, unrelated meanings (Piantadosi, Tily, & Gibson, 2012). Several papers (Trott & Bergen, 2020; Caplan, Kodner, & Yang, 2020) have adopted the approach of building *phonotactic baselines* (Dautriche et al., 2017) to ask how much homophony one should expect to find purely as a function of a language's phonotactics. That is, if wordforms were randomly sampled (with replacement) from phonotactic space, how frequently would different meanings be assigned to the same wordform?

These phonotactic baselines have been able to account for both the amount and distribution of homophony. But surprisingly, real human lexica actually have *fewer* homophones per wordform than their artificial, phonotactic counterparts (Trott & Bergen, 2020), and this homophony is more evenly distributed across the lexicon, i.e., across longer and more ill-formed wordforms, than one would expect (Trott & Bergen, 2020; Caplan et al., 2020).

A natural explanation for the gap in homophony is that real lexica are subject to a pressure against saturating the same wordform with too many meanings. A few notes of clarification are required here. First, all spoken languages appear to display homophony, so any hypothesized pressure against homophony must not be categorical (Sampson, 2013). Second, what all languages studied to date share is an apparent resistance to *over-saturation*, i.e., the number of meanings loaded onto the same wordform, relative to what would be expected from a phonotactic baseline. This is despite the fact that some languages (English and Japanese in particular) have a higher *rate* of homophony (i.e., more wordforms with at least one meaning) than baselines (Trott & Bergen, 2020).[2] Taken together, these facts suggest that real lexica may be subject to a *smoothing* process: rather than concentrating many meanings in the highest-probability wordforms—which could impede communication—real lexica may distribute these meanings more evenly across phonotactic space (Trott & Bergen, 2020), which could result in larger neighborhoods.

Could smoothing create larger neighborhoods? If real lexica prefer wordforms with high phonotactic probability, as they appear to, and if at the same time they also select against over-saturating the same high-probability wordform, then they should be more likely to instead select other high-probability (but not overly homophonous) wordforms in adjacent phonological space. Under this account, the distribution of wordforms across phonological space would be determined by two primary factors:

1. A pressure to use well-formed phonological sequences, i.e., those with high phonotactic probability.

2. A pressure against over-saturating the same wordform with an excess of meanings.

Critically, this pair of pressures together could result in larger phonological neighborhoods than either of them would independently, even while not directly selecting for dense neighborhoods. Instead of sampling the same high-probability wordform (e.g., "gap") many times, this process would sample from similarly high-probability regions of phonotactic space, which—simply because of the previously established connection between phonotactic probability and neighborhood density (Dautriche et al., 2017)—would select wordforms that are more likely than chance to be neighbors of existing words. In the aggregate, this would indirectly produce denser neighborhoods.

This explanation—a pressure against oversaturation of individual wordforms increases neighborhood density—has several things to recommend it a priori. First, the pressure against oversaturation is itself independently motivated, as described above. But second, it could also account for a *dissocation* between homonymy and neighborhood size observed in past work (Dautriche et al., 2017; Trott & Bergen, 2020). Across five languages tested (English, Dutch, German, French, and Japanese) by two groups, real human lexica consistently have larger neighborhoods but fewer homonyms than their phonotactic baselines. Finding a single explanation for both effects is desirable from the perspective of theoretical parsimony; rather than positing multiple, distinct pressures to explain different results—a pressure *against* homophony (Trott & Bergen, 2020) and a pressure *for* denser neighborhoods (Dautriche et al., 2017)—a single pressure could in principle explain two apparently unrelated phenomena, i.e., "filling two needs with one deed".[3]

Under this alternative account, dense neighborhoods may still provide benefits to word learning and production (Storkel, 2004; Storkel et al., 2006; Vitevitch, 2002). However, these advantages would not be causally responsible for larger neighborhoods, but rather, would be a kind of "positive externality" created by a selection pressure against homophones.

