Cognition 267 (2026) 106372

Contents lists available at ScienceDirect

Cognition

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

Full length article

# The lexicon adapts to competing communicative pressures: Explaining patterns of word similarity

Aislinn Keogh [a,b,*], Jennifer Culbertson [a], Simon Kirby [a]

[a] Centre for Language Evolution, University of Edinburgh, United Kingdom
[b] Department of Psychology, Durham University, United Kingdom

ARTICLE INFO

Dataset link: https://osf.io/vsy6z/

*Keywords:*
Language evolution
Communication
Efficiency
Lexicon
Computational modelling
Artificial language learning

ABSTRACT

Cross-linguistically, lexicons tend to be more phonetically clustered than required by the phonotactics of the language; that is, words within a language are more similar to each other than they need to be. In this study, we investigate how this property evolves under the influence of competing communicative pressures: a production-side pressure to re-use more easily articulated sounds, and a comprehension-side pressure for distinctiveness of wordforms. In an exemplar-based computational model and a communication experiment using a miniature artificial language, we show that natural-language-like levels of clustering emerge from a trade-off between these pressures. With only one pressure at work, the resulting lexicons tend to inhabit an extreme region of the possible design space: production pressures alone give rise to maximally clustered lexicons, while comprehension pressures alone give rise to maximally disperse lexicons. We also test whether clustering emerges more strongly for high-frequency items, but our results lend support only to a weak relationship between frequency and clustering. Overall, this study adds to a growing body of evidence showing that mechanisms operating at the level of individual language users and individual episodes of communication can give rise to emergent structural properties of language.

## 1. Introduction

Different languages have different rules about how sounds can be combined to form words. For example, “zad” is an unattested but possible word of English, whereas “zbad” is both unattested and impossible (but could be a word of Polish). Naturally, the fact that these rules differ between languages means that words within a language generally sound more similar to each other than they do to words of other languages. Indeed, both infants (Jusczyk et al., 1993; Mehler et al., 1988; Moon et al., 1993) and adults (Lorch & Meara, 1989; Marks et al., 2003; Stockmal et al., 1996) can discriminate surprisingly well between languages, even ones they don’t know.

Perhaps less obvious is the fact that, even within a language, possible sounds and sound combinations are not necessarily equally frequent. Fig. 1 gives a sense that, while “zad” is a phonotactically legal sound sequence in English, it is perhaps not very likely to be coined as a new word: the [z] phoneme is relatively uncommon in English (especially in word-initial position), and the [zæ] biphone is extremely low-frequency. This skewed distribution is not unique to English: it is a common property across languages that not all possible sounds or sound sequences are equally frequent (Krevitt & Griffith, 1972; Macklin-Cordes & Round, 2020; Martindale et al., 1996). As a result, words within a language are actually more similar *to each other* than they really need to be. In other words, lexicons are *phonetically clustered*.

Naively, we might expect languages to use up their available phonotactic space more uniformly; that is, words could be evenly distributed in this space to avoid repeating sound sequences where possible. Successful communication depends on listeners being able to perceive and interpret a speaker’s message with a high degree of accuracy. And since communication takes place over a noisy channel (Gibson et al., 2013; Levy, 2008; Shannon, 1948), there is always a possibility that information will be lost; a lexicon that maximised the distance between words would reduce this possibility (Flemming, 2004). Indeed, we know that comprehension is easier when words are more distinct: in line with the Neighbourhood Activation Model (Luce & Pisoni, 1998), words from sparser phonological neighbourhoods and less densely connected areas of the lexical network (i.e. words that are less similar to other words) are recognised more quickly and accurately, especially in noisy conditions (Chan & Vitevitch, 2009; Cluff & Luce, 1990; Goldinger et al., 1989; Magnuson et al., 2007; Siew & Vitevitch, 2016; Vitevitch & Luce, 1998).

