Cognition 225 (2022) 105094

Contents lists available at ScienceDirect

Cognition

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

Languages are efficient, but for whom?

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:
Ambiguity
Efficiency
Language evolution
Frequency
Homophones

ABSTRACT

Human languages evolve to make communication more efficient. But efficiency creates trade-offs: what is efficient for *speakers* is not always efficient for *comprehenders*. How do languages balance these competing pressures? We focus on Zipf's meaning-frequency law, the observation that frequent wordforms have more meanings. On the one hand, this law could reflect a speaker-oriented pressure to reuse frequent wordforms. Yet human languages still maintain thousands of distinct wordforms, suggesting a countervailing, comprehender-oriented pressure. What balance of these pressures produces Zipf's meaning-frequency law? Using a neutral baseline, we find that frequent wordforms in real lexica have *fewer* homophones than predicted by their phonotactic structure: real lexica favor a comprehender-oriented pressure to reduce the cost of frequent disambiguation. These results help clarify the evolutionary drive for efficiency: human languages are subject to competing pressures for efficient communication, the relative magnitudes of which reveal how individual-level cognitive constraints shape languages over time.

1. Introduction

Languages adapt to the needs of the people who use them. In particular, there is increasing evidence that human languages have evolved in part to facilitate *efficient communication* (Piantadosi, Tily, & Gibson, 2009; Gibson et al., 2019; Mahowald, Dautriche, Gibson, & Piantadosi, 2018; Zaslavsky, Kemp, Regier, & Tishby, 2018; Regier, Carstensen, & Kemp, 2016; Kemp, Xu, & Regier, 2018). Pressure for efficiency has been used to explain various features of language, like how they carve up semantic domains among words (Conway, Ratnasingam, Jara-Ettinger, Futrell, & Gibson, 2020; Gibson et al., 2017; Kemp & Regier, 2012; Zaslavsky et al., 2018), as well which wordforms a lexicon contains (Mahowald et al., 2018; Meylan & Griffiths, 2017; Piantadosi, Tily, & Gibson, 2011). But efficiency involves trade-offs: features that make a language more efficient for *speakers* sometimes make it less efficient for *comprehenders*, and vice versa (Zipf, 1949). How do languages balance the interests of speakers and comprehenders when those interests are misaligned?

In the case of a language's grammatical rules, there is an emerging consensus that languages reflect a trade-off between reducing complexity (i.e., minimizing difficulties in production) and reducing ambiguity (i.e., minimizing difficulties in comprehension). The need to balance these pressures may explain cross-linguistic patterns in word order (Hahn, Jurafsky, & Futrell, 2020), person marking (Zaslavsky, Maldonado, & Culbertson, 2021), case marking (Mollica, Bacon, Xu, Regier, & Kemp, 2020), and more. Moreover, some theories argue that efficiency is best achieved by prioritizing the needs of speakers specifically (Levinson, 2000; MacDonald, 2013). Planning and producing utterances is cognitively expensive: speakers must ultimately translate the concepts they wish to convey into a series of complex motor commands, a process that involves selecting the correct lexical items and arranging them in an appropriate syntactic configuration (Ferreira, 2008; MacDonald, 2013). The architecture of the language production system is largely tuned towards reducing speaker effort (Ferreira, 2008), and as a consequence, the form of human languages themselves may also be oriented towards *producibility*, rather than *comprehensibility* (Levinson, 2000; MacDonald, 2013). On this view, comprehension is nevertheless possible because comprehenders have a sufficiently less taxing task than speakers (MacDonald, 2013; MacDonald, 2015), and rely on pragmatic inference to decipher under-specified or ambiguous utterances (Levinson, 2000). Of course, some grammatical features may also reflect a pressure for efficient comprehension, such as grammatical gender (Wasow, 2013; Dye, Milin, Futrell, & Ramscar, 2017; Dye, Milin, Futrell, & Ramscar, 2018). Similarly, the mere fact of grammatical regularity in the first place likely makes communication more robust to noise, which helps with both comprehension and production (Gibson et al., 2013).

There is substantively less consensus when it comes to the lexicon. Although many researchers agree that human lexica are shaped for

* 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).

https://doi.org/10.1016/j.cognition.2022.105094
Received 11 August 2021; Received in revised form 19 January 2022; Accepted 7 March 2022
Available online 24 March 2022

efficient communication (Piantadosi et al., 2009; Mahowald et al., 2018), it remains unclear whether they favor a pressure for efficient production or efficient comprehension, or whether they are shaped equally by both pressures. The paradigm example of these pressures in conflict is Zipf's *meaning-frequency* law (Piantadosi, Tily, & Gibson, 2012; Zipf, 1945), the empirical observation that more frequent words are more ambiguous.

On the one hand, this distribution could be interpreted as serving the speaker's needs. It is easier to produce frequent words than infrequent ones (Dell, 1990; Oldfield & Wingfield, 1965), so a lexicon that concentrates meanings among its most frequent wordforms would be more efficient for speakers than a lexicon that distributes its meanings more evenly across wordforms (Piantadosi et al., 2012; Zipf, 1949). Under this view, the meaning-frequency law reflects a pressure for efficient production, which Zipf (1945) termed *unification*. Taken to the extreme, this pressure—sometimes called *compressibility* (Kirby, Tamariz, Cornish, & Smith, 2015)—leads to a *degenerate* lexicon, "in which every meaning is associated with a single, shared, maximally ambiguous signal" (Kirby et al., 2015, pg. 88). A maximally degenerate lexicon is often taken as a speaker's ideal because it requires a speaker to remember and produce only a single word, and thus imposes minimal costs on speakers, i.e., it is minimally complex (Zaslavsky et al., 2018; Zipf, 1949).

On the other hand, real lexica are far from maximally degenerate. This is because lexica are also subject to a countervailing pressure, alternatively termed *diversification* (Zipf, 1945), *expressivity* (Kirby et al., 2015), or *informativity* (Zaslavsky et al., 2018), to reduce the burden of comprehension (Wasow, 2013; Zipf, 1945) and ensure clarity of communication (Piantadosi et al., 2012). An incomprehensible language is not particularly efficient—suggesting that the cost of disambiguation should in principle also shape the development of communicative systems. Oversaturating frequent wordforms with many meanings likely incurs costs for comprehenders: even if disambiguation is less costly than production (Levinson, 2000), it does appear to impose at least a marginal increase in processing difficulty (Blott, Rodd, Ferreira, & Warren, 2020; Rayner & Duffy, 1986; Rayner & Frazier, 1989). And if the most frequent wordforms are also the most ambiguous, then comprehenders will be required to disambiguate more often. But while real lexica do exhibit a relationship between frequency and ambiguity, their most frequent wordforms are not maximally ambiguous, i.e., this relationship is weaker than would be expected by a purely speaker-centric lexicon. Thus, under this view, the empirical relationship between ambiguity and frequency also reflects a pressure to reduce the burden on comprehenders.

Zipf's interpretation is that the empirical distribution of meanings across wordforms represents a *compromise* between these purported pressures (Zipf, 1945; Zipf, 1949). Yet identifying an equilibrium is only part of an explanation, as it leaves the relative magnitudes of the countervailing pressures indeterminate. It is possible that the pressures are equal in size, as Zipf (1945) suggests. But it could be that a speaker-oriented pressure has ultimately won out—that the equilibrium point is closer to the speaker's ideal than the comprehender's. This view of a Speaker-Oriented Pressure is similar to claims that grammar shows an equivalent bias (MacDonald, 2013). Alternatively, the lexicon may be driven primarily by a Comprehender-Oriented Pressure, biased towards reducing the cost of disambiguation.