Current Work

The central goal of the current work was to ask whether a pressure against homophony—coupled with phonotactic constraints—could explain the distribution of neighborhood sizes observed in real lexica. To our knowledge, this account has not been directly tested.

We followed the approach taken in past work (Dautriche et al., 2017; Trott & Bergen, 2020; Caplan et al., 2020);

[2] This may also partially account for the mixed results reported in more recent work (Pimentel, Meister, Teufel, & Cotterell, 2021).

[3] In principle, a pressure for larger neighborhoods may also explain why real lexica have fewer homophones. This issue is explored in the General Discussion.

for each language of interest, we simulated a series of *baselines*, matched for the phonotactics and distribution of word lengths (as defined by number of syllables) of the target lexicon. Unlike past work, however, we also introduced novel constraints for some of these baselines. Specifically, we introduced an Anti-Homophone pressure, which prevented a wordform from acquiring too many meanings and forced the baselines to conform to the rank distribution of homophones found in the real lexicon.[4]

We then compared two measures of neighborhood size (Mean and Maximum Neighborhood Size) across the real lexica and their baselines. Our question was to what extent these constraints—phonotactics, and a pressure against homophony—were *sufficient* to account for neighborhood density in real lexica. Critically, a demonstration of sufficiency would not *disconfirm* the possibility that real lexica are subject to a pro-neighborhood pressure. Rather, it would serve as a proof-of-concept that there are alternative (and possibly more parsimonious) routes that could account for the size of neighborhoods in real lexica.

All materials and code are available on GitHub: `https://github.com/seantrott/neighbors_lexica`.

Methods

Materials

We analyzed five languages: English, Dutch, German, French, and Mandarin. To do this, we relied on lexical resources that contained phonological information for each *lemma* of a lexicon. We used CELEX (Baayen, Piepenbrock, & Gulikers, 1996) for English, Dutch, and German; Lexique (New, Pallier, Brysbaert, & Ferrand, 2004) for French; and the Chinese Lexical Database for Mandarin (Sun, Hendrix, Ma, & Baayen, 2018).

To ensure that our analyses were consistent with previous work (Trott & Bergen, 2020; Piantadosi et al., 2012), we restricted our analysis to lemmas. We also removed wordforms containing hyphens, spaces, or apostrophes, as well as proper nouns. The final number of lexical entries (i.e., lemmas) for each real lexicon was: 41887 entries in English, 67583 entries in Dutch, 51718 entries in German, 43782 in French, and 45552 in Mandarin.

Building Phonotactic Models

To model the phonotactic rules of each language, we fit a series of *n*-phone Markov Models to each lexicon (Dautriche et al., 2017; Trott & Bergen, 2020; Caplan et al., 2020). By observing the entire set of wordforms in a language, an *n*-phone model can learn statistical contingencies such as which phonemes are most likely to start and end a wordform, and which phonemes are most likely to follow the previous *n - 1* phonemes.

[4]We did not attempt to model the specific cognitive or diachronic mechanisms by which homophony avoidance might come about, e.g., through the inhibition of homophony-producing sound changes (Wedel, Kaplan, & Jackson, 2013; Wedel, Jackson, & Kaplan, 2013); this topic is explored more in the General Discussion.

Following past work (Trott & Bergen, 2020), we identified the optimal *n* for each lexicon using a cross-validation procedure. For each lexicon, we performed a train/test split (75% train, 25% test). Then, we fit a series of *n*-phone models ranging from $n = 1$ to $n = 6$ on the training set, and used these trained models to calculate the phonotactic probability of wordforms in the test set. Importantly, we performed this procedure 10 times for each value of *n*, to ensure that the results were not too sensitive to a particular train/test split. The optimal *n* was defined as the value that maximized the probability of wordforms in the held-out test set—i.e., large enough to capture the appropriate dependencies, but not so large that it overfit to the training set. This procedure resulted in $n = 5$ for English, Dutch, and German; and $n = 4$ for French and Mandarin. (Note that tones were treated as phonemes in the phonotactic model; exploratory analyses suggest that the *n*-phone model captured statistical regularities in which tones co-occurred with the internal structure of the corresponding syllable, but future work could ask about the impact of conditioning tones on particular segments of the preceding syllable (Kirby, 2021).)