[*] Correspondence to: Rowan House, Mountjoy Centre, Durham, DH1 3TN, United Kingdom.
*E-mail address:* aislinn.k.keogh@durham.ac.uk (A. Keogh).

https://doi.org/10.1016/j.cognition.2025.106372
Received 24 December 2024; Received in revised form 9 September 2025; Accepted 4 November 2025
Available online 7 November 2025

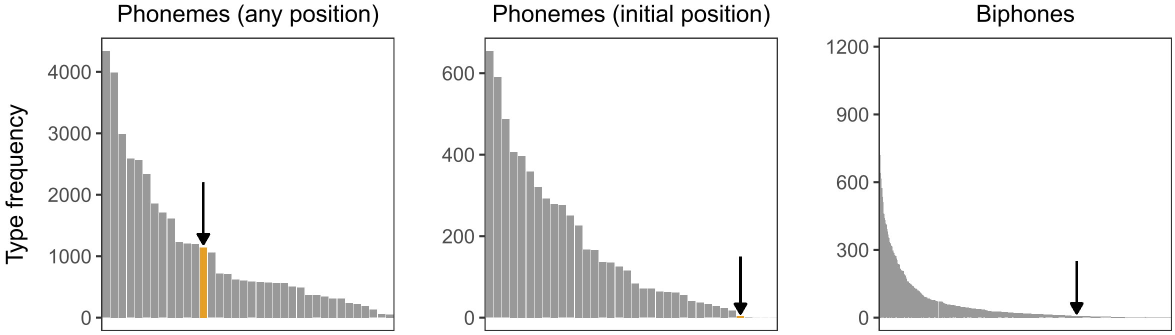

**Fig. 1.** Type frequency of all phonemes and biphones of English, derived from the British National Corpus (BNC Consortium, 2007) using List 1.2 (rank frequency list for the whole corpus, limited to words with a frequency of at least 100 per million) from Leech et al. (2001), converted to IPA using the `eng-to-ipa` package in Python (https://pypi.org/project/eng-to-ipa/). Yellow bars and arrows indicate the [z] phoneme in the left-hand and middle panels, and the [zæ] biphone in the right-hand panel. The specific identity of other phonemes/biphones is not shown on the *x*-axis for ease of presentation; there are 36 unique phonemes and 670 unique biphones represented in the word list. The key observation is that the shape of all these distributions is skewed: certain sounds and sound sequences are considerably more frequent than others.

However, the effect of word similarity on comprehension is not completely straightforward. In particular, increases in phonotactic probability (which reflects the existence of high-frequency sound sequences within a word[1]) have been found to be beneficial for word recognition (Vitevitch & Luce, 1998, 1999; Vitevitch et al., 1997, 1999). Furthermore, there is good evidence that spoken word *production* is facilitated by increases in both neighbourhood density *and* phonotactic probability (Chen & Mirman, 2012; Gahl et al., 2012; Goldrick & Larson, 2008; Goldrick & Rapp, 2007; Munson, 2001; Stemberger, 2004; Vitevitch et al., 2004; Vitevitch & Luce, 1998, 2005; Vitevitch & Sommers, 2003). That is, words that are more similar to other words are generally pronounced more quickly and accurately.

This suggests that communication involves a complex interplay of different functional pressures coming from both production and perception, and taken together these do not straightforwardly point to an overall advantage or disadvantage of word similarity. How might language users balance these competing pressures in a way that leads to phonetically clustered lexicons? Almost 80 years ago, the linguist George Kingsley Zipf claimed that the organisational structure of languages is shaped by a trade-off between a pressure for accurate communication on the one hand, and a pressure for efficiency on the other (Zipf, 1949). Although this claim is most famously instantiated in the “Law of Abbreviation” — whereby more frequent words tend to be shorter — Zipf also argued that languages should preferentially re-use easy-to-articulate sounds over more difficult sounds (Zipf, 1935). A related argument was made by Piantadosi et al. (2012), who suggest that an efficient communication system should re-use more easily produced words and sounds, even if doing so results in some ambiguity.