Unfortunately, we cannot adjudicate between these competing accounts using the empirical distribution of word meanings alone. In the absence of a suitable baseline, it is impossible to determine whether Zipf's meaning-frequency law is attributable to a bias towards production or a bias towards comprehension, or even whether it can be explained without either such pressure (Caplan, Kodner, & Yang, 2020; Trott & Bergen, 2020). To date, the observed relationship between wordform frequency and ambiguity has only been compared with a baseline in which there is *no* relationship between wordform frequency and ambiguity (Piantadosi et al., 2012; Zipf, 1949). But such a baseline is indistinguishable from one version of a purely Comprehender-Oriented Pressure, in which meanings are distributed evenly across wordforms, and is thus inappropriate for adjudicating between Speaker-Oriented and Comprehender-Oriented Pressures. Instead, a baseline is required that establishes how many meanings those same wordforms should be expected to accrue just on the basis of other known factors. Previous work has established that even controlling for frequency, shorter and more phonotactically probable words have more meanings (Piantadosi et al., 2012). Using a baseline that incorporates these effects, we can then ask whether the positive empirical relationship between wordform frequency and ambiguity is larger (reflecting a Speaker-Oriented pressure) or smaller (reflecting a Comprehender-Oriented pressure) than what would be expected without either such pressure.

Here, a conceptual parallel can be drawn to work in evolutionary biology; many traits that appear adaptive for a particular function may have emerged from other, more indirect selective pressures, or even genetic drift (Gould & Lewontin, 1979). This has led to the use of so-called "neutral" models (Alonso, Etienne, & McKane, 2006) to establish baselines of what to expect in the absence of selection pressures. More recently, neutral models have been applied to cultural evolution as well, to understand which aspects of language change are due to explicit selection and which are better explained by stochastic drift (Newberry, Ahern, Clark, & Plotkin, 2017). There is some controversy around the question of whether neutral models can be used to provide positive evidence of a causal mechanism (Leroi, Lambert, Rosindell, Zhang, & Kokkoris, 2020; Bentley, Carrignon, Ruck, Valverde, & O'Brien, 2021); however, there is general agreement that they are useful for establishing a "null" baseline, against which alternative theoretical models can be compared (Leroi et al., 2020).

Consonant with this line of reasoning, recent work has shown that when the observed distribution of homophony is compared with an appropriate baseline, other apparently efficient distributions of meanings show up in lexica without any explicit pressure for efficiency (Caplan et al., 2020; Trott & Bergen, 2020). Indeed, Trott and Bergen (2020) find that when compared against a suitable baseline that incorporates a lexicon's phonotactics and distribution of word lengths, real human lexica actually have *fewer* homophones than one would expect. Strikingly, this result is consistent with a Comprehender-Oriented Pressure, i.e., one in which homophones are avoided during the course of language change (Wedel, Jackson, & Kaplan, 2013; Wedel, Kaplan, & Jackson, 2013). Importantly, however, because this work used a simulated baseline (i.e., not using real words in the lexicon), it was unable to investigate whether the frequency of actual wordforms in a lexicon shaped a pressure for or against homophony. This leaves a gap in the literature: could a Comprehender-Oriented Pressure explain Zipf's meaning-frequency law as well?

The logic of our approach below is as follows. First, we establish a suitable baseline that characterizes the expected relationship between wordform frequency and ambiguity in the absence of either a direct production-oriented pressure or a comprehension-oriented pressure. The distribution obtained in this baseline is then compared to the attested distribution in real lexica. If the relationship between frequency and homophony is stronger in real lexica than in the baseline, it is consistent with production-oriented pressures shaping the language; in contrast, a weaker relationship in real lexica is consistent with the language being shaped by a comprehension-oriented pressure. Finally, if the real relationship between frequency and homophony is indistinguishable from the baseline, it suggests either that both pressures are equal in magnitude, or that neither pressure is required to explain how many meanings words of different frequencies have.

This hinges on first establishing a procedure for distributing meanings that is *neutral* with respect to whether it privileges a speaker-oriented pressure to accumulate meanings among frequent wordforms, or a comprehender-oriented pressure to reduce ambiguity among those wordforms. That is, given M meanings and W wordforms, how ought those meanings to be distributed across wordforms in a neutral manner? One candidate for such a neutral procedure is to assign meanings to

wordforms according to their phonotactic probability. Although all wordforms of a language must obey the phonotactic rules of that language—i.e., which sounds can begin and end a word, which sounds can occur in which sequence, and so on—some phonological sequences are nonetheless more common across wordforms than others. Wordforms containing very common phonological sequences are considered to have a higher *phonotactic probability* (Vitevitch & Aljasser, 2021). Critically, phonotactic probability appears to facilitate word production (Goldrick & Larson, 2008; Vitevitch, Armbrüster, & Chu, 2004), word recognition and processing (Vitevitch & Luce, 1999; Vitevitch, Luce, Pisoni, & Auer, 1999), and word learning (Jusczyk, Luce, & Charles-Luce, 1994; Munson, 2001; Coady & Aslin, 2004; Storkel, 2001). To our knowledge, there is no evidence that phonotactic probability disproportionately benefits speakers over listeners, or vice versa. Thus, it is reasonable to expect that both speakers and listeners would prefer a lexicon that privileged phonotactically probable wordforms, as opposed to phonotactically improbable ones. (Of course, according to Zipf (1949), speakers might prefer that every meaning is conveyed by a single, high-probability wordform—while listeners might prefer no ambiguity at all. However, the goal of this baseline is not to implement the ideal speaker-oriented or listener-lexicon—it is to construct a lexicon according to neutral principles.)

A second, related reason to distribute meanings according to the phonotactic probability of wordforms is that in real lexica, homophones are disproportionately concentrated among phonotactically probable wordforms (Piantadosi et al., 2012; Trott & Bergen, 2020). This lends further plausibility to the approach being taken: empirically, meanings are attracted to high-probability regions of phonotactic space.

Finally, phonotactic probability correlates with frequency across a number of languages (Bentz & Ferrer Cancho, 2016; Mahowald et al., 2018; Meylan & Griffiths, 2017). While this is not itself a reason to adopt this baseline, it does tell us a priori that even in the absence of a frequency bias, a preferential distribution of meanings according to phonotactic probability would produce a positive correlation between frequency and ambiguity. Importantly, this baseline correlation with frequency would be epiphenomenal in the sense that it emerged from other principles of lexicon design. The central question of the current work is whether the correlation between frequency and ambiguity in the baseline is *weaker* than the one observed in real lexica (implying a speaker-oriented pressure), or *stronger* than the one observed in real lexica (implying a comprehender-oriented pressure).

2. Current work

Using a neutral baseline, we calculated the Homophony Delta for each wordform in the real lexicon: the difference between how many homophones a wordform *actually* has, and how many homophones it would be *expected* to have, assuming that meanings distributed purely according to phonotactic probability. We then asked whether the relationship between Homophony Delta and Frequency was positive (as predicted by a speaker-centric pressure) or negative (as predicted by a comprehender-centric pressure).

To calculate the expected number of homophones, we first calculated the phonotactic probability of each wordform using an n-phone model.[1] We then multiplied each wordform's phonotactic probability by the number of meanings for words of that length (see the Methods section below for more details on how the number of meanings was calculated). This ensured that the distribution of meanings across word lengths was matched across each of the real lexica and their neutral baselines; for example, if the real English lexicon has 7706 meanings distributed among its monosyllabic wordforms, the English baseline would do the same. Finally, we subtracted a wordform's expected number of homophones from the number of homophones a wordform actually has. A positive value of Homophony Delta indicates that a wordform has *more* homophones than expected, and a negative value indicates that it has *fewer*. We repeated this process across six target languages: English, Dutch, German, French, Japanese, and Mandarin.

The accounts outlined above make opposing predictions about the relationship between Frequency and Homophony Delta. Given that more frequent wordforms are easier and faster to produce (Dell, 1990; Oldfield & Wingfield, 1965), a pressure to minimize speaker effort should result in frequent wordforms acquiring more meanings than their phonotactics would predict. Thus, Frequency should exhibit a *positive* relationship with Homophony Delta. On the other hand, concentrating meanings in the most frequent wordforms results in a language requiring more frequent disambiguation by comprehenders. Such a lexicon would impose a larger average disambiguation cost than one that distributed its meanings more evenly across wordforms. Thus, a pressure to minimize comprehender effort predicts a *negative* relationship between Frequency and Homophony Delta. Finally, it is possible that these pressures are roughly equal in size, or even that neither pressure plays a role at all—i.e., that phonotactic plausibility and length is the sole determinant of homophony. In both cases, the relationship between Frequency and Homophony Delta should be statistically indistinguishable from zero.

