Cognition 205 (2020) 104449

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

# Why do human languages have homophones?

Sean Trott*, Benjamin Bergen

*Department of Cognitive Science, UC San Diego, 9500 Gilman Dr., La Jolla, CA 92093, United States of America*

ARTICLE INFO

*Keywords:*
Homophones
Ambiguity
Efficiency
Language evolution
Phonotactics

ABSTRACT

Human languages are replete with ambiguity. This is most evident in homophony—where two or more words sound the same, but carry distinct meanings. For example, the wordform "bark" can denote either the sound produced by a dog or the protective outer sheath of a tree trunk. Why would a system evolved for efficient, effective communication display rampant ambiguity? Some accounts argue that ambiguity is actually a design feature of human communication systems, allowing languages to recycle their most optimal wordforms (those which are short, frequent, and phonotactically well-formed) for multiple meanings. We test this claim by constructing five series of artificial lexica matched for the phonotactics and distribution of word lengths found in five real languages (English, German, Dutch, French, and Japanese), and comparing both the quantity and concentration of homophony across the real and artificial lexica. Surprisingly, we find that the artificial lexica exhibit higher upper-bounds on homophony than their real counterparts, and that homophony is even more likely to be found among short, phonotactically plausible wordforms in the artificial than in the real lexica. These results suggest that homophony in real languages is not directly selected for, but rather, that it emerges as a natural consequence of other features of a language. In fact, homophony may even be selected against in real languages, producing lexica that better conform to other requirements of humans who need to use them. Finally, we explore the hypothesis that this is achieved by "smoothing" out dense concentrations of homophones across lexical neighborhoods, resulting in comparatively more minimal pairs in real lexica.

## 1. Introduction

Human languages are replete with ambiguity. This is most evident in *homophony*—where two or more words sound the same, but carry distinct meanings. For example, the wordform "bark" can denote either the sound produced by a dog or the protective outer sheath of a tree trunk. Estimates of the rate of homophony in English range from 7.4% (Rodd, Gaskell, & Marslen-Wilson, 2002) to over 15%[1] (Baayen, Piepenbrock, & Gulikers, 1995). Dautriche (2015) estimates the average homophony rate across languages to be 4%, with considerable cross-linguistic variability, ranging from approximately 3% in Dutch to 15% in Japanese. The prevalence of homophony, like other kinds of ambiguity, is confounding on its face. Human languages are generally thought to be shaped by pressures for efficient, effective communication (Gibson et al., 2019; Zipf, 1949). Yet ambiguity increases both the effort required for comprehension and the likelihood of miscommunication. A comparison between human and programming languages places this into relief. Programming languages, designed for efficient and errorless communication, generally abide no ambiguity at all. Why then do human languages insist on encoding distinct messages identically? Why are homophones so common?

Part of the answer appears to be that human comprehenders are adept at disambiguating ambiguous input using various contextual cues (Ferreira, 2008; Levinson, 2000; Piantadosi, Tily, & Gibson, 2012; Wasow, Perfors, & Beaver, 2005). In the case of homophones, a wide array of cues to meaning are available, including the syntactic structures that words are embedded in (Dautriche, Fibla, Fievet, & Christophe, 2018), gestures that accompany speech (Holle & Gunter, 2007; Holler & Beattie, 2003; Kidd & Holler, 2009), and statistical aspects of linguistic context (Aina, Gulordava, & Boleda, 2019). The human capacity for disambiguation thus creates a tolerant environment for ambiguous wordforms—explaining why as languages evolve, homophones might not be strictly selected against.

But might homophones also be selected for? Zipf (1949) argues that ambiguity is a design feature of any human communication system, resulting from a direct pressure for efficiency. A growing body of evidence is consistent with the claim that lexica are optimized for efficient communication between humans (Gibson et al., 2019; Piantadosi, Tily,

* Corresponding author at: Department of Cognitive Science, 9500 Gilman Dr., La Jolla, CA 92093-0515, United States of America.
*E-mail address:* sttrott@ucsd.edu (S. Trott).

[1] Estimates of the rate of *polysemy* (wordforms with related meanings) are considerably higher: up to 80% of wordforms in English are thought to be polysemous (Rodd et al., 2002).

https://doi.org/10.1016/j.cognition.2020.104449
Received 14 January 2020; Received in revised form 11 August 2020; Accepted 23 August 2020

& Gibson, 2009), from the way they carve up semantic domains (Kemp & Regier, 2012; Kemp, Xu, & Regier, 2018; Regier, Kay, & Khetarpal, 2007; Xu & Regier, 2014; Zaslavsky, Kemp, Regier, & Tishby, 2018) to the wordforms that they contain (Mahowald, Dautriche, Gibson, & Piantadosi, 2018; Piantadosi et al., 2012; Piantadosi, Tily, & Gibson, 2011). This pressure for an efficient lexicon could result in a selective bias for wordforms that are particularly easy to produce and comprehend, where *ease* reflects properties such as a word's length, phonotactic plausibility, and frequency. Combined with a tolerance for ambiguity, a bias for easy wordforms could exert a pressure on lexica to "recycle" particularly optimal wordforms for multiple meanings. This pressure, termed "unification" by Zipf (1949), would increase efficiency by reducing the number of unique wordforms that speakers need to learn and encode. Furthermore, by preferentially re-using the most optimal wordforms, such a lexicon would arguably involve less effort in speaking or writing than an unambiguous linguistic system. If such a pressure exists, it should produce concentrations of homophony in optimal regions of phonotactic space—the "easiest" wordforms should be the most ambiguous. Indeed, Piantadosi et al. (2012) find that English, German, and Dutch count more homophones among wordforms that are short, frequent, and phonotactically well-formed. This finding is consistent with the idea that ambiguity arises out of a pressure for efficiency.

However, homophony could also emerge in a lexicon without being directly selected for, as an indirect consequence of other factors affecting how words are distributed in a lexicon. Two indirect mechanisms could also partially (or even fully) account for the uneven distribution of homophony across a lexicon.

First, the *proportion of occupied phonotactic space* (i.e., the ratio of actual wordforms to possible wordforms) for English and every other language we are aware of will always be higher for shorter wordforms than for longer wordforms. This is because the number of possible wordforms of a given length grows exponentially with each added syllable. If a language's phonotactics permit $n$ unique syllables, then there are $n$ possible monosyllabic wordforms, approximately $n^2$ possible bisyllabic wordforms, approximately $n^3$ possible trisyllabic wordforms, and so on. In contrast, the number of actual wordforms does not grow exponentially with word length (e.g., the CELEX set of English lemmas contains approximately 7706 monosyllabic words, 15,247 disyllabic words, and 11,379 trisyllabic words). This means that the proportion of occupied phonotactic space will always be greater among short wordforms than long wordforms. Thus, even if words were randomly added to a lexicon, homophony would by chance be more likely to occur among short wordforms than long wordforms.

Second, the existence of *phonotactic constraints* results in a lexicon that is not uniformly distributed across the space of possible wordforms. All languages appear to impose idiosyncratic constraints on sounds and their combinations—for example, English does not allow the velar nasal /ŋ/ in syllable onsets, unlike Vietnamese; but English does allow consonant clusters like /st/, unlike Japanese. Phonotactic regularities narrow the space of possible wordforms considerably (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017). By limiting the range of possible wordforms and biasing the formation and evolution of the lexicon, these phonotactic constraints could also increase the prevalence of homophones. Critically, they could do so even without a direct pressure to reuse entire wordforms. Even a pressure to merely statistically reuse certain phonological sequences more often would increase the likelihood of homophones overall, and particularly among the most phonotactically probable wordforms.