Finally, we fit an *n*-phone model to each lexicon using all unique word types (rather than the 75% training set). (Word types, rather than tokens, were chosen to be consistent with past work (Piantadosi et al., 2012; Trott & Bergen, 2020), and to avoid conflating phonotactic probability with word frequency.)

Phonotactic Baselines

Following past work (Dautriche et al., 2017; Trott & Bergen, 2020; Caplan et al., 2020), we used the trained phonotactic models to simulate a series of phonotactic baselines for each language. Unlike past work, we built three different types of baselines (described below), with ten versions for each baseline (for a total of thirty baselines per language).

Neutral Baselines The procedure for generating Neutral baselines was identical to the procedure adopted in past work (Trott & Bergen, 2020). We first identified the number of lemmas (not wordforms) per word length (e.g., the English lexicon has 7,706 monosyllabic lemmas). Then, we used the phonotactic model to generate novel wordforms; each artificial lexicon was constrained to have the same distribution of words per word length as the real lexicon. For example, if an artificial lexicon already had the maximum number of monosyllabic words allowed, future monosyllabic words generated by the model would be discarded. This procedure was continued until the artificial lexicon had the same number as lemmas (not necessarily wordforms) as the real lexicon. Importantly, there was no constraint on the number of "meanings" a given wordform could acquire (i.e., the same wordform could be sampled an arbitrary number of times, provided more words of that length were required).

Anti-Homophony Baselines The Anti-Homophony Baselines followed an identical procedure as the Neutral Baselines, with one additional constraint: no wordform was al-

lowed to acquire more meanings than the equivalently-ranked wordform in the real lexicon's rank distribution of homophones. That is, if the most homophonous wordform in English had eight meanings, then no wordform in the baseline would be allowed to acquire more than eight meanings—and if the tenth most homophonous wordform had only three meanings, then the tenth most homophonous wordform in the baseline could acquire at most three meanings.

Conceptually, this pressure is akin to "blocking" new meanings from being attached to overly homophonous wordforms, and finding an alternative wordform instead. This is similar (though not identical) to instead adding a new word to the lexicon with some probability p, where p decays with the number of meanings already assigned to that wordform.

Anti-Homophony+ Baselines Finally, we considered an alternative implementation of an Anti-Homophony pressure. Rather than simply discarding overly homophonous wordforms, we applied a *sound change* to one of the phonemes in the target wordform. First, we randomly selected a phoneme in the target wordform to change. Then, we replaced it with a random vowel or consonant (depending on the identity of the phoneme). Finally, to ensure that the resulting wordform was sensible, we evaluated its phonotactic probability; if the wordform's probability was higher than the least-probable wordform in the real lexicon, we added it to the lexicon (provided it also did not have too many homophones).

The motivation for this procedure was that a pressure against homophony may not manifest as "blocking" the offending wordform entirely—overly homophonous wordforms likely have many desirable properties as wordforms of that language (i.e., they are short and well-formed). Thus, this anti-homophony pressure would preserve many of these desirable properties (most of the wordform remains intact) while also avoiding an excess of ambiguity.

Note that this procedure could arguably be interpreted as also implementing an indirect, *pro-neighbor* pressure, given that offending wordforms are directly converted to minimal pairs. However, this pro-neighbor pressure need not necessarily be *pro-neighborhood* per se—if the offending homophones are converted to existing wordforms, the distribution of meanings across wordforms could change without altering the distribution of neighborhood sizes.