All data and code necessary to reproduce the analyses described here can be found on GitHub: https://github.com/seantrott/homophony_delta.

3. Methods

3.1. Materials

We analyzed lexica from six languages: English, Dutch, German, French, Japanese, and Mandarin Chinese. Importantly, we restricted our analysis to the unique *lemmas* of each language. This means that inflectional variants (e.g., "dogs") would not be included as distinct entries, whereas distinct meanings of the same wordform (e.g., *water.n* and *water.v*) would be listed separately, with separate frequency estimates for each lemma. For determining which meanings counted as distinct lemmas, as well as the frequencies of those lemmas, we relied on lexical resources for each language.

For English, Dutch, and German, we used the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). For French, we used the French Lexique (New, Pallier, Brysbaert, & Ferrand, 2004). For Japanese, we used the Japanese CallHome Lexicon (Kobayashi, Crist, Kaneko, & McLemore, 1996). For Mandarin Chinese, we used the Chinese Lexical Database (Sun, Hendrix, Ma, & Baayen, 2018); we also conducted the same analysis (and obtained qualitatively identical results) using the Mandarin CallHome Lexicon (Huang, Bian, Wu, & McLemore, 1996), which is included in the Supplementary Materials. We removed wordforms containing hyphens, spaces, or apostrophes, as well as proper nouns (in the case of the Mandarin Chinese lexica). The number of unique wordforms (i.e., after collapsing across distinct entries) in each lexicon was as follows: 35,107 English wordforms, 50,435 German wordforms, 65,260 Dutch wordforms, 37,278 French wordforms, 40,449 Japanese wordforms, and 41,009 Mandarin Chinese wordforms (with 45,871 in the Mandarin CallHome lexicon).

Frequency estimates for English, Dutch, and German were taken from CELEX; respectively, these frequency estimates were in turn based on the COBUILD (approximately 18 million words), INL (approximately 40 million words), and Mannheim (approximately 5 million words) corpora (Kruyt & Dutilh, 1997; Kupietz & Keibel, 2009; Sinclair, 1987). Note that we also replicated the analyses described here using the SUBLTEX estimates of word frequency, and obtained qualitatively identical results (i.e., a negative relationship between Log Frequency and Homophony Delta; see *Supplementary Analysis 4*) for a description of those results. The lexica for French and Mandarin Chinese already contained by-lemma frequency estimates. The corpus sizes from which

[1] See *Supplementary Analysis 5* for a replication of the primary results using a measure of phonotactic probability calculated using an LSTM.

these estimates were obtained were, approximately: 14.8 M (for French) and 120 M (for Mandarin). The frequency estimates for Japanese wordforms were taken from the Japanese CallHome Lexicon, with a total of approximately 690 K tokens. In each language, if by-lemma frequency estimates were available for a given wordform, we summed these estimates to calculate the total frequency of that wordform. Note that the Japanese lexicon did not contain reliable by-lemma frequency measures—thus, for Japanese, we used the *mean* frequency for each lemma corresponding to a given wordform. However, the results reported below are qualitatively identical using the sum of lemma frequencies. Additionally, because we would eventually calculate the log of each frequency, we incremented each frequency value by 1, to ensure that no wordforms had a frequency of 0. Additionally, for the French lexicon specifically, frequency values were multiplied by 14.8 (given that Lexique normalized the book frequency estimates to 14.8).

Finally, the frequency estimates reflect a mixture of spoken and written text, depending on the language. The English COBUILD corpus consists primarily of written language (approximately 5% is spoken), as do the Dutch INL (approximately 9% is spoken) and German Mannheim (0% is spoken). The Chinese Lexical Database frequency estimates combine two written sources: the Leiden Weibo Corpus (van Esch, 2012) and the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). For French, we relied on frequency estimates from a corpus of written books (New et al., 2004). Finally, frequency estimates for the Japanese CallHome Lexicon are based solely on spontaneous spoken speech (Kobayashi et al., 1996).

3.2. Calculating phonotactic probability

For each lexicon, we built an *n*-phone Markov Model that approximated the phonotactics of the target language. We adapted the code and procedure used in previous work (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017; Trott & Bergen, 2020).

Given some value of *n* (e.g., 2), an *n*-phone model can use the set of wordforms[2] in a lexicon to learn which phoneme characters occur in which positions and in which sequence; for example, in English, such a model would learn that the sequence *bn*– never occurs at the start of a wordform. Such a model can then be used to compute the probability of an entire wordform, which is defined as the product of all the transitional probabilities between each phoneme in that wordform (including the START and END symbols). We identified the appropriate chain length (i.e., value of *n*) for each language using a cross-validation procedure—the optimal *n* was defined as the model that, when trained on a set of real wordforms (e.g., 75% of a lexicon), maximizes the probability of held-out wordforms (e.g., the remaining 25%). This cross-validation procedure was identical to the one described in Trott and Bergen (2020), and determined the optimal models to be 5-phone models for English, Dutch, and German, and 4-phone models for Japanese, French, and Mandarin Chinese.

We then calculated the phonotactic probability of each wordform in each lexicon using 1000-fold cross-validation. We divided each lexicon into 1000 "folds" (each containing roughly 0.1% of the entire set of wordforms). Then, for each fold, we trained an *n*-phone model on the remaining 99.9% of the lexicon, and evaluated the phonotactic probability of the wordforms in the target fold. This allowed us to produce estimates of phonotactic plausibility from a model that never directly observed the wordforms in question—only other wordforms resembling them to varying degrees. As in past work (Dautriche et al., 2017; Trott & Bergen, 2020), we also assigned non-zero probability to unobserved phoneme sequences using Laplace smoothing with the parameter set to 0.01.

Finally, we used these probabilities to calculate the phonotactic surprisal of each wordform, which is defined as the negative log probability (note that we used log_{10})—i.e., less probable phonotactic sequences will have higher phonotactic surprisal. Because phonotactic surprisal is correlated with length, we divided surprisal by the number of phonemes in the wordform to obtain a Normalized Phonotactic Surprisal measure, as in Piantadosi et al. (2012).

Note that recent work (Pimentel, Meister, Teufel, & Cotterell, 2021) has found that an LSTM provides a better measures of phonotactic probability, and is less prone to overfitting, than an n-gram model. We have replicated the primary results described below using an LSTM with qualitatively identical results; see *Supplementary Analysis 5* for more details.

3.3. Calculating actual number of homophones

Following past work (Piantadosi et al., 2012; Trott & Bergen, 2020), we calculated the Actual Number of Homophones for a given wordform, $A(w_i)$, by identifying the number of distinct lexical entries with the same phonological form, then subtracting one. Note that this measure would include both homographic (e.g., "baseball *bat*" vs. "furry *bat*") and heterographic (e.g., "juicy *steak*" vs. "wooden *stake*") homophones. In the latter case, the wordform */steɪk/* has three entries, so the Actual Number of Homophones is two.

3.4. Estimating expected number of homophones

To estimate a wordform's Expected Number of Homophones, we calculated the number of meanings each wordform should be assigned if meanings were assigned purely on the basis of phonotactic plausibility alone. We also sought to control for word length, so the procedure described below was performed separately for words of varying lengths (e.g., 1-syllable, 2-syllable, etc.).