Both of these mechanisms offer indirect causal pathways whereby a drive for efficiency could lead to increased homophony. For example, more phonotactically regular words could be easier to learn (Coady & Aslin, 2004; Gathercole, Willis, Emslie, & Baddeley, 1991; Jusczyk, Luce, & Charles-Luce, 1994; Munson, 2001), which would lead to more phonotactically probable words being more likely to be transmitted across generations, or less phonotactically words becoming more phonotactically probable through imperfect intergenerational transmission. This in turn could result in increased homophony, particularly among highly probable wordforms. Once again, though, both phonotactics and the distribution of word lengths in a lexicon could in principle lead to the emergence of homophony without a direct, selective pressure for the preferential reuse of specific, optimal wordforms (as hypothesized by Zipf, 1949). Furthermore, both factors should be most likely to produce homophones in exactly those regions of phonotactic space reported by Piantadosi et al. (2012): among short, phonotactically plausible wordforms.

It is currently unknown, however, how much homophony exists due to these simple, distributional characteristics of languages alone. As a consequence, no evidence exists for or against an efficiency-motivated direct pressure for homophony, as hypothesized by Zipf. The current work asks two primary questions. First, to what extent is the **amount** of homophony found in real human lexica attributable to indirect and uncontroversial factors such as length and phonotactic regularities, without a direct pressure to reuse existing wordforms? And second, to what extent are these indirect factors responsible for the **concentration** of homophony within optimal regions of the lexicon?

To answer these questions, we constructed five series of artificial lexica, designed to mirror the phonotactic regularities and word lengths of the real lexica of English, Dutch, German, French, and Japanese. The generative model was an adaptation of the model used by Dautriche et al. (2017), in which a language's phonotactics were learned by training an $n$-phone Markov Model on the set of unique wordforms in a lexicon. By observing the patterns of sounds and sound combinations in a language, such a model can learn to encode phonotactic rules about which sounds a word can start and end with, which sounds can occur in what sequence, and so on. For each language, this model was then used to generate 10 artificial lexica, all matched for the total number of words as well as the distribution of word lengths. For example, if the real lexicon has 5000 monosyllabic words, then each of the artificial lexica will also have 5000 monosyllabic words. Furthermore, the distribution of sounds within and across those words will approximate the phonotactics of the real language. These artificial lexica had no constraints regarding homophones, reflecting a general tolerance for ambiguity; however, they also did not contain a parameter biasing them *towards* the reuse of existing wordforms. Each artificial lexicon thus represents one answer to the questions: 1) **how much** homophony can be expected to emerge in a lexicon as a function of just the real, observed phonotactic regularities and the real, observed distribution of word lengths; and 2) where should we expect to find the largest **concentrations** of homophony as a function of these factors? They thus serve as a baseline characterization of the effects of indirect causes of homophony. Comparing the real lexica to these artificial ones reveals how much more or less homophony the real languages display—and how much more or less concentrated it is—than would be expected without any direct pressure for or against homophony.[2]

Note that these artificial lexica are not intended to serve as plausible models of lexicon formation and change. Rather, as described above, they serve as statistical baselines in the attempt to understand which theoretical parameters are necessary to explain the existence and distribution of homophony in real lexica. For this reason, the artificial lexica are parameterized solely by each particular language's phonotactics and distribution of word lengths.

The data and code to reproduce these analyses can be found on GitHub (https://github.com/seantrott/homophone_simulations).

[2] Note that our statistical models do not include a measure of frequency, even though this is included in the original model built in Piantadosi et al. (2012). This is because it would not be meaningful to estimate frequency for the words in the artificial lexica.

## 2. Current work

### 2.1. Materials and methods

#### 2.1.1. Data

The English, German, and Dutch lexica were sourced from the CELEX lexical database (Baayen et al., 1995). For French, we used the French Lexique (New, Pallier, Brysbaert, & Ferrand, 2004). For Japanese, we used the Japanese CallHome Lexicon (Kobayashi et al., 1996). We restricted our analysis to lemma-only forms. Additionally, following Piantadosi et al. (2012), we also excluded any words containing spaces, hyphens, or apostrophes. This resulted in 41,887 entries for English (with 35,107 unique phonological forms), 51,719 entries for German (with 50,435 unique phonological forms), 67,477 entries for Dutch (with 65,260 unique phonological forms), 47,782 entries for French (with 37,278 unique phonological forms), and 51,147 entries for Japanese (with 40,449 unique phonological forms). As in Piantadosi et al. (2012), words with multiple parts of speech were counted as homophones.[3]

### 2.2. Methods

#### 2.2.1. Estimating number of syllables

Our primary determinant of word length was Number of Syllables (or Number of Morae, in the case of Japanese; see below). While the real lexica annotated this information for each lexical entry, it had to be estimated for the artificial lexica. To ensure a fair comparison, we applied the same estimation procedure to wordforms in the real lexica and wordforms in the artificial lexica.

For English, Dutch, German, and French, Number of Syllables was estimated by counting the number of vowels occurring in a wordform's phonetic transcription. The set of possible vowel characters for a given language was transcribed by hand and can be found on the project's GitHub page.[4]

Since Japanese has been characterized as a mora-timed, rather than syllable-timed language (Port, Dalby, & O'Dell, 1987), we calculated Number of Morae instead of Number of Syllables. In addition to counting the number of vowels in a Japanese wordform, we counted the number of nasal codas, as well geminate consonants (e.g., "kk" in *Hokkaido*, or "gg" in *doggu*). It should be noted that the results we report below—both the replication of Piantadosi et al. (2012), and the comparison to the artificial lexica—are qualitatively similar whether word length in Japanese is estimated using Number of Syllables or Number of Morae.

#### 2.2.2. Counting number of homophones

Following Piantadosi et al. (2012), we defined Number of Homophones as the number of lexical entries with an identical phonological form as some target entry.[5] This means the smallest possible value for Number of Homophones would be 0 (i.e., there are no other words with the same form in a given lexicon), and the largest possible value would be one less than the size of the lexicon (i.e., all words share the same form).

After identifying the number of homophones for each entry in a lexicon, we reduced each lexicon to the set of unique phonological wordforms (e.g., the 41,887 entries in English were reduced to 35,107 unique forms).

#### 2.2.3. Building the phonotactic model

In order to estimate the phonotactic plausibility of wordforms in a lexicon, as well as to generate phonotactically plausible novel wordforms (see below), it was first necessary to model the phonotactics of each language. We adapted the procedure used in Dautriche et al. (2017),[6] which is described briefly below.

The phonotactics of a target language can be learned by observing, for all wordforms in that language, which phonemes appear in what position and in what sequence. Specifically, an *n*-phone model calculates the probability of observing some phoneme in position *i* given the previous *n-1* phonemes. For example, a 2-phone (biphone) model would condition the probability of observing some phoneme as a function of the previous phoneme, i.e., $p(X_i \mid X_{i-1})$. We included special symbols for the START and END of a word so that the model would also learn which phonemes are most likely to begin and end a word in a given language. Note that unlike Piantadosi et al. (2012), these models were trained using the set of unique *types* (i.e., wordforms), rather than *tokens* (i.e., the actual instances of each wordform); this is because training on tokens conflates phonotactic probability with frequency. This is analogous to the main approach taken in Dautriche et al. (2017).