Results

Replication of Homophony Results

Past work (Trott & Bergen, 2020; Caplan et al., 2020) found that phonotactic baselines without a pressure against homophones exhibited a higher upper-bound of homophony: the Maximum Number of Homophones (i.e., the number of meanings assigned to the most homophonous wordform, minus one) was larger in the baselines than their real counterparts. As depicted in Figure 1, we replicated this effect: Neutral baselines consistently contained higher levels of homophony than the real lexica.[5]

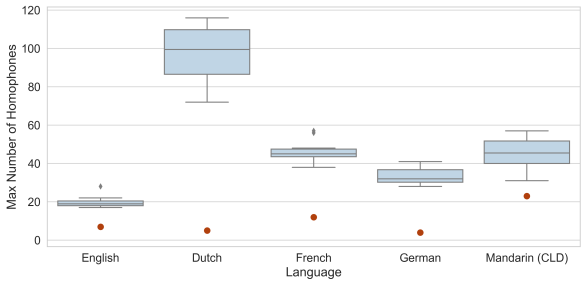

Figure 1: Maximum Number of Homophones across the real lexica and Neutral baselines. Red circles represent the values for the real lexicon.

Comparing Neighborhood Sizes

We used two primary dependent variables to compare the relative density of neighborhoods across real and artificial lexica: Mean Neighborhood Size and Maximum Neighborhood Size.[6] The neighborhood size of a given wordform was defined as the number of wordforms that were exactly one *edit* away, i.e., using either insertion, deletion, or substitution. Thus, the Mean Neighborhood Size was the average phonological neighborhood size across the entire lexicon, while the Maximum Neighborhood Size was the size of the densest neighborhood in a given lexicon.

Consistent with past work (Dautriche et al., 2017; Trott & Bergen, 2020), the real lexica had larger Mean Neighborhood Sizes and Maximum Neighborhood Sizes, compared to the Neutral baselines. For example, the Mean Neighborhood Size in English was 2.51, while the Neutral English baselines ranged from 2.23 to 2.32 ($M = 2.28, SD = 0.03$). Similarly, the Maximum Neighborhood Size in Dutch was 42, while the Neutral Dutch baselines ranged from 25 to 30 ($M = 27.3, SD = 1.89$). This demonstrates that phonotactics alone cannot account for neighborhood density in real lexica.

Yet as depicted in Figure 2, this gap largely disappeared (or in some cases, reversed) with the introduction of a pressure against over-saturation. Across all languages, the Mean Neighborhood Size was at least as large in the Anti-Homophony baselines. For example, in English, the Mean Neighborhood Size of the Anti-Homophony baselines ranged from 2.52 to 2.59 ($M = 2.54, SD = 0.03$) (recall that the value for the real English lexicon was 2.51). In some languages (e.g., Dutch and German), the Anti-Homophone baselines actually had *larger* neighborhoods on average. The gap was also attenuated for Maximum Neighborhood Size (see Figure 3). However, the largest neighborhoods in real lexica tended

[5]The Anti-Homophony and Anti-Homophony+ baselines are excluded from this figure, given that their levels of homophony were constrained not to exceed the real lexicon.

[6]Equivalent results were obtained using the Total Number of Minimal Pairs within a lexicon, as in past work (Dautriche et al., 2017).

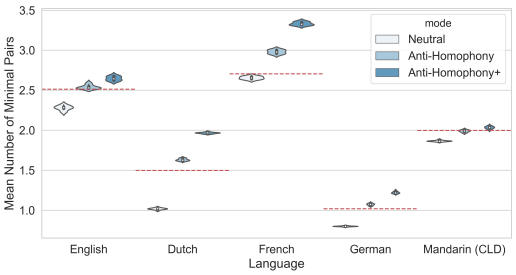

Figure 2: Mean Neighborhood Size as a function of language and lexicon type. Red lines represent the value for each real lexicon.

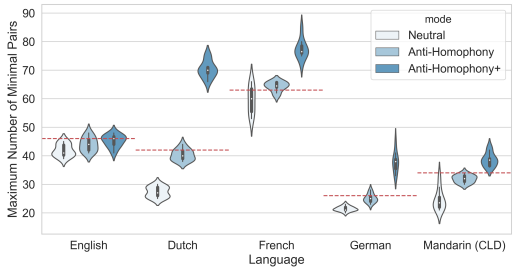

Figure 3: Maximum Neighborhood Size as a function of language and lexicon type. Red lines represent the value for each real lexicon.

to be slightly larger than the median value in the baselines (with the exception of French).