First, we normalized a wordform's phonotactic probability, p_i, to the number of meanings, *M*, distributed among wordforms of that length. To do this, we calculated the sum of those wordforms' probabilities—typically much less than 1, depending on the smoothing parameter and number of wordforms in question—then divided each probability p_i by that sum. This produced a set of normalized wordform probabilities such that they summed to 1, which ensured that the sum of expected number of meanings (*M'*) distributed among some set of wordforms would equal the actual number of meanings (*M*). After this normalization procedure, the monosyllabic wordform */steɪk/* ends up with a normalized probability of 0.0009.

Then, for each wordform, we multiplied its normalized probability by *M*, the number of meanings available for wordforms of that length. This yielded the expected number of meanings. For example, the normalized probability for the wordform */steɪk/* (0.0009) would be multiplied by the number of meanings available for monosyllabic wordforms (7706), yielding the expected number of meanings (approximately 6.94).

Finally, to calculate the Expected Number of Homophones, we simply subtracted one from the expected number of meanings (as in the real lexicon); the wordform */steɪk/* would thus have approximately 5.94 homophones. This is illustrated in the equation below, where w_i refers to a given wordform, M_i refers to the number of meanings expressed by wordforms of that length, p_i refers to the normalized probability of that wordform and $E(w_i)$ refers to the Expected Number of Homophones.

$$\mathrm{E}(w_i) = M_i{}^{*}p_i - 1$$

Note that unlike in the real lexicon, this procedure can yield non-integer values for a wordform's Expected Number of Homophones; occasionally these values are even negative, if the expected number of meanings is below 1. We chose not to "correct" these values (i.e., round them to the nearest integer), because doing so would no longer ensure equivalence between the *actual* and *expected* number of meanings distributed among wordforms of a certain length. Because our primary

[2] Note that these models were trained using the set of unique *types* (individual wordforms), rather than *tokens* (actual instances of each wordform in a text corpus), to avoid conflating phonotactic probability with frequency.

interest is in the relative differences between expected and actual numbers of meanings, the absolute value of Expected Number of Homophones should not impact the interpretation of results. (See *Supplementary Analysis 6* for an alternative approach ensuring that wordforms are assigned an integer number of meanings.)

3.5. Calculating homophony delta

Homophony Delta, i.e., $HD(w_i)$, was defined as the difference between a wordform's Actual Number of Homophones, i.e., $A(w_i)$, and that wordform's Expected Number of Homophones, i.e., $E(w_i)$:

$$HD(w_i) = A(w_i) - E(w_i)$$

We subtracted the latter estimate (described above) from the former, obtained from the real lexica. Thus, a negative value means that wordform has *fewer* homophones than predicted by its phonotactics, while a positive value means that a wordform has *more* homophones than predicted by its phonotactics. For the wordform */steɪk/*, the Actual Number of Homophones is 2, while the Expected Number of Homophones is 5.94, so the Homophony Delta would be −3.94. Put another way: the wordform */steɪk/* has approximately 3.94 fewer homophones than predicted by its phonotactics.

4. Results

4.1. Homophony and frequency

For each language, we constructed a linear regression model with Homophony Delta as the dependent variable, and Log Frequency, Number of Syllables, and Normalized Phonotactic Surprisal[3] as predictors. We were primarily interested in the effect of Log Frequency, which we focus on below; given that frequency is correlated with word length and phonotactic probability, we included Number of Syllables and Normalized Phonotactic Surprisal as covariates to identify and isolate the variance explained by Frequency specifically.[4] All analyses were performed in R version 3.6.3 (R Core Team, 2020).

Log Frequency exhibited a significant, negative relationship with Homophony Delta across all six languages: English [$\beta = -0.49$, $SE = 0.07$, $p < .001$], Dutch [$\beta = -1.85$, $SE = 0.07$, $p < .001$], German [$\beta = -1.28$, $SE = 0.1$, $p < .001$], French [$\beta = -0.45$, $SE = 0.05$, $p < .001$], Japanese [$\beta = -1.73$, $SE = 0.11$, $p < .001$], and Mandarin Chinese [$\beta = -0.28$, $SE = 0.02$, $p < .001$]. The magnitude of this relationship, and the absolute values of Homophony Delta, varied considerably across languages; for example, the most frequent wordforms in Dutch have much larger negative values of Homophony Delta than the most frequent wordforms in French or Japanese. Crucially, however, the overall relationship was negative in each of the languages we considered: frequent wordforms consistently have fewer homophones than predicted by their phonotactics. Because we modeled frequency as logarithmic, these coefficients can be interpreted as representing the expected reduction in homophony (relative to a wordform's phonotactics), given each order of magnitude increase in frequency. For example, in English, the coefficient estimate for Frequency is *−0.49*; this means that an increase in frequency from 10 to 100 would predict a *0.49* decrease in how many homophones a given wordform has, relative to its phonotactics.

This is best illustrated by Fig. 1, which directly compares the actual and expected number of homophones for each of 20 frequency bins. Across all languages, frequent wordforms have fewer homophones in actuality than expected. That is, although each language exhibits the well-attested, positive relationship between wordform frequency and ambiguity[5]—i.e., Zipf's *meaning-frequency law* (Zipf, 1945)—this relationship is considerably weaker than one would expect if meanings were assigned purely on the basis of phonotactic probability and length.

In addition to the negative relationship between Log Frequency and Homophony Delta, Normalized Phonotactic Surprisal exhibited a significant, positive correlation with Homophony Delta across all six languages: English [$\beta = 3.58$, $SE = 0.13$, $p < .001$], Dutch [$\beta = 4.04$, $SE = 0.17$, $p < .001$], German [$\beta = 3.42$, $SE = 0.18$, $p < .001$] French [$\beta = 3.18$, $SE = 0.11$, $p < .001$], Japanese [$\beta = 2.92$, $SE = 0.09$, $p < .001$], and Mandarin Chinese [$\beta = 2.71$, $SE = 0.07$, $p < .001$]. The most phonotactically plausible wordforms in real lexica have *fewer* homophones than predicted by their phonotactics alone. This is not surprising, given that our baselines assumed that phonotactic probability was the sole determinant of homophony—if the distribution of homophones in real lexica is influenced by any other factors, then the resulting relationship should be weaker than in our baselines.

More surprising is the observation that Number of Syllables was positively correlated with Homophony Delta across five of the six languages (all but Mandarin Chinese): English [$\beta = 0.73$, $SE = 0.07$, $p < .001$], Dutch [$\beta = 0.4$, $SE = 0.07$, $p < .001$], German [$\beta = 0.36$, $SE = 0.07$, $p < .001$], French [$\beta = 0.56$, $SE = 0.05$, $p < .001$], and Japanese [$\beta = 0.13$, $SE = 0.02$, $p < .001$]. In other words, short wordforms in these languages were less ambiguous than expected, given their phonotactics. The coefficient in Mandarin was not significant after correcting for multiple comparisons ($p > .1$). Across all languages, however, short wordforms were no *more* homophonous than one would expect (i.e., no language had a negative coefficient for Number of Syllables); this finding is consistent with past work (Caplan et al., 2020; Trott & Bergen, 2020) suggesting that the empirical relationship between length and homophony is not necessarily a product of a speaker-centric pressure to reuse short wordforms—indeed, in some languages, short wordforms have fewer homophones than one would otherwise expect. See Fig. 2 for the complete distribution of parameter estimates (and standard errors) across lexica.

4.2. Homophony and neighborhood size

If real lexica are indeed subject to a pressure against homophony in high frequency words, that pressure should have detectable consequences elsewhere in a language. We pursued this line of reasoning by focusing on the distribution of phonological neighborhood sizes in the real lexicon. Phonological neighbors are defined as two wordforms that can be converted into one another via a single edit, i.e., a substitution, deletion, or addition (Luce & Pisoni, 1998; Vitevitch & Luce, 1999). For example, under this definition, "pot" and "pit" would be neighbors, as would "bat" and "cat". Previous work (Dautriche et al., 2017; Trott & Bergen, 2020) has found that real languages have *larger* neighborhoods than artificial lexica matched for their phonotactics and distribution of word lengths, despite having a *smaller* number of homophones. Trott and Bergen (2020) argue that these results could arise from a pressure to avoid homophones, combined with a pressure to use high-probability phoneme sequences. Together, these pressures could create dense pockets of phonological neighborhoods in the place of a single, high-probability wordform over-saturated with meanings. If this interpretation is correct, then the wordforms most resistant to acquiring homophones should also have larger neighborhoods—i.e., controlling for

[3] Because Number of Syllables is correlated with Phonotactic Surprisal, we followed the procedure described in Piantadosi et al. (2012) and divided Phonotactic Surprisal by the number of phonemes in a wordform, which we called Normalized Phonotactic Surprisal.