Of course, there are several reasons why lexicons might re-use particular sounds more than others (as in Fig. 1), not all of which point to an adaptive explanation. For example, we would expect certain sounds to reoccur across many words in languages with productive morphology: *unkind*, *unsatisfying* and *unpleasant* all sound somewhat similar because of a shared prefix, while *tangled*, *entangle* and *disentangle* all sound extremely similar because of a shared root. Words that sound similar may also tend to have similar meanings (Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Monaghan et al., 2014) or syntactic functions (Kelly, 1992), although form-meaning correspondences are generally very subtle; phonaesthemes are a notable exception (Bergen, 2004). And many words that map to distinct categories in their modern form trace their origins back to a shared ancestor; for example, *skirt* and *shirt* sound similar because they both come from the Old Norse *skyrta*.

Naturally, phonotactic constraints are also a major source of phonetic clustering: sounds and sound sequences that can appear in more contexts will be more frequent across a language. Nonetheless, corpus analysis reveals a cross-linguistic tendency for lexicons to be *even* more clustered than required by the phonotactics of the language (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017). That is, the skewed distributions we saw in Fig. 1 are not simply a byproduct of phonotactic restrictions in English: these distributions are, in fact, *surprisingly* skewed. In particular, across a range of word lengths, high-frequency words tend to be more tightly clustered — both in terms of neighbourhood density and phonotactic probability — while lower frequency words tend to be more distinctive (Frauenfelder et al., 1993; King & Wedel, 2020; Landauer & Streeter, 1973; Mahowald et al., 2018; Meylan & Griffiths, 2024). This pattern is suggestive of adaptation for efficient communication (Gibson et al., 2019; Jaeger & Tily, 2011), since it minimises production effort for items that are produced most often, and maximises understandability for low-frequency items, which are often harder to process in comprehension (Brysbaert et al., 2018). More generally, the fact that lexicons are observably less disperse than they could be suggests that, overall, the advantages associated with word similarity outweigh the disadvantages. However, corpus data alone cannot provide causal evidence of a relationship between particular functional pressures and the structure of language.

In this study, we investigate how production and comprehension pressures compete to shape the degree of phonetic clustering in the lexicon. First, we set out an agent-based computational model of sound change (Section 2). In line with the psycholinguistic evidence reviewed above, we model production and comprehension pressures that pull in opposite directions. We test the prediction that natural-language-like lexicons will emerge only under the combined influence of both. In particular, we test whether clustered lexicons emerge, and whether this clustering is found particularly for high frequency words. To further explore the role of production and comprehension in shaping the lexicon, we then model a similar process in a behavioural experiment in which human participants communicate with a partner using a miniature artificial language (Section 3). To preview our results, the

[1] More precisely, phonotactic probability for a given word is usually calculated by extracting from a corpus the probability with which each phoneme (or biphone) occurs — either overall, or in a particular position — across the words of the language, and then averaging over all segments in the word. Some metrics (e.g. Mayer et al., 2022) instead take the product of the individual probabilities, which means that decreases in word length automatically result in *increased* phonotactic probability. Naturally, word length is controlled for in all the studies we cite here, such that the increases in phonotactic probability we refer to are above and beyond the effect of word length.

lexicons that emerged from our model when both production and comprehension pressures were at play were more clustered than those generated by comprehension pressures alone, but more disperse than those generated by production pressures alone. Similarly, in the experiment, manipulating the difficulty of only the production task or only the comprehension task gave rise to behaviours at one extreme or the other. When both tasks were difficult, participants adopted a variety of strategies, but overall there was more of a balance between ease of production and ease of perception. However, the effect of frequency on emergent lexicons was less clear; there was a subtle tendency in the model for more frequent words to become more clustered, but this pattern did not robustly materialise in the experiment.

## 2. Computational model

We use an agent-based exemplar model (Nosofsky, 1986; Wedel, 2006) to test how mechanisms operating during individual episodes of production and comprehension might influence the degree of phonetic clustering present in a lexicon over time. In this model, pairs of agents use a miniature artificial language to communicate with each other over repeated rounds. In each communication round, agents take turns producing and interpreting signals, with some mechanisms that would be expected to favour or disfavour word similarity encoded within these processes (described in Section 2.1.3). Signals that result in successful communication are strengthened over time, while unsuccessful signals are more likely to drop out of the agents' memory. At the end of every round, we observe the state of the lexicon. The following section describes all of these components in detail; an overview is given in Fig. 2. Readers wishing to skip the technical details can move on to Section 2.3 to see the results.