Surprisingly, the Anti-Homophony+ baselines exceeded both the Mean and Maximum Neighborhood Sizes of their real counterparts, sometimes to a very large degree (e.g., in French and Dutch). Further, the Anti-Homophony+ baselines *overestimated* the average neighborhood size across all languages tested.

In order to quantify which baseline produced the best *fit*, we calculated the Mean Error (ME) between the rank distribution of neighborhood sizes for each real lexicon and its artificial baselines. ME is defined as:

$$ME = \frac{\sum_{i=1}^{n} y_i - x_i}{n}$$

Where x_i was the neighborhood size from the real lexicon, and y_i was the neighborhood size of an equivalently ranked wordform from the baseline, i.e., the "predicted" neighborhood size. Mean Error was used (rather than mean absolute or squared error) to reveal the direction of the average error, i.e., whether a given baseline tended to underestimate or overestimate neighborhood sizes on average. As depicted in Figure 4, the Neutral baselines generally exhibited the worst fit (with the exception of French), and tended to underestimate neighborhood sizes. The Anti-Homophony baselines produced better predictions, and in fact, had the best fit for every language

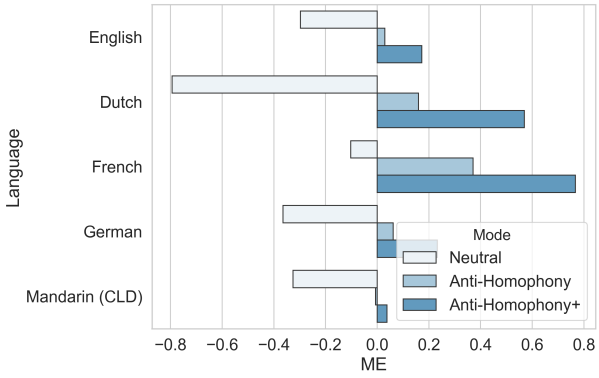

Figure 4: Mean Error (ME) for each baseline. Mean Error was computed by comparing the neighborhood sizes across each real lexicon and its artificial baselines; a score closer to zero corresponds to better fit.

but French (in which the predicted neighborhood sizes were too large on average). Finally, the Anti-Homophony+ baselines erred on the side of *overestimating* neighborhood sizes, to an even greater degree than the Anti-Homophony baselines.

General Discussion

We asked whether the distribution of neighborhood sizes in real lexica could be explained by the combination of phonotactic constraints and a pressure against homophony. Past work (Dautriche et al., 2017) found that phonotactics alone were insufficient to account for neighborhood sizes in real lexica, suggesting that real lexica are shaped by a positive selection pressure for larger neighborhoods. This pro-neighborhood pressure would also be consistent with evidence that dense neighborhoods confer benefits on learning (Coady & Aslin, 2003; Storkel, 2004; Fourtassi et al., 2020) and production (Vitevitch, 2002). The current work replicated this effect, as well as the finding that phonotactics alone tends to *overestimate* the degree of homophony compared to real lexica (Trott & Bergen, 2020).

Critically, however, we found that introducing a pressure against homophony in the baselines resulted in substantially larger neighborhood sizes on average—eroding or even reversing (in French and Dutch) the gap between the real lexica and their baselines (see Figure 2). This also resulted in a larger *upper-bound* on neighborhood sizes in the baselines, though not always larger than the real lexica (see Figure 3). Finally, a pressure that converted overly homophonous wordforms to minimal pairs resulted in larger neighborhoods across the board—surpassing the Mean Neighborhood Size of real lexica, and attaining or surpassing the Maximum Neighborhood Size of real lexica.