[4] Note that Frequency is correlated with Number of Syllables, and the presence of collinearity between predictors can sometimes lead to suppression or enhancement of parameter estimates (Wurm & Fisicaro, 2014). To check for collinearity, we calculated the variance inflation factor (VIF) for the complete model for each language, and found that all VIF scores were below 1.5, which suggests that collinearity is not necessarily a concern in this case.

[5] See *Supplementary Analysis 3* for an analysis illustrating that Zipf's meaning-frequency law replicates across all six languages.

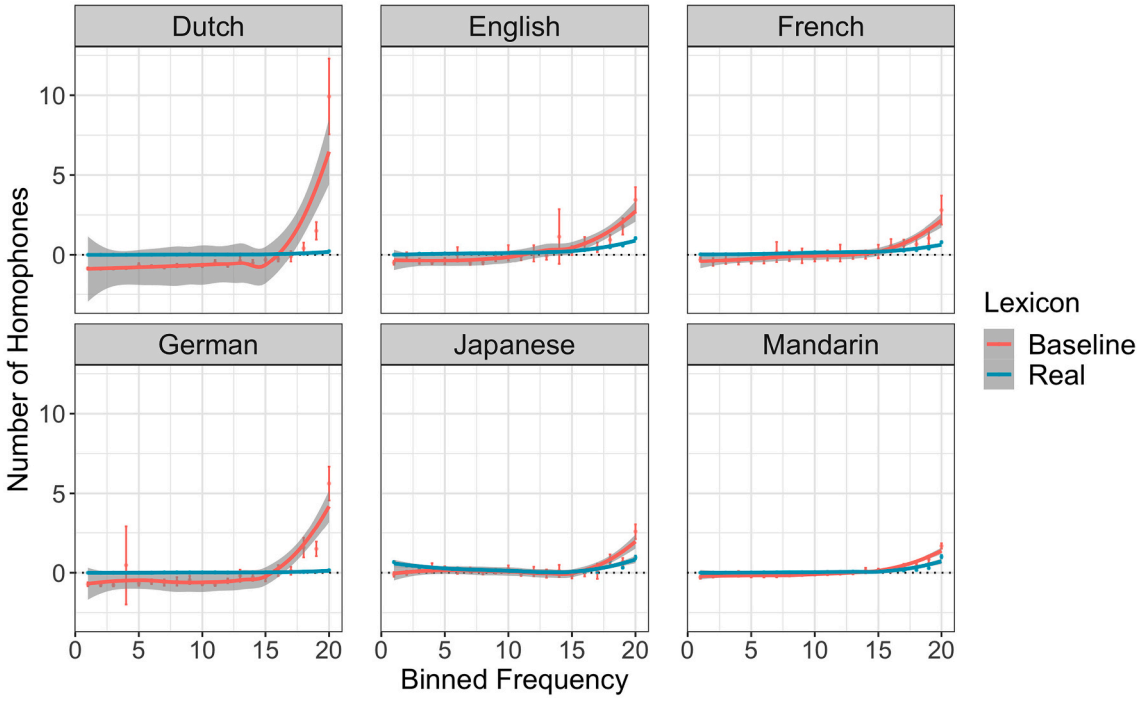

Fig. 1. Across all six languages, the most frequent wordforms have fewer homophones in actuality (Real) than predicted by their phonotactics (Baseline). Higher values of Binned Frequency correspond to more frequent words. Error bars are one standard error.

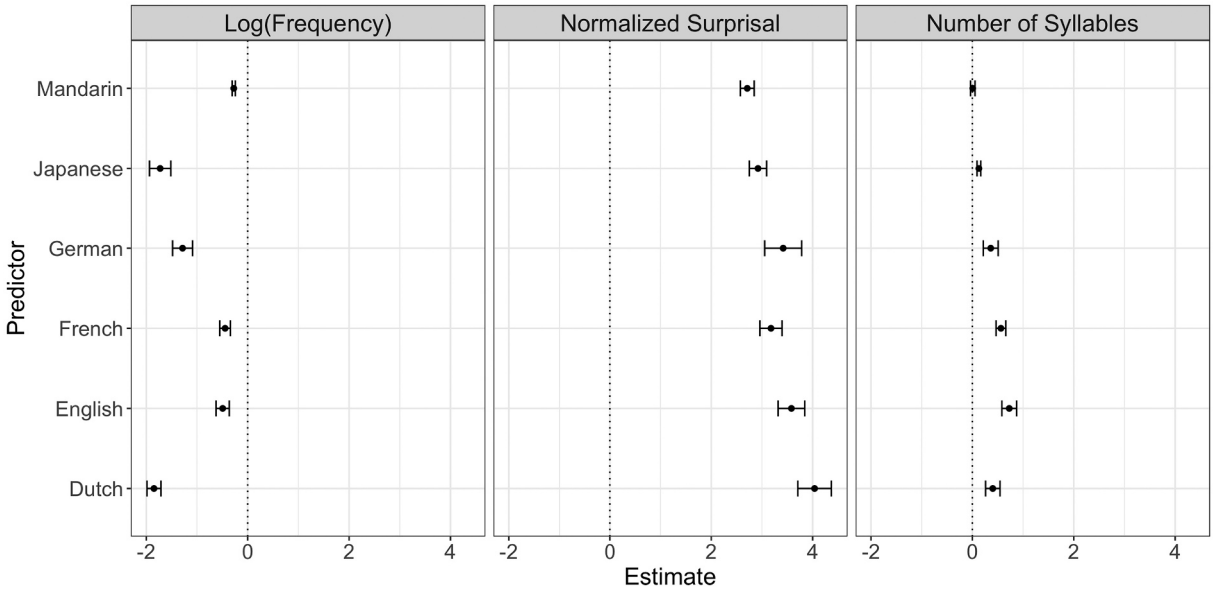

Fig. 2. Parameter estimates of Homophony Delta for Log Frequency, Normalized Phonotactic Surprisal, and Number of Syllables across all six languages. Importantly, the estimates for Log Frequency are negative for each of the six languages tested. Error bars are two standard errors.

other factors, Homophony Delta should be negatively correlated with Neighborhood Size.

To test this hypothesis, we added Log Neighborhood Size as a covariate to the models described above. Even accounting for the effects of Log Frequency, Normalized Phonotactic Surprisal, and Number of Syllables, the relationship between Log Neighborhood Size and Homophony Delta was significantly negative across all six languages: English [$\beta = -1.59, SE = 0.08, p < .001$], Dutch [$\beta = -1.1, SE = 0.23, p < .001$], German [$\beta = -1.85, SE = 0.29, p < .001$], French [$\beta = -3.49, SE = 0.14, p < .001$], Japanese [$\beta = -2.5, SE = 0.07, p < .001$], and Mandarin Chinese [$\beta = -2.1, SE = 0.05, p < .001$]. Wordforms with larger neighborhoods tended to have fewer homophones than predicted by their phonotactics (see also Fig. 3). (Critically, the effect of Log Frequency remained significant across all six languages even with the addition of Log Neighborhood Size.)