While previous work (Dautriche et al., 2017) found that a 5-phone model effectively captured phonotactic dependencies in English, Dutch, German, and French, we sought to independently determine the optimal *n* for each language, particularly because Japanese has notably shorter syllables than the other four languages. To do this, we followed a similar procedure as reported in Dautriche et al. (2017) and Futrell, Albright, Graff, and O'Donnell (2017). For each real lexicon, we first extracted the set of unique wordforms (e.g., 35,107 wordforms in English), then performed a series of train/test splits (75% train, 25% test). For each split, we trained a series of *n*-phone models ranging from $n = 1$ to $n = 6$ on the wordforms in the training set, then evaluated the probability of wordforms in the held-out test set. The basic motivation for this approach is as follows: the optimal *n*-phone model for a language's phonotactics should be the model that, when trained on a set of real wordforms, maximizes the probability of held-out wordforms that also appear in that lexicon. Following Futrell et al. (2017), we ran a series of one-tailed two-sample *t*-tests on the set of log-likelihoods of held-out wordforms obtained from each successive *n*-phone model—i.e., the log-likelihoods obtained from the 2-phone model were compared to the 1-phone model, those from the 3-phone model were compared to the 2-phone model, and so on. The optimal *n* for a given lexicon was the smallest *n* that represented a significant improvement over the *n-1* model for the same set of wordforms. Note that $\log_{10}$ was used to calculate log likelihoods (and subsequently, surprisal); the results are not qualitatively different when using $\log_2$ instead.

The mean log-likelihood calculated for held-out wordforms in each language are visualized in Fig. 1 below. Critically, we found that for English, Dutch, and German, the 5-phone model represented a significant improvement over the 4-phone model. That is, held-out wordforms were significantly more likely under the 5-phone model than the 4-phone model for English ($t = 4.05$, $p < .001$), Dutch ($t = 3.55$, $p < .001$), and German ($t = 7.31$, $p < .001$). However, the 6-phone model either did not improve or actually decreased model fit (suggesting overfitting) in each language (all $t \leq 0$). The 4-phone model was optimal for French ($t = 8.67$, $p < .001$) and Japanese ($t = 4.08$, $p < .001$). Thus, a 5-phone model was used to evaluate the probabilities of wordforms in English, Dutch, and German (and

[3] Importantly, this should only serve to *inflate* the estimated amount of homophony in naturally-occurring languages relative to the amount of homophony in the artificial lexica. Thus, it would actually work against the effects reported below (i.e., the artificial lexica exhibiting more homophony than the real lexica).

[4] Link: https://github.com/seantrott/homophone_simulations.

[5] As pointed out by an anonymous reviewer, it is possible that the lexical resources we used, including CELEX, count as homophony some meanings that are actually polysemous. If this is the case, our estimates of homophony should actually be inflated for the real lexica, which would work against the effects reported below (i.e., the artificial lexica displaying higher incidences of homophony overall).

[6] Link to GitHub associated with Dautriche et al. (2017): https://github.com/SbllDtrch/NullLexicons.

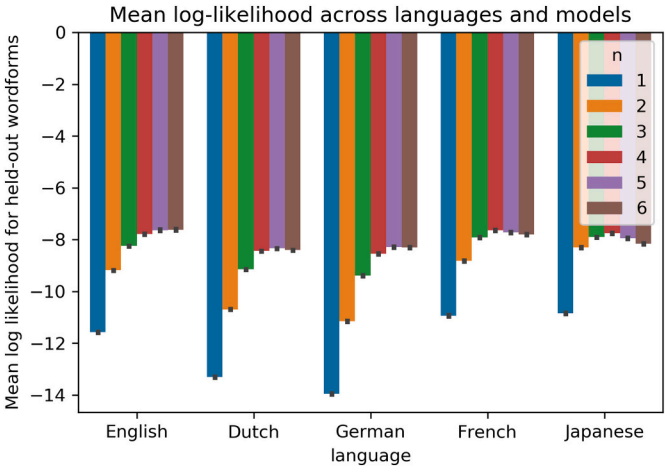


**Fig. 1.** Mean log-likelihood of held-out wordforms for each *n*-phone model, across languages. Higher values (i.e., less negative) indicate higher probability under that model. For English, Dutch, and German, increasing *n* up to 5 significantly improved model fit over the 4-phone model; a 6-phone model did not improve fit. For French and Japanese, a 4-phone model was the highest *n* representing an improvement over the (*n-1*) model.

generate artificial lexica for those languages), and a 4-phone model was used for French and Japanese.

This model allows us to evaluate the probability of a given wordform, which can be defined as the product of all the transitional probabilities between each phoneme in that wordform (including the start and end symbols). The Surprisal of a given wordform is thus defined as the negative log probability of observing that particular sequence of phonemes: Surprisal(word) = −log(p(word)). As in Piantadosi et al. (2012), we normalized this measure to the number of phonemes in a word to ensure that surprisal could be compared across words of different length: Normalized Surprisal = Surprisal(word) / Length(word).

Once the model was built for each language, it was then used to generate novel wordforms in an iterative manner. For each word, the model began with the START symbol, then generated a phoneme conditioned on that start symbol (i.e., one of the phonemes likely to occur at the beginning of the word). The next phoneme was then conditioned on the first phoneme and the START symbol, and so on, until the model produced the END symbol, signaling the end of the word.

Finally, as in Dautriche et al. (2017), we assigned non-zero probability to unobserved phoneme sequences using an identical smoothing procedure; they report that "optimal smoothing was obtained with Laplace smoothing with parameter .01" (pg. 132), so this was the value we used in configuring the phonotactic model.

#### 2.2.4. Generating artificial lexica

We generated 10 artificial lexica for each real lexicon. First, we identified the number of words in the real lexicon, as well as the distribution of their lengths, as measured by Number of Syllables (see above for the estimation procedure). Each artificial lexicon was constrained to have the same overall number of *words* (not *wordforms*) as the corresponding real lexicon, as well as the same distribution of word lengths. For example, since the real English lexicon has 7706 monosyllabic words, each artificial English lexicon was also constrained to have 7706 monosyllabic words.

We then built a phonotactic model for the real lexicon as described above, and used this model to generate wordforms for each artificial lexicon. For each potential wordform, we estimated the Number of Syllables to determine whether to add it to the artificial lexicon—e.g., if the word had 1 syllable and the artificial lexicon still had fewer monosyllabic words than the real lexicon, the word was added to the lexicon; otherwise, it was discarded. No other constraints were placed on the generation of wordforms; we allowed the model to generate real wordforms, as well as wordforms that were homophonous with wordforms already in the lexicon. This process continued until the artificial lexicon had the same number of words of each length as the real lexicon.

Note that the models used to generate the artificial lexica were trained on the entire set of unique wordforms for the target lexicon; however, qualitatively similar results were obtained using a 50/50 split of the target lexicon to generate and evaluate wordform phonotactic probability (see Supplementary Analysis 5).

### 2.3. Results

#### 2.3.1. Replication and extension of previous findings

First, we replicated the primary analysis reported by Piantadosi et al. (2012) on the real lexica of English, Dutch, and German, and extended this analysis to two non-Germanic languages: French and Japanese. Using a Poisson regression, we asked whether a wordform's #Homophones (the number of additional, distinct meanings) was related to its length in syllables (#Syllables) and its phonotactic plausibility (Surprisal). As in Piantadosi et al. (2012), we used the Normalized Surprisal measure described above, obtained by dividing a wordform's Surprisal by its length in phones.

We found significant, negative relationships in the real lexica between #Homophones and #Syllables (or #Morae[7] in Japanese) for English [$\beta = -0.72$, SE = 0.03, $p < .001$], German [$\beta = -0.69$, SE = 0.04, $p < .001$], Dutch [$\beta = -1.11$, SE = 0.03, $p < .001$], French [$\beta = -0.35$, SE = 0.02, $p < .001$], and Japanese [$\beta = -1.01$, SE = 0.01, $p < .001$]. That is, for all five languages, shorter wordforms were more likely to have more homophones—consistent with the notion that lexica recycle short wordforms for multiple meanings.