Thus, a pressure against homophony was in many cases *sufficient* to account for average neighborhood sizes. This means that an explanation for average neighborhood sizes in real lexica need not posit a direct selection pressure for

these neighborhoods: the distribution of neighborhood sizes observed in real languages may be the sole result of phonotactics and a pressure against over-saturation. Additionally, the Anti-Homophone+ pressure actually overestimated neighborhood sizes in many cases.

While these results cannot rule out the possibility that neighborhoods are directly selected for (see below), they do demonstrate that a pro-neighborhood pressure may not be a *necessary* part of an explanation. Importantly, this would not be inconsistent with evidence that dense neighborhoods provide benefits to learning and production—but under this account, these benefits would simply be "positive externalities" of a causally unrelated pressure against over-saturation.

Limitations and Future Work

The work described here is limited in certain ways. First, the languages tested represent a limited subset of the world's languages. The sample was biased towards Indo-European languages (English, Dutch, German, and French), with one Sino-Tibetan language (Mandarin), and did not include languages from other major language families such as Austronesian or Niger-Congo. The languages reflect a convenience sample; they are the languages for which we could obtain lexical resources that contained phonological information at the level of individual lemmas.

A second limitation lies in the measures of neighborhood density used. We used the average and maximum neighborhood size in a lexicon. However, past work (Dautriche et al., 2017) also used more sophisticated measures of the network structure in a lexicon, such as the degree of *transitivity*. Future work in this vein could better quantify how exactly neighborhoods distribute across the lexicon, using tools from network analysis.

Third, as in past work (Dautriche et al., 2017; Caplan et al., 2020; Trott & Bergen, 2020), we used an *n*-phone model to learn the phonotactics of the target language. Recent work has used more sophisticated approaches, such as a generative model (Futrell, Albright, Graff, & O'Donnell, 2017) or LSTM neural network (Pimentel et al., 2021). Future work could ask how adopting an alternative approach to modeling phonotactics changes the distribution of neighborhood sizes in the baselines. That said, recent work (Trott & Bergen, 2022) did find comparable results using an LSTM and *n*-phone approach.

Fourth, our approach cannot cannot directly *disconfirm* the theory that real lexica are shaped by a pro-neighborhood pressure. At best, the baselines demonstrate the *sufficiency* of a particular set of constraints in explaining the distribution of neighborhood sizes, absent a direct pro-neighborhood pressure; there may still be *a priori* reasons to prefer a theory that posits such a pressure. The results do suggest that a pressure against homophony can in principle explain two seemingly independent facts—namely, that real lexica have fewer homophones, and larger neighborhoods, than predicted by their phonotactics—but they do not rule out the possibility of alternative explanations.

A fifth, related limitation is that the baselines do not illuminate the causal mechanisms by which an anti-homophony pressure could operate, either at the level of individual communicative constraints or diachronic language change. Future research would benefit from experimental work directly probing these causal mechanisms, e.g., whether errors made during learning homophones (Casenhiser, 2005) could result in minimal pairs. Similarly, researchers could build computational models of how these local pressures interact with changes operating over longer timescales, such as sound change (Wedel, Jackson, & Kaplan, 2013).

Sixth, this work did not consider other important variables, such as *frequency*—both of individual wordforms, and of the distinct lemmas conveyed by those wordforms. This is in part due to limitations of the simulation method used. Employing a different approach, recent work (Trott & Bergen, 2022) discovered several relevant findings: homophony resistance is positively correlated with the frequency of particular *wordforms*, though not necessarily with the relative frequency of their meanings; and further, homophony resistance is highest among wordforms with high neighborhood density—consistent with the results presented here.

Finally, these analyses made two simplifying assumptions. First, meanings were implicitly assumed to be discrete units, with no relation between them. However, meanings are likely at least partially continuous (Elman, 2009; Trott & Bergen, 2021; Li & Joanisse, 2021); further, some meanings are more related (as in polysemy) than others (as in homonymy). Second, forms were assumed to be arbitrarily related to meanings—however, there is considerable evidence (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; Gutiérrez, Levy, & Bergen, 2016) that form-meaning relationships may be less arbitrary than previously thought. Future work could integrate both lines of thought by using a continuous representation of the meaning space, and exploring different ways of assigning form-meaning pairings in either systematic or arbitrary ways.