This relationship could be the product of a pressure to avoid homophones, which creates larger neighborhoods in their stead. But an alternate possibility exists—with reverse causality. Neighborhood size might affect the cost of disambiguation. Psycholinguistic research suggests that wordforms with larger neighborhoods are more likely to be confused with other wordforms in that language (Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Vitevitch & Luce, 1999; though see other work

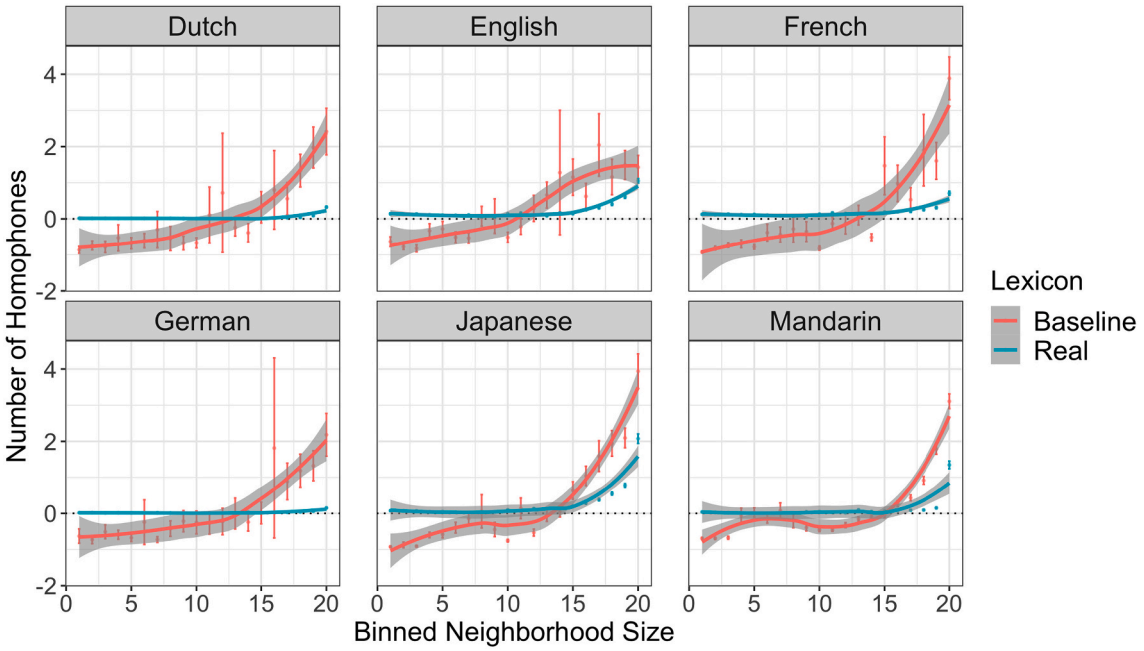

Fig. 3. Real vs. predicted number of homophones, by binned neighborhood size. Wordforms with larger phonological neighborhoods tend to have more homophones in Real lexica, but this is still fewer than predicted on the basis of their phonotactics (Baseline). Higher values of Binned Neighborhood Size correspond to larger neighborhoods. Error bars are one standard error.

(Arutiunian & Lopukhina, 2020; Vitevitch & Rodríguez, 2005) for evidence that this effect varies across languages). If high-density wordforms are already confusable, one might expect those wordforms to display a stronger resistance to acquiring additional meanings. On this explanation, larger neighborhoods—like frequency—are a *cause* of a selection pressure against homophones. The current results do not allow us to adjudicate between these possibilities; however, a prediction derived from the latter interpretation is explored in the General Discussion.

5. General discussion

Our central question was the extent to which human lexica are adapted to minimize effort for speakers or comprehenders. The uneven distribution of lexical ambiguity provides a useful test case for this question: a lexicon optimized for production ease should concentrate its meanings among the easiest wordforms to produce, such as highly frequent wordforms (Piantadosi et al., 2012; Zipf, 1949). Yet such a lexicon would require frequent disambiguation on the part of comprehenders—thus, a pressure for comprehension ease would favor a lexicon with its meanings more uniformly distributed. Adjudicating between these accounts requires the use of a "neutral" baseline, i.e., a lexicon that is agnostic with respect to the relationship between wordform frequency and ambiguity and distributes its meanings according to other known factors. We used such a baseline to estimate what the magnitude of this relationship would be if meanings were assigned to wordforms with no direct pressure *for* or *against* concentrating meanings among frequent wordforms—in this case, meanings were assigned purely as a function of a wordform's phonotactic probability and length. This allowed us to compare how many meanings each wordform actually has to the number of meanings predicted by its phonotactics.

Across six languages, we found that frequent wordforms have *fewer* homophones than predicted by their phonotactics (see Fig. 1), and in many cases, infrequent wordforms have slightly *more* homophones than expected. We also replicated this result using an alternative measure of phonotactic probability (see *Supplementary Analysis 5*). These findings are most consistent with a Comprehender-Oriented Pressure—alternatively termed *diversification* (Zipf, 1949) or *expressivity* (Kirby et al., 2015). If each additional meaning of a wordform imposes some marginal cost for comprehenders, then a lexicon whose meanings are disproportionately concentrated among frequent wordforms will impose a larger average cost than a lexicon whose meanings are more evenly distributed. Thus, from the standpoint of minimizing comprehender effort, a selection pressure *against* homophony should manifest particularly strongly among the most frequent wordforms of a lexicon. Altogether, these results suggest that any pressure to optimize production ease is weaker than a countervailing pressure to reduce the cost of frequent disambiguation. The results of *Supplementary Analysis 6*, which formalized measures of speaker and listener effort across the lexicon, are also consistent with this conclusion. Of course, these results do not entail that human lexica are entirely shaped by comprehender-centric pressures; after all, lexica do tolerate some degree of ambiguity, even among the most frequent wordforms. Thus, a speaker-centric pressure is likely at play is well—our results simply suggest that at least when it comes to the distribution of meanings across wordforms, the comprehender-centric pressure is larger.

Further, along with other recent work (Caplan et al., 2020; Ferrer-i-Cancho, Bentz, & Seguin, 2020; Gibson et al., 2019; Pimentel, Nikkarinen, Mahowald, Cotterell, & Blasi, 2021; Trott & Bergen, 2020), these results emphasize the importance of developing formal baselines when investigating questions about the relative optimality of the lexicon.

6. Limitations

One limitation of the present work is the number and identity of languages considered. We analyzed six languages, spanning three language families (Indo-European, Japonic, and Sino-Tibetan); this sample was biased towards Indo-European languages, and did not include languages from major families like Austronesian or Niger-Congo. We selected these languages since they are the only ones that have widely available lexical resources including information about individual meanings or lemmas, as opposed to wordforms; this was necessary for the current analyses. If similar resources become available for other languages, these analyses (and others) could be extended to a larger and more diverse set of languages.

Another potential concern is our choice of baseline, which itself might be divided into several lines of critique. The first critique is that n-gram models are prone to overfitting (see, e.g., Pimentel, Meister, et al., 2021). This is a valid concern, but we have replicated the primary results using an LSTM to model phonotactics (see *Supplementary Analysis 5*), following Pimentel, Nikkarinen, et al. (2021). Thus, the finding that frequent wordforms have fewer homophones than predicted by their phonotactics appears robust to the phonotactic model chosen. A second concern might be that the neutral baseline is somehow not neutral—i.e., that assigning meanings to wordforms on the basis of their phonotactic probability is disproportionately biased towards speakers (or towards listeners). If this were true, it would pose a serious problem for our theoretical interpretation, which hinges on the neutrality of this assignment procedure. However, as described in the Introduction, phonotactic probability is known to facilitate both word production (Goldrick & Larson, 2008; Vitevitch et al., 2004) and word recognition and processing (Vitevitch et al., 1999; Vitevitch & Luce, 1999). To our knowledge, there is no reason to believe that phonotactic probability disproportionately benefits speakers (or listeners). This supports the neutrality of our procedure for assigning meanings to wordforms.