However, we found positive[8] relationships between Normalized Surprisal and #Homophones across all real languages but Japanese, i.e., *less* phonotactically plausible wordforms (as measured by a 5-phone model or 4-phone model, as appropriate) were more likely to have more homophones. This was true for English [$\beta = 0.78$, SE = 0.03, $p < .001$], German [$\beta = 0.86$, SE = 0.06, $p < .001$], Dutch [$\beta = 0.997$, SE = 0.04, $p < .001$], French [$\beta = 0.73$, SE = 0.04, $p < .001$], but not Japanese [$\beta = 0.0004$, SE = 0.031, $p = .99$]. This is in contrast to the original result reported by Piantadosi et al. (2012), who found a negative relationship between Normalized Surprisal and #Homophones in German and Dutch.

There are several possible explanations for the disparity between our results and those of Piantadosi et al. (2012). First, while Piantadosi et al. (2012) used a 3-phone model to determine phonotactic plausibility, we used 4-phone and 5-phone models to estimate wordform probability, which were found to improve model fit over a 3-phone model (see Fig. 1). Second, our models were trained using lexical types, as opposed to tokens (which would conflate frequency with phonotactic probability). And third, our estimates were not calculated using held-out wordforms, as they were in Piantadosi et al. (2012). This final explanation is explored in Supplementary Analysis 3; using 10-fold cross-validation to obtain our surprisal estimates, we found that the coefficients for Normalized Surprisal were closer to 0 for all the real lexica, and negative in Japanese. Thus, a likely reason for the disparity is that

[7] Like syllables, a mora is a unit of timing, and is usually considered the basis of the sound system in Japanese. A single mora in Japanese is constituted by a vowel (or an onset and a vowel); nasal codas also constitute a separate mora, as does the first part of a geminate consonant.

[8] Note that negative relationships were obtained between the non-normalized Surprisal measure and Number of Homophones across each language; these results are described in Supplementary Analysis 2. However, this non-normalized Surprisal measure conflates phonotactic plausibility with word length, which is why Normalized Surprisal may be a better measure overall.

the surprisal estimates given here were not calculated using held-out wordforms.

However, the central question of the current work concerns the comparison between the real and artificial lexica. The results of these comparisons are described in detail below, both concerning the **amount** of homophony across the real and artificial lexica, as well as where those homophones are **concentrated**.

### 2.3.2. Simulated lexica exhibit higher upper-bounds on homophony

We operationalized the **amount** of homophony in three ways. First, we measured the Maximum Number of Homophones per wordform—that is, in a given lexicon, how many homophones does the most homophonous wordform have? Second, we measured the Mean Number of Homophones per wordform. And third, we measured Homophony Rate: how many wordforms in a lexicon have *at least* 1 homophone? In all cases, more positive values reflect a greater amount of homophony. For each measure in each language, we compared the distribution of values obtained from the simulated lexica to the value in the real lexicon. This enabled us to ask the question: to what extent can the **amount** of homophony in a language be attributed to a selective pressure for lexical ambiguity, as opposed to an emergent outcome of a language's phonotactics and distribution of word lengths? Note that for all of these measures, the values obtained for the real and artificial lexica were significantly different[9] ($p < .001$), except where noted otherwise.

Across all five languages, the simulated lexica had a significantly larger Maximum Number of Homophones on a single wordform (see Fig. 2 below). For example, the most homophonous wordforms in the real English lexicon had at most 7 homophones, while the most homophonous wordforms in the simulated English lexica had anywhere from 17 to 28 homophones (M = 19.8, SD = 3.3). This difference was particularly stark for Dutch: the most homophonous wordform in the Dutch lexicon had 5 homophones, while the maximum number of homophones per wordform in the simulated lexica ranged from 72 to 116 (M = 97.1, SD = 15.13).

As expected, there was considerable variability across the five languages in how much homophony was tolerated per wordform. For example, the real Japanese lexicon exhibited a much higher upper-bound on homophony (33) than the real German lexicon (4); this is not surprising, given the limited syllable inventory of Japanese (on the order of 100 possible syllables) relative to German (over 1000 possible syllables, conservatively). Importantly, however, the simulated Japanese lexica still had more homophones per wordform than their real counterpart, ranging from 71 to 92 (M = 81.6, SD = 6.67). In other words, despite inter-linguistic variability, the simulated lexica in each language all exhibited higher upper-bounds on how much homophony was tolerated for a given wordform—the most homophonous wordforms were considerably more ambiguous, sometimes by an order of magnitude (e.g., in Dutch).

Similarly, with the exception of Japanese ($p = .5$), wordforms in the simulated lexica had a significantly larger Mean Number of Homophones than wordforms from their real counterparts (see Fig. 3 below for an illustration); in Japanese, the Mean Number of Homophones per wordform was at least as high in the artificial lexica as it was in the real lexicon.[10] For example, wordforms in English have on

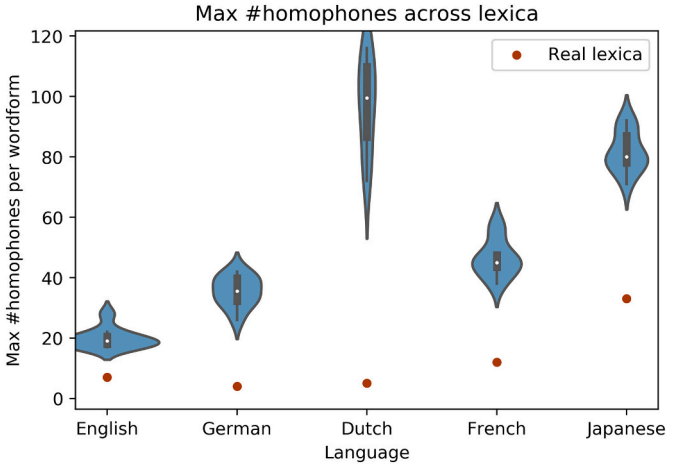


**Fig. 2.** For each language, the most homophonous wordforms in the artificial lexica (shown by the violin plots) have more homophones than the most homophonous wordforms in the real lexica (shown by the orange dots). The artificial lexica uniformly exhibit a higher upper-bound (Maximum Number of Homophones) on homophony. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

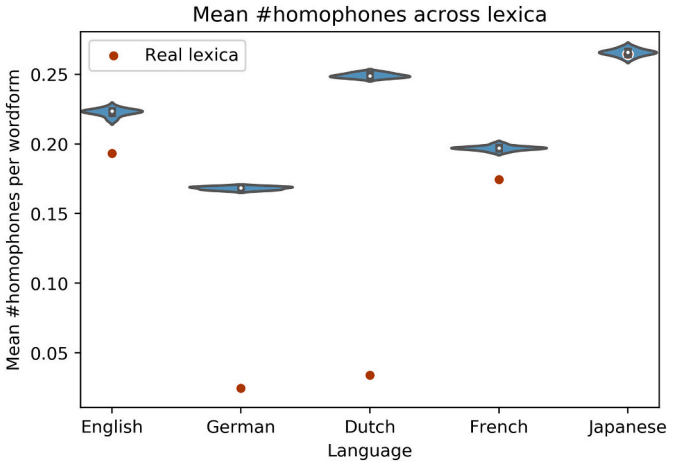


**Fig. 3.** In every language but Japanese, wordforms in the artificial lexica (shown by violin plots) have more homophones (Mean Number of Homophones) on average than wordforms in the real lexica (shown by orange dots). In Japanese, the Mean Number of Homophones per wordform is at least as high in the artificial lexica (M = 0.27, SD = 0.002) as the real lexica (0.26). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

average 0.19 homophones; in contrast, the average number of homophones per wordform in the simulated English lexica ranged from 0.22 to 0.23 (M = 0.22, SD = 0.003). Again, there was considerable inter-linguistic variability; wordforms in the real Japanese lexicon have more homophones on average (0.26) than wordforms in the real German lexicon (0.02). However, in each language, the average number of homophones per wordform was at least as large in the simulated lexica as the real counterparts—and for four of the five languages, wordforms in the simulated lexica were, on average, *more* ambiguous than those in the real lexica.