Conclusion

Why do real lexica have such large phonological neighborhoods? One explanation is that real lexica are subject to a selection pressure for dense neighborhoods, possibly because dense neighborhoods facilitate word learning (Storkel, 2004; Coady & Aslin, 2003) and production (Vitevitch, 2002; Vitevitch & Sommers, 2003). We pursued another possibility—that dense neighborhoods emerge from the interaction of other constraints operating over real lexica, namely phonotactics and a pressure against individual wordforms acquiring too many meanings (Trott & Bergen, 2020). We tested the sufficiency of this latter account using simulated baselines. Crucially, the combination of phonotactic constraints and an anti-homophony pressure was *sufficient* to account for average neighborhood sizes in real human lexica—demonstrating that a direct selection pressure for neighborhood density is not a *necessary* part of an explanation.

Acknowledgments

We are grateful to the reviewers for their valuable feedback. We also thank members of the Language and Cognition Lab (James Michaelov, Cameron Jones, Tyler Chang) for helpful comments and discussion.

References

Arutiunian, V., & Lopukhina, A. (2020). The effects of phonological neighborhood density in childhood word production and recognition in russian are opposite to english. *Journal of Child Language*, *47*(6), 1244–1262.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The celex lexical database. *Distributed by the Linguistic Data Consortium: University of Pennsylavia.*.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823.

Caplan, S., Kodner, J., & Yang, C. (2020). Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition*, *205*, 104466.

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, *32*(2), 319–343.

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child language*, *30*(2), 441–469.

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

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, *33*(4), 547–582.

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: Insight from 10 languages. *Cognitive Science*, *44*(7), e12847.

Frisch, S. A., Large, N. R., & Pisoni, D. 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.

Futrell, R., Albright, A., Graff, P., & O'Donnell, T. J. (2017). A generative model of phonotactics. *Transactions of the Association for Computational Linguistics*, *5*, 73–86.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389–407.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the royal society of London. Series B. Biological Sciences*, *205*(1161), 581–598.

Gutiérrez, E. D., Levy, R., & Bergen, B. (2016). Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2379–2388).

Jones, S. D., & Brandt, S. (2019). Do children really acquire dense neighbourhoods? *Journal of child language*, *46*(6), 1260–1273.

Jones, S. D., & Brandt, S. (2020). Density and distinctiveness in early word learning: Evidence from neural network simulations. *Cognitive science*, *44*(1), e12812.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kirby, J. (2021). Incorporating tone in the calculation of phonotactic probability. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 32–38).

Levshina, N., & Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, *7*(s3).

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), e12955.

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

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

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

Pimentel, T., Meister, C., Teufel, S., & Cotterell, R. (2021). On homophony and r\'enyi entropy. *arXiv preprint arXiv:2109.13766*.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.

Richie, R. (2016). Functionalism in the lexicon: Where is it, and how did it get there? *The Mental Lexicon*, *11*(3), 429–466.

Sampson, G. (2013). A counterexample to homophony avoidance. *Diachronica*, *30*(4), 579–591.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in

lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning.

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (cld). *Behavior Research Methods*, *50*(6), 2606–2629.

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

Trott, S., & Bergen, B. (2021). Raw-c: Relatedness of ambiguous words in context (a new lexical resource for english). In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 7077–7087).

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

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735.

Vitevitch, M. S., & Aljasser, F. M. (2021). Phonotactics in spoken-word recognition. *The Handbook of Speech Perception*, 286–308.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325–329.

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in spanish. *Journal of Multilingual Communication Disorders*, *3*(1), 64–73.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & cognition*, *31*(4), 491–504.

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and speech*, *56*(3), 395–417.

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

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.