A third limitation or objection is that our theoretical interpretation hinges on a crucial assumption—namely, that speakers prefer a lexicon that concentrates its meanings among a few, frequent wordforms, while comprehenders prefer a lexicon that distributes its meanings more evenly. Although this assumption is consistent with past theoretical work (Kirby et al., 2015; Zipf, 1949), we did not ground it in an explicit mathematical operationalization. Recent work (Mollica et al., 2020; Zaslavsky et al., 2018; Zaslavsky et al., 2019) has used information-theoretic tools to formalize the notions of speaker and listener effort. In *Supplementary Analysis 6*, we adopted these tools and found that, consistent with the work above, the real arrangement of wordforms and meanings is associated with lower listener effort (and higher speaker effort) than the arrangement obtained if meanings were assigned to wordforms as a function of their phonotactic probability. While this Supplementary Analysis has some limitations of its own, it is encouraging that a different methodological paradigm yielded qualitatively similar results. Future work would benefit from a more explicit grounding of the underlying semantic space.

Finally, our analyses focused on homophony. Another well-known kind of lexical ambiguity is polysemy, in which the same wordform has multiple, related meanings. Unlike homophony, polysemous words appear to enjoy advantages in both word learning (Floyd & Goldberg, 2021; Rodd et al., 2012; Srinivasan, Berner, & Rabagliati, 2019) and processing (Klepousniotou, 2002; Klepousniotou, Pike, Steinhauer, & Gracco, 2012; Rodd, Gaskell, & Marslen-Wilson, 2002). Thus, it is possible that polysemy—unlike homonymy—may even be selected for (Xu, Duong, Malt, Jiang, & Srinivasan, 2020). If this is true, then one might also expect the opposite pattern of results to the ones reported here: the most frequent wordforms should also be *even more* polysemous than predicted by their phonotactics. In contrast, if the cost of disambiguation is still too high, a comprehender-oriented pressure for expressivity may win out even in the case of polysemy. However, one challenge to analyzing polysemy in this way is the lack of consensus about what exactly constitutes a distinct "sense" (Kilgarriff, 2007; Brown, 2008; Krishnamurthy & Nicholls, 2000). Some resources make relatively fine-grained distinctions, while others aim for more coarse-grained sense inventories (Lacerra, Bevilacqua, Pasini, & Navigli, 2020). Future work in this area would thus benefit from additional resource development.

7. Future research

Our findings point to other promising directions for future research. A first step would be to identify other factors that contribute to disambiguation cost. For example, many homophones are unbalanced, such that one meaning is used much more frequently than others. From the perspective of minimizing cost, unbalanced homophones might be preferred—if one meaning is much more frequent than another, comprehenders could simply assume the dominant meaning was intended, generally avoiding the need to disambiguate. This is consistent both with psycholinguistic evidence, which suggests that comprehenders tend to activate the dominant meaning of a homophone (Blott et al., 2020; Duffy, Morris, & Rayner, 1988), as well as work on historical sound change (Wedel, Kaplan, & Jackson, 2013), which finds that phoneme mergers are especially unlikely if those mergers would create homophones among balanced minimal pairs. This interpretation also makes a testable prediction: homophones with a more uniform distribution over meanings should be more resistant to acquiring additional meanings. We tested this prediction in a supplementary analysis (see *Supplementary Analysis 2*), operationalizing meaning uncertainty as the Shannon entropy over possible senses of a wordform (Meylan, Mankewitz, Floyd, Rabagliati, & Srinivasan, 2021). We found no significantly negative relationship between Sense Entropy and Homophony Delta in three of the five languages tested, though we did find a significantly negative relationship in German and Mandarin.

One explanation for these results is that disambiguation cost is driven primarily by *contextual* discriminability—i.e., how much information context provides about the intended meaning of an ambiguous wordform. In other words, the critical factor may not be the entropy over meanings in isolation, $H(X)$, but the conditional entropy over meanings given some informative context, $H(X \mid C)$ (Piantadosi et al., 2012). Presumably, the homophones that *do* persist in a lexicon are those whose distinct meanings are sufficiently distinguishable in context (Dautriche, Fibla, Fievet, & Christophe, 2018; Piantadosi et al., 2012). Human comprehenders exploit a number of contextual cues to disambiguate, including grammatical class (Dautriche et al., 2018), co-speech gesture (Holle & Gunter, 2007; Holler & Beattie, 2003), linguistic context (Aina, Gulordava, & Boleda, 2019), and even the speaker's accent (Cai et al., 2017). Contextual discriminability should reduce the cost of disambiguation for a given wordform, thus easing the selection pressure against adding more meanings to that wordform. If this is true, a pressure against homophony should be *weaker* among wordforms with more contextually discriminable meanings, and *stronger* among wordforms whose meanings are less discriminable. Measuring contextual discriminability at scale is challenging, but future work could rely on sense-annotated corpora (Langone, Haskell, & Miller, 2004; Meylan et al., 2021), or use neural language models to derive an estimate of the residual uncertainty over meanings, given context (Pimentel, Maudslay, Blasi, & Cotterell, 2020).

The main findings reported above also inform accounts of how individual-level cognitive and communicative constraints produce emergent, lexicon-wide trends at longer timescales through language change. The presence of lexical ambiguity might elicit errors among language learners (Casenhiser, 2005) or adult comprehenders (Blott et al., 2020)—either because the cost of disambiguation was too high, or because they selected a meaning other than the one intended. Through a process of online, interactive repair (van Arkel, Woensdregt, Dingemanse, & Blokpoel, 2020), speakers might then use a different word (or series of words) to convey their intended meaning. Over many interactions, a population of speakers might drift towards using a different word in the first place, avoiding the need for disambiguation or repair. This decision need not involve explicit or conscious ambiguity avoidance on the part of speakers, which is known to be challenging and rare (Ferreira, 2008; Wasow, 2015). Rather, it might reflect a form of implicit learning or routinization (Ferreira, 2019); the language production system might learn that when trying to convey meaning m, wordform w_2 (as opposed to ambiguous wordform w_1) is often used successfully. Correspondingly, the use of wordform w_1 to convey meaning m should eventually decrease, as an appropriate and less ambiguous substitute has been identified. In this way, failures of comprehension could drive future production decisions, which in turn shape lexicon structure.

Across longer timescales, one might look to processes like sound

change, which are known to generate homophony (Ke, 2006; Sampson, 2013; Sampson, 2015), yet which also appear to be sensitive to a pressure to *avoid* homophones (Wedel, Kaplan, & Jackson, 2013; Yin & White, 2018; Ceolin, 2020). For example, phoneme mergers are statistically less likely for pairs of phonemes that carry higher functional load, i.e., which distinguish more minimal pairs (Wedel, Kaplan, & Jackson, 2013). Here, the findings above lead to another concrete prediction: phoneme mergers should be especially unlikely if they would create homophones among the most *frequent* wordforms of a language. In other words, a pressure for homophony avoidance should be strongest among frequent wordforms. To our knowledge, such a prediction has not been directly tested. A second testable prediction regarding historical sound change comes from the relationship observed above between neighborhood size and homophone resistance. One interpretation of this finding is that high-density wordforms are more perceptually confusable, and thus display a stronger resistance to acquiring more meanings. If perceptual confusability plays a role in homophone avoidance during historical sound change, phoneme mergers should be less likely if they would create homophones among high-density wordforms. Both predictions could be tested using historical data about phoneme mergers across time and languages (Wedel, Jackson, & Kaplan, 2013; Wedel, Kaplan, & Jackson, 2013).

8. Conclusion

Overall, our results are consistent with the claim that languages are well-designed for human use (Gibson et al., 2019; Mahowald et al., 2018; Piantadosi et al., 2009): lexica distribute their meanings in a way that reduces the cost of disambiguation. But they also support a nuanced view of "efficiency". As others (Piantadosi et al., 2012; Zipf, 1949) have noted, minimizing the effort of certain processes (e.g., production) can make other processes more challenging (e.g., disambiguation). Humans have limited cognitive resources at their disposal (Lieder & Griffiths, 2020), and these limitations create trade-offs across many domains of communication. Identifying these tension points allows us to ask more targeted questions about how this pressure for efficiency operates within and across languages. Thus, when we assess the claim that language is efficient, we might do well to begin by asking: efficient for whom?