The results for the Homophony Rate (i.e., the proportion of wordforms with at least one homophone) across real and simulated lexica were more mixed (see Fig. 4 below).

In two languages (German and Dutch), the simulated lexica had significantly more homophonous wordforms, sometimes by a factor of 2× or 3×; for example, the homophony rate in the real Dutch lexicon was 0.03, while the rate in the simulated lexica ranged from 0.108 to

[9] Significance was determined by comparing a given test statistic for the real lexicon $t_{real}$ to the corresponding distribution of test statistics obtained from the artificial lexica, $T_{artificial}$. Each of these values was centered according to the mean of $T_{artificial}$, denoted here as $T'_{artificial}$ and $t'_{real}$. We then conducted a two-tailed significance test, i.e., calculating the proportion of values in $|T'_{artificial}|$ that were greater than or equal to $|t'_{real}|$. This proportion corresponds to a $p$-value; e.g., if all the values in $|T'_{artificial}|$ are less than $|t'_{real}|$, $p = 0$.

[10] Note that for Japanese, the Mean Number of Homophones per wordform is actually higher in the artificial lexica than the real lexicon with the use of a 5-phone model, rather than a 4-phone model.

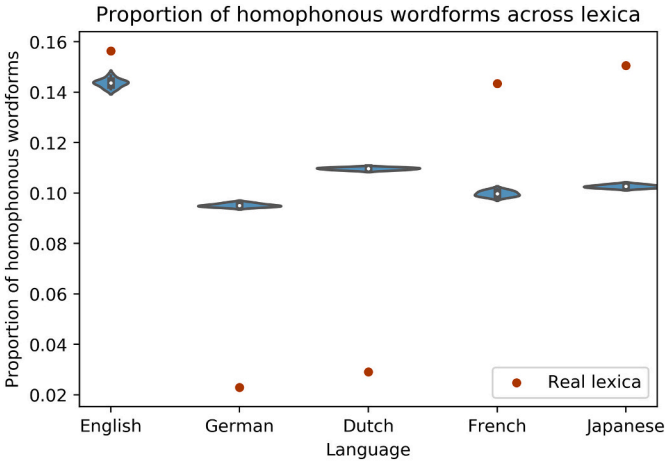


**Fig. 4.** The artificial Dutch and German have a higher proportion of wordforms with at least one homophone (shown by the violin plots) than their real counterparts (shown by the orange dots). However, the artificial French, Japanese, and English artificial lexica have lower Homophony Rates than the real lexica. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

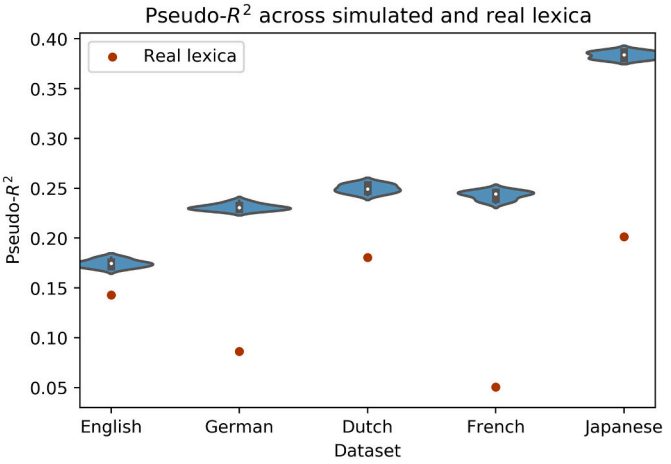


**Fig. 5.** We built a series of Poisson regression models predicting #Homophones from #Syllables and Normalized Surprisal. In each language, the models constructed for the artificial lexica (shown by violin plots) exhibit better model fit (larger pseudo-$R^2$) than the models constructed for the real lexicon (shown by orange dots). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

0.11 (M = 0.11, SD = 0.0004). On the other hand, the Homophony Rate in the real English lexicon (0.156) was significantly higher than the rate the simulated lexica (M = 0.143, SD = 0.002); similarly, the Homophony Rate for the real French and Japanese lexica were significantly higher than that for the artificial lexica.

Together, these results suggest that the **amount** of homophony in the five real lexica is not the result of a direct pressure for ambiguity. In fact, the real lexica actually display *less* homophony than the artificial ones in some measures, particularly the upper-bound of homophones tolerated for a given wordform and the mean number of homophones per wordform. This means that merely the pressure for highly probable phonotactic sequences, combined with the observed distribution of word lengths, can produce concentrations of homophony in a lexicon that are as dense or denser than in real lexica, without a direct pressure to recycle entire wordforms.

### 2.3.3. Simulated lexica exhibit more efficient reuse of optimal wordforms

We then asked whether homophones were more concentrated in optimal regions of phonotactic space in the simulated lexica or their real counterparts. That is, to what extent do the phonotactics of a language, as well as its distribution of word lengths, account for the finding that more optimal wordforms tend to have more homophones?

In order to assess the degree to which homophony was optimally distributed in a lexicon, we regressed a wordform's #Homophones against two operationalizations of wordform optimality: its length (#Syllables) and its phonotactic plausibility (Normalized Surprisal). For each lexicon, we extracted the following information from the model: 1) pseudo-$R^2$, as a measure of overall model fit; 2) the coefficient for #Syllables; and 3) the coefficient for Normalized Surprisal. A larger, more positive value for (1) reflects more efficient reuse overall, and more negative values for (2) and (3) reflect more efficient reuse along those particular dimensions of wordform optimality. Then, for each language, we compared each of these test statistics from the real lexicon to the distribution of test statistics obtained from the corresponding simulated lexica. The significance for each of these comparisons was assessed in the same way as above. All of the comparisons described revealed significant difference. To preview the overall finding, in all cases, the simulated lexica exhibited *stronger* effects (i.e., more optimally distributed wordforms) than their real counterparts.

Across all five languages, the distribution of pseudo-$R^2$ values obtained from the simulated lexica was significantly higher than the pseudo-$R^2$ value from the real lexicon (see Fig. 5 below). Pseudo-$R^2$ reflects a model's goodness-of-fit, i.e., how well the predictors in a model explain variance in the dependent variable. Thus, this indicates that two operationalizations of wordform optimality—its length, and its phonotactic plausibility—were better predictors of homophony across all of the simulated lexica than their real counterparts, for each language. For example, the pseudo-$R^2$ for the model constructed on the real English lexicon was 0.143, while the mean for the simulated lexica was 0.17 (SD = 0.004). Some differences were even starker: the pseudo-$R^2$ for the real German lexicon was 0.09, while the distribution of pseudo-$R^2$ values for the simulated German lexica averaged more than twice that (M = 0.231, SD = 0.003). Concretely, this means that homophony is better predicted by wordform optimality in the artificial than real lexica.

Further evidence comes from direct comparison of the coefficients for both predictors (Number of Syllables and Surprisal) across the real and artificial lexica. As reported earlier, the real lexica all exhibited negative relationships between Number of Syllables and Number of Homophones—i.e., short wordforms have more homophones in all five languages. However, the simulated lexica exhibited significantly stronger relationships, as depicted in Fig. 6 below.