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to Tamara Rhodes for helping to procure the Japanese CallHome Lexicon. We also thank Catherine Arnett for her guidance to using the Mandarin CallHome Lexicon, Vic Ferreira for his help on understanding speaker-oriented accounts of language design, Tyler Marghetis for his valuable pointers to the use of neutral models in evolutionary biology, members of the Language and Cognition Lab (James Michaelov, Cameron Jones, and Tyler Chang) for their comments and feedback, and Isabelle Dautriche and Kyle Mahowald for making the code for modeling phonotactic probability available on GitHub. We are also grateful to Leon Bergen and Tyler Chang for lengthy and valuable discussions of information theory. Finally, we thank Tiago Pimentel for making the code to model phonotactics using an LSTM available on GitHub, and for his very helpful guidance to implementing and adapting the code.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105094.

References

Aina, L., Gulordava, K., & Boleda, G. (2019). Putting words in context: LSTM language models and lexical ambiguity. *arXiv preprint*, 1–7. arXiv:1906.05149.
Alonso, D., Etienne, R. S., & McKane, A. J. (2006). The merits of neutral theory. *Trends in Ecology & Evolution, 21*(8), 451–457.
van Arkel, J., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020, November). Explaining the efficiency of communication: How communicators can reduce their computational burden through interaction. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 177–194).
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. (1995). *The CELEX lexical database (release 2). Distributed by the Linguistic Data Consortium*. University of Pennsylvania.
Bentley, R. A., Carrignon, S., Ruck, D. J., Valverde, S., & O'Brien, M. J. (2021). Neutral models are a tool, not a syndrome. *Nature Human Behaviour*, 1–2.
Bentz, C., & Ferrer Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics* (pp. 1–4). University of Tübingen.
Blott, L. M., Rodd, J. M., Ferreira, F., & Warren, J. E. (2020). Recovery from misinterpretations during online sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition., 47*(6), 968–997.
Brown, S. W. (2008, June). Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers* (pp. 249–252).
Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One, 5*(6), Article e10729.
Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology, 98*, 73–101.
Caplan, S., Kodner, J., & Yang, C. (2020). Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition, 205*, Article 104466.
Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language, 32*(2), 319.
Ceolin, A. (2020). On functional load and its relation to the actuation problem. *University of Pennsylvania Working Papers in Linguistics, 26*(2), 6.
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.
Conway, B. R., Ratnasingam, S., Jara-Ettinger, J., Futrell, R., & Gibson, E. (2020). Communication efficiency of color naming across languages provides a new framework for the evolution of color terms. *Cognition, 195*, Article 104086.
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. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5*(4), 313–349.
Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27*(4), 429–446.
Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In *Perspectives on Morphological Organization* (pp. 212–239).
Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science, 10*(1), 209–224.
van Esch, D. (2012). Leiden Weibo Corpus. http://lwc.daanvanesch.nl.
Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation, 49*, 209–246.
Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology, 70*, 29–51.
Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2020). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, 1–30. https://doi.org/10.1080/09296174.2020.1778387
Floyd, S., & Goldberg, A. E. (2021). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(1), 29.
Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., … Conway, B. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences, 114*(40), 10785–10790.
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.
Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science, 24*(7), 1079–1088.
Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition, 107*(3), 1155–1164.
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.
Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences, 117*(5), 2347–2353.

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.
Huang, S., Bian, X., Wu, G., & McLemore, C. (1996). *CALLHOME Mandarin Chinese Lexicon LDC96L15. Web Download.* Philadelphia: Linguistic Data Consortium.
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), 630–645.
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.
Kilgarriff, A. (2007). Word senses. In *Word Sense Disambiguation* (pp. 29–46). Dordrecht: Springer.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87–102.
Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language, 81*(1–3), 205–223.
Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language, 123*(1), 11–21.
Kobayashi, M., Crist, S., Kaneko, M., & McLemore, C. (1996). *CALLHOME Japanese Lexicon LDC96L17. Web Download.* Philadelphia: Linguistic Data Consortium.
Krishnamurthy, R., & Nicholls, D. (2000). Peeling an onion: The lexicographer's experience of manual sense-tagging. *Computers and the Humanities, 34*(1), 85–97.
Kruyt, J. G., & Dutilh, M. W. F. (1997). A 38 million words Dutch text corpus and its users. *Lexikos, 7*, 229–244.
Kupietz, M., & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In *, 3. Working Papers in Corpus-Based Linguistics and Language Education* (pp. 53–59).
Lacerra, C., Bevilacqua, M., Pasini, T., & Navigli, R. (2020, April). CSI: A coarse sense inventory for 85% word sense disambiguation. In *, 34, No. 05. Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 8123–8130).
Langone, H., Haskell, B. R., & Miller, G. A. (2004). *Annotating wordnet.* NJ, Cognitive Science Lab: Princeton University.
Leroi, A. M., Lambert, B., Rosindell, J., Zhang, X., & Kokkoris, G. D. (2020). Neutral syndrome. *Nature Human Behaviour, 4*(8), 780–790.
Levinson, S. (2000). *Presumptive meanings: The Theory of Generalized Conversational Implicature.* MIT Press.
Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*.
Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1.
MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*, 226.
MacDonald, M. C. (2015). The emergence of language comprehension. In B. MacWhinney, & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 81–99). Chichester, UK: Wiley.
Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science, 42*(8), 3116–3134.
Meylan, S., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). *Quantifying Lexical Ambiguity in Speech To and From English-Learning Children.*
Meylan, S. C., & Griffiths, T. L. (2017). Word forms-not just their lengths-are optimized for efficient communication. *arXiv preprint*, 1–16. arXiv:1703.01694.
Mollica, F., Bacon, G., Xu, Y., Regier, T., & Kemp, C. (2020, August). Grammatical marking and the tradeoff between code length and informativeness. In *CogSci*.
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.
Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *Nature, 551*(7679), 223–226.
Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*(4), 273–281.
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. In *, Vol. 2582. The 31st Annual Meeting of the Cognitive Science Society (CogSci09)* (p. 2587). Austin, TX: Cognitive Science Society.
Pimentel, T., Maudslay, R. H., Blasi, D., & Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. *arXiv preprint*, 1–12. arXiv:2010.02172.
Pimentel, T., Meister, C., Teufel, S., & Cotterell, R. (2021). On homophony and Renyi entropy. *arXiv preprint*, 1–10. arXiv:2109.13766.
Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021). How (non-) optimal is the lexicon? *In Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 4426–4438. arXiv:2104.14279.
R Core Team. (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.
Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*(3), 191–201.
Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(5), 779.
Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS One, 11*(4), Article e0151138.
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.
Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition, 40*(7), 1095–1108.
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.
Sinclair, J. (1987). *Collins COBUILD English Language Dictionary*. London: William Collins.
Srinivasan, M., Berner, C., & Rabagliati, H. (2019). Children use polysemy to structure new word meanings. *Journal of Experimental Psychology: General, 148*(5), 926.
Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research., 44*(6), 1321–1337.
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*, Article 104449.
Vitevitch, M. S., & Aljasser, F. M. (2021). Phonotactics in spoken-word recognition. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The Handbook of Speech Perception* (pp. 286–308). Hoboken, NJ, USA: John Wiley & Sons.
Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 514.
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., & 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., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*(1–2), 306–311.
Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders, 3*(1), 64–73.
Wasow, T. (2013). The appeal of the PDC program. *Frontiers in Psychology, 4*, 236.
Wasow, T. (2015). Ambiguity avoidance is overrated. In S. Winkler (Ed.), *Ambiguity: Language and Communication* (pp. 21–51). Berlin: DeGruyter.
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.
Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37–48.
Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition, 201*, Article 104280.
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.
Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021, May). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd annual meeting of the cognitive science society*. Cognitive Science Society.
Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. *arXiv preprint*, 1–7. arXiv:1905.04562.
Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology, 33*(2), 251–256.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press.