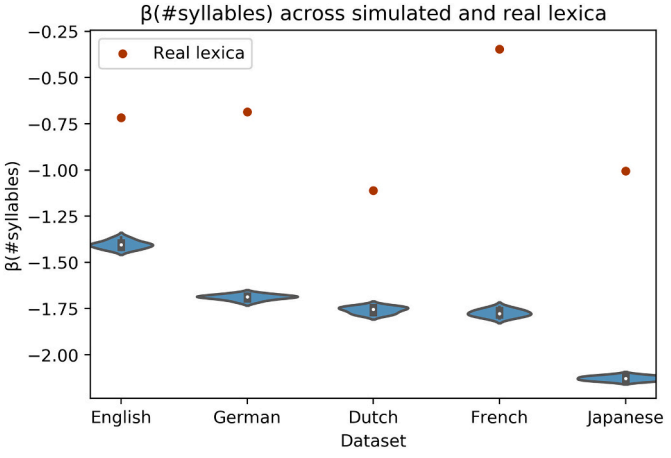


**Fig. 6.** Word length (as measured in #Syllables) is a better predictor of homophony in the artificial lexica (shown by the violin plots) than the real lexica (shown by orange dots). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

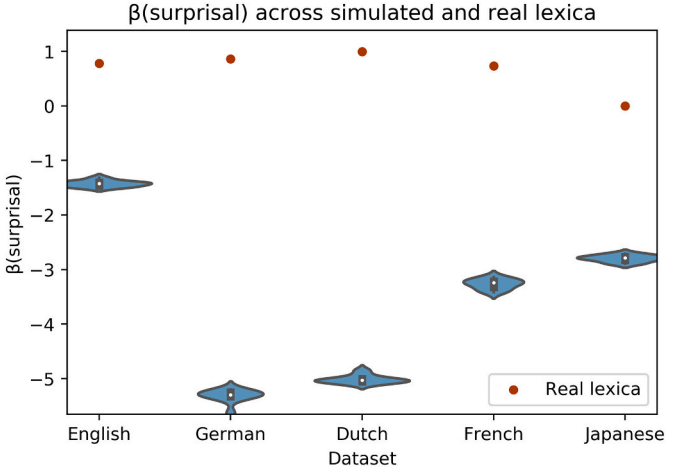


**Fig. 7.** Phonotactic Surprisal was more negatively correlated with Number of Homophones (i.e., more probable wordforms had comparatively more homophones) in the artificial lexica (shown by violin plots) than real lexica (shown by orange dots). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For example, the coefficient for Number of Syllables in the real English lexicon was −0.717, but the coefficients for the simulated English lexica were approximately twice as large (M = −1.4, SD = 0.02). In some cases, the difference was even larger, as in French: here, the coefficients for the simulated lexica (M = −1.77, SD = 0.02) were approximately five times as large as the coefficient for the real lexicon (−0.35).

Even more striking results were obtained for Surprisal: the real lexica actually exhibited positive relationships between Surprisal and Number of Homophones, while the artificial lexica all demonstrated negative relationships (see Fig. 7); these differences were significant for each language. In other words, the artificial lexica reused short, phonotactically plausible wordforms to a greater extent than did their real counterparts.

## 3. General discussion

In the current work, we asked whether the prevalence of homophony across five languages—English, German, Dutch, French, and Japanese—could be plausibly attributed to a direct pressure to recycle optimal wordforms. We reasoned that even without a direct pressure *for* ambiguity, an absence of a pressure *against* ambiguity should result in some amount of homophony in a lexicon, simply as a result of a language's phonotactic constraints and the distribution of words across different lengths. Under this view, the selective pressure is for well-formed phonotactic sequences as opposed to entire wordforms; the pressure to use well-formed sequences could result in homophony, particularly for the most phonotactically probable wordforms. Furthermore, given that the proportion of occupied phonotactic space will always be highest for short wordforms, homophony should also be most likely to occur in short words.

We tested this view by simulating a series of artificial lexica for each of the five languages. Across all five languages, we found that wordforms in the real lexica had either fewer or an equivalent number of homophones on average as wordforms in the artificial lexica (in every language but Japanese, wordforms in the artificial lexica were *more* ambiguous on average than those in the real lexica (see Fig. 2)). The real lexica also uniformly exhibited lower upper-bounds on the number of homophones tolerated per wordform (see Fig. 3). This was true despite considerable cross-linguistic variability in the propensity towards homophony overall (e.g., Japanese vs. Dutch); in each language, the artificial lexica surpassed their real counterparts in terms of the degree to which a wordform could be saturated with many meanings. The main exception to this trend was Homophony Rate (the proportion of wordforms with at least one homophone): for English, French, and Japanese (but not German and Dutch), the real lexica had higher Homophony Rates than the artificial lexica. This will be discussed in more detail below. Finally, statistical analyses of *where* these homophones were distributed revealed that homophones in the real lexica were concentrated less efficiently in "optimal" regions of phonotactic space: across all languages, word length and phonotactic plausibility—taken as operationalizations of wordform optimality—were better predictors of homophony in the artificial lexica than the real lexica (see Figs. 5–7).

There are two conclusions to be drawn from these results. First, neither the amount of homophony in these five real languages, nor the apparent concentration of homophones among optimal regions of phonotactic space, requires explanation by a direct pressure to recycle entire wordforms. Rather, homophony appears to be a natural and perhaps inevitable consequence of other features of a language—i.e., its phonotactics and distribution of word lengths. Of course, these features may themselves be related to efficiency, as noted in the Introduction—but indirectly so.

Second, real lexica may actually be the product of a pressure *against* homophony. The artificial lexica were modeled using only two parameters: the phonotactics of the target lexicon and a particular distribution of word lengths. They were not designed to explicitly select *for* homophony, nor did they contain a parameter selecting *against* homophony. In other words, they reflect the consequence of allowing the phonotactics of a language to determine its space of realized wordforms, under the assumption that the speakers of that language place no upper limit on how many homophones are tolerated per wordform. This resulted in considerably more homophones per wordform than observed in real languages. For example, wordforms in the real Dutch lexicon had at most 5 homophones, whereas the average upper-bound in the Dutch lexica was 97—more than 16 times as high. Furthermore, homophony in the artificial lexica was more likely to be found among more optimal wordforms.

One explanation for this result is that real lexica are subject to a pressure against *oversaturating* the same wordform with too many unrelated meanings—no matter how "optimal" it is. Clearly this pressure is not absolute: homophony does still exist (to varying degrees) in real languages—and in fact, some languages (French, English, and Japanese) had a higher proportion of wordforms with *at least one* homophone than their artificial counterparts. This suggests that the pressure is not against the existence of homophony per se, but rather, could reflect a constraint on the extent to which any given wordform can be saturated with distinct, unrelated meanings. Assigning too many unrelated meanings to the same signal could impede communication or learning (Casenhiser, 2005; though see Dautriche et al., 2018), and may thus be selected against. Such a pressure against oversaturation is roughly analogous to what others have termed *diversification* (Zipf, 1949) or a pressure for *clarity* (Piantadosi et al., 2012). However, unlike Zipf (1949), we find no opposing pressure towards *unification*; instead, homophony appears to emerge naturally as a function of other pressures (e.g., phonotactics), and is attenuated in particular wordforms (i.e., it does not reach the potential predicted by that wordform's phonotactics) due to a pressure against oversaturation.

There are a number of explanations for how this direct or indirect pressure against over-saturation might come about. For example, the attenuation of homophony could manifest as a kind of *smoothing* of high-probability phoneme sequences across phonological neighborhoods as opposed to being concentrated in a specific wordform. (A wordform's neighborhood is the set of wordforms differing from it in only one phoneme.) This could satisfy the pressure to reuse well-formed phonotactic sequences while also avoiding potential impediments to communication caused by overloading the same high-probability wordform with too many meanings.

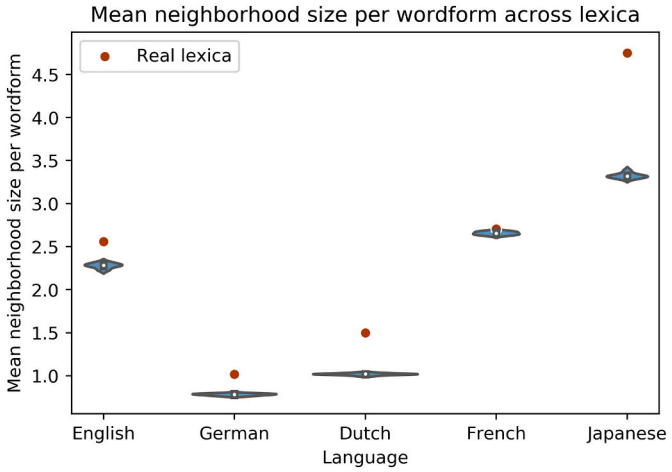


**Fig. 8.** Consistent with previous work (Dautriche et al., 2017), wordforms in the real lexica (shown by orange dots) have larger lexical neighborhoods (i.e., the set of words differing in exactly one phoneme) on average than wordforms in the artificial lexica (shown by violin plots). Note that this is true even in French, where the values are closest: wordforms in the real French lexicon have 2.71 neighbors on average, whereas wordforms in the artificial lexica have approximately 2.66 neighbors on average (SD = 0.02). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This account leads to testable predictions. If real lexica are subject to this smoothing process, they should have larger phonological neighborhoods than the artificial lexica, which were placed under no pressure against ambiguity. Indeed, previous work using an identical generative model (Dautriche et al., 2017) found exactly this: across four languages (English, German, Dutch, and French), real lexica exhibit more "clumpiness" (i.e., larger and more densely connected neighborhoods) than ought to be expected merely as a function of those languages' phonotactics. We extended a subset of their analyses to the set of artificial lexica we constructed, counting as "neighbors" any two wordforms that could be converted into each other via one phoneme substitution, deletion, or insertion (Dell & Gordon, 2003; Luce & Pisoni, 1998; Vitevitch & Luce, 1999). Under this definition of neighbor, the neighbors of the word *cat* would include *rat* (substitution), *at* (deletion), and *cast* (insertion). Consistent with prior work, and despite a different operationalization of neighborhoods from Dautriche et al. (2017), we found that wordforms in the real lexica had larger average neighborhood sizes than wordforms in the artificial lexica (see Fig. 8 below). For example, wordforms in the real English lexica averaged 2.56 neighbors, whereas the mean neighborhood sizes in the artificial English lexica ranged from 2.23 to 2.32 (M = 2.28, SD = 0.03). This result is the inverse of our finding regarding homophony—wordforms in the artificial lexica have *more* homophones on average than wordforms in the real lexica. In other words, the artificial lexica appear to optimize for dense concentrations of homophony, while the real lexica appear to optimize for larger neighborhoods. This apparent trade-off can also be illustrated by comparing both the rank-distribution of homophone counts and rank-distribution of neighborhood sizes across the real and artificial lexica (see Fig. 9).

Taken together, these findings are broadly consistent with the hypothesis that real lexica could be subject to a pressure against oversaturating the same wordform with too many meanings, and this selection against homophony could instead result in the creation of lexical neighbors. Of course, a similar effect could be achieved not through selection against high levels of homophony but rather from a positive pressure towards large neighborhoods, i.e., a "clumpy" lexicon. As Dautriche et al. (2017) argue, dense lexical neighborhoods may have many beneficial consequences, e.g., for word learning (Coady & Aslin, 2003; Storkel, Armbrüster, & Hogan, 2006; though see also Swingley & Aslin, 2007) and word production (Vitevitch, 2002; Vitevitch & Sommers, 2003). It is impossible to know from the current work whether the disparity between the real and artificial lexica is due to a direct pressure in real lexica against oversaturation that results in dense neighborhoods, or a positive selection for dense neighborhoods that results in less homophony. Future work could explore this potential trade-off at both the psychological level of explanation (e.g., whether learners make errors when learning homophones that lead to the creation of near neighbors), and by simulating such pressures during the lexicon generation process (e.g., whether a direct pressure in favor of large neighbors reduces the number of homophones, or whether a direct pressure against over-saturation increases neighborhood size).

Homophones could conceivably be reduced in real lexica through other, more indirect mechanisms as well. Notably, many human languages have rich morphological structure, allowing them to flexibly combine existing morphemes to construct novel meanings. While the real lexica we analyzed excluded wordforms derived via inflectional morphology, they did not exclude derivational morphology (e.g., adding the suffix *-ify* to the adjective *humid* creates the verb *humidify*; adding the suffix *-ness* to the adjective *happy* creates the noun *happiness*). Morphological compositionality allows speakers to convey new meanings without coining entirely new wordforms—but it also avoids the need to reuse existing wordforms for new, unrelated meanings (i.e., homophony). Thus, compositionality represents an efficient mechanism for recycling existing lexical materials that also avoids outright ambiguity. Clearly, wordforms in the artificial lexica were not constructed via processes of morphological composition. Future work could also explore whether parameterizing these artificial lexica according to the morphology of the underlying real lexicon would decrease the overall homophony, and if so, how. (See Supplementary Analysis 4 for further exploration of the relationship between derivational morphology and homophony in real lexica.)

In addition to real lexica exhibiting a lower upper-bound on homophony overall, we found that their homophones were less optimally distributed—that is, homophones were much more concentrated among short, phonotactically likely wordforms in the artificial lexica than in their real counterparts. This result is surprising on its face: why do real lexica apparently prefer (at least relative to the phonotactic baselines) to distribute their homophones across less optimal regions of the lexicon? Even if real lexica select against over-saturation, intuition suggests that the homophones that *are* preserved should be concentrated among short, phonotactically likely wordforms. One possible explanation for this result is that the pressures that ordinarily select against homophony are reduced for longer wordforms—there are at least two accounts as to why this may be the case. The first account is that longer wordforms might be more contextually discriminable than short wordforms and are thus more likely to be preserved in the lexicon. If this is true, the distinct senses of homophonous wordforms should be better disambiguated by contextual cues (e.g., some representation of the linguistic context) for longer wordforms. The second account holds that because longer wordforms are comparatively less common than short wordforms, they require less frequent disambiguation. Even if longer wordforms are no more contextually discriminable than short wordforms, they are encountered less often. If a frequent need to disambiguate is one of the factors that selects against homophony—e.g., because disambiguation may incur processing costs, no matter how marginal—homophones should be relatively *more* likely to be preserved among infrequent wordforms than frequent ones. Note that this does not predict that short wordforms have less homophones overall; past work (Piantadosi et al., 2012) has shown empirically that this is not the case. Rather, the penalty against accruing multiple meanings will be proportionately less for longer, less frequent wordforms. Therefore, less frequent wordforms should experience less of a reduction in their *projected* homophony (relative to their phonotactics) than more frequent ones.

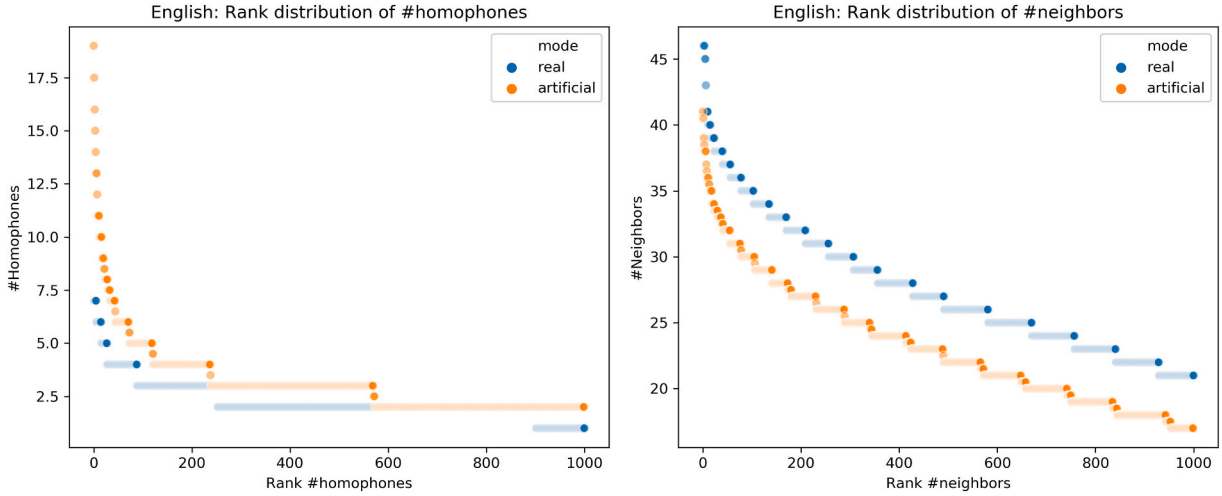


**Fig. 9.** Rank-distribution of homophone counts (left) and rank-distribution of neighborhood sizes (right) across the real and artificial English lexica. The most homophonous wordforms in the artificial lexica are more homophonous than equivalently ranked wordforms in the real lexica. Conversely, the wordforms with the largest neighborhoods in the artificial lexica still have smaller neighborhoods than equivalently ranked wordforms in the real lexica.

As noted above, the artificial lexica are intended as statistical baselines to determine which theoretical parameters are required to explain homophony, not as models of the many other pressures that real lexica are subject to. Thus, our work does not elucidate the developmental or historical mechanisms by which homophones arise, nor the processes by which they might be selected against or be preserved in a lexicon. There are a number of known sources of homophony in real lexica, including sound change and lexical borrowing (Ke, 2006; Ogura & Wang, 2006). Despite some debate about the extent to which homophony-generating sound changes are avoided (Sampson, 2013; Sampson, 2015; Wedel, Kaplan, & Jackson, 2013; Yin & White, 2018), there are many attested examples of phoneme losses and mergers resulting in homophony, such as *knight* and *night* in English, or as a consequence of the many phoneme mergers experienced in Middle Chinese (Ke, 2006; Sampson, 2013; Sampson, 2015). Similarly, lexical borrowing can lead to homophony; for example, the English words *sheik* and *chic* were both borrowed from different languages at different time points (16th century Arabic vs. 19th century French, respectively), and both have an identical phonological form (Ke, 2006). A satisfying explanation of homophony at a mechanistic level should incorporate these generative processes—i.e., the *mutations* by which potential homophones are introduced into a lexicon. Such a model should also predict which potential homophones will be selected against (and what form this selection process takes, i.e., whether it is via the avoidance of homophony-inducing mergers (Wedel et al., 2013; Yin & White, 2018) or something else) and which will be preserved. Homophones should be more likely to survive in a lexicon if their meanings are systematically made sufficiently discriminable by context (Dautriche et al., 2018). A better understanding of this process would also yield insights into which sources of contextual information human speakers and comprehenders routinely sample and deploy for disambiguation, and therefore influence language change.

We began by asking why a system that appears to be optimized for efficient communication (Gibson et al., 2019) contains apparently inefficient properties such as lexical ambiguity. A series of simulations suggests no evidence for a direct selection pressure in favor of homophones. Rather, the concentration of homophony among short, high-probability wordforms can be explained purely as a function of a language's phonotactics and distribution of word lengths, which perhaps themselves are the result of a pressure for efficiency. In fact, real lexica may even select *against* dense concentrations of homophony. We have suggested one mechanism: they might "smooth out" high-probability phonotactic sequences across lexical neighborhoods instead of concentrating these sequences in a single wordform. The product is lexica that are slightly less optimal in phonotactic terms but may better conform to other requirements of humans who need to use them.

## CRediT authorship contribution statement

**Sean Trott:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Benjamin Bergen:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision.

## Acknowledgments

We are very grateful to Tamara Rhodes for helping to secure access to the Japanese CallHome Lexicon. We also thank Steven Piantadosi for helping us understand the operationalization of surprisal in his model, Pamela Rivière for her help in concatenating the figures, Sarah Creel for advice on relevant literature about lexical neighborhoods in language acquisition, and Isabelle Dautriche for making the code for learning and generating phonotactic sequences available on GitHub. We are grateful to the anonymous reviewers for their advice both on the theoretical framing of the paper, as well as the construction and evaluation of the phonotactic models. Finally, we thank members of the Center for Research in Language and the Language and Cognition Lab at UC San Diego for their helpful comments on earlier versions of these analyses.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104449.

## References

Aina, L., Gulordava, K., & Boleda, G. (2019). Putting words in context: LSTM language models and lexical ambiguity. arXiv preprint arXiv:1906.05149.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (release 2)*. Distributed by the Linguistic Data Consortium: University of Pennsylvania.

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.

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology, 89*(3), 183–213.

Dautriche, I. (2015). *Weaving an ambiguous lexicon.* Sorbonne Paris Cité: Doctoral

dissertation.
Dautriche, I., Fibla, L., Fievet, A. C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology, 104*, 83–105.
Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition, 163*, 128–145.
Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes. *Phonetics and phonology in language comprehension and production: Differences and similarities, 6*, 9–37.
Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation, 49*, 209–246.
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.
Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. *Applied PsychoLinguistics, 12*(3), 349–367.
Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407.
Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience, 19*(7), 1175–1192.
Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture, 3*(2), 127–154.
Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*(5), 630.
Ke, J. (2006). A cross-linguistic quantitative study of homophony. *Journal of Quantitative Linguistics, 13*(01), 129–159.
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.
Kidd, E., & Holler, J. (2009). Children's use of gesture to resolve lexical ambiguity. *Developmental Science, 12*(6), 903–913.
Kobayashi, M., Crist, S., Kaneko, M., & McLemore, C. (1996). *CALLHOME Japanese Lexicon LDC96L17. Web Download*Philadelphia: Linguistic Data Consortium1996.
Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.
Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1.
Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science, 42*(8), 3116–3134.
Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research. 44*(4), 778–792.
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.
Ogura, M., & Wang, W. S. (2006). *Ambiguity and language evolution: Evolution of homophones and syllable number of words*.
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.
Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. *The 31st annual meeting of the Cognitive Science Society (CogSci09)* (pp. 2582–2587). .
Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America, 81*(5), 1574–1585.
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.
Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*(2), 245–266.
Sampson, G. (2013). A counterexample to homophony avoidance. *Diachronica, 30*(4), 579–591.
Sampson, G. (2015). A Chinese phonological enigma. *Journal of Chinese Linguistics, 43*(2), 679–691.
Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research. 49*(6), 1175–1992.
Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54*(2), 99–132.
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., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408.
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.
Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe,* 265–282.
Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*(2), 179–186.
Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 36, No. 36*.
Yin, S. H., & White, J. (2018). Neutralization and homophony avoidance in phonological learning. *Cognition, 179*, 89–101.
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.
Zipf, G. (1949). *Human behavior and the principle of least effort.* New York: Addison-Wesley.