### 2.1. Details of the model

The model is implemented in Python 3.11; full code is available at https://osf.io/vsy6z/.

#### 2.1.1. The agents

Each agent maintains their own independent internal representation of the lexicon, based on prior evidence. An agent's internal representation consists of 20 atomic meaning categories (represented by integers), each associated with a collection of signals. In the most basic version of the model, all meanings are equally frequent; we implement a simple frequency manipulation in Section 2.3.1. Each meaning category has a memory limit $S$ (default value = 10) which constrains the number of signals that can be associated with it at any given time-point. When a new signal needs to be added to a category that is already at this limit, a random older signal is deleted first.

Since the model is exemplar-based, there is no abstract representation for agents to infer from the evidence they receive; rather, they store concrete exemplars of linguistic behaviour they've observed. As in Wedel (2012), we do not intend to make any claims about the specific nature of humans' mental lexicons[2]; this architecture is simply a convenient and transparent way to capture the fact that there is always fine-grained phonetic variation below the level of "the lexicon", and to show how this variation can provide the fodder for lexical evolution (Winter, 2014). More specifically, while we might perceive words as having categorical boundaries, in reality, subtle variations in pronunciation mean that word boundaries are at least somewhat fuzzy, even within the same individual; different exemplars in our model can be thought of as representing this fuzziness.

One consequence of this exemplar-based implementation is that it is perfectly possible for an agent's internal representation to contain duplicates, because they are quite likely to observe the same signal multiple times across the course of interaction. Storing multiple copies of a given exemplar effectively corresponds to having higher baseline activation for that exemplar; for other methods of implementing activation, see e.g. Hintzman (1986); Wedel (2012); Wedel and Fatkullin (2017).

#### 2.1.2. The lexicon

The "words" agents store in our model are character strings. Because we are interested in how clustering might emerge above and beyond the effects of word length (since shorter words are, necessarily, more similar to each other than longer words), word length is a constant in our model: all words are of length 8. For simplicity, the individual segments that make up a word are represented simply by letters, rather than by bundles of features or some other more phoneme-like representation (cf. Wedel, 2012). Because of this simplification, it is not the case that segments can be more or less similar to each other: two segments are either identical, or they are different. Although this makes comparisons between words less nuanced, it is a reasonable simplification to improve model tractability, particularly given the lack of evidence that natural language lexicons are more clustered around highly distinctive contrasts than around more confusable contrasts (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017).

At the start of each run of the model, we generate 20 words (one per meaning category) by randomly combining letters from the set of English consonants. Letters are drawn from a uniform distribution, meaning that there is no pressure towards clustering coming from the initial lexicons. We use these words to seed a process of exemplar creation: specifically, the starting set of exemplars in each meaning category is a collection of $S$ strings (where $S$ is the memory limit for that category), each of which is created by randomly substituting a single character from the seed word assigned to that category. For example, if the seed word for a category was "tam", it could generate exemplars like "zam", "tum", and "tak".

Although agents therefore store a considerable amount of variation in their internal representation, we are treating exemplars as pronunciation variants of the same word, so we want to smooth out this within-category variation when we examine the state of the lexicon. To collapse an agents' internal representation down to a single word per meaning category — the canonical or 'average' form of the word — we simply concatenate the most common character in each position across all exemplars in that category. For example, given a set of exemplars {"miq", "mas", "taq", "maq"}, this process of concatenation would yield the word "maq", since "m" is the most common first letter, "a" is the most common second letter, and "q" is the most common final letter.

In order to analyse how the lexicon changes over time, and whether words are becoming more or less similar to each other, we calculate the *average pairwise edit distance* between words at each time step, including for the initial lexicon. Average pairwise edit distance, $D(L)$, is given by:

$$D(L) = \frac{\sum_{i,j \in L, i \neq j} LD(i,j)}{|L| \cdot (|L| - 1)} \tag{1}$$

where $L$ is the lexicon, $|...|$ indicates cardinality (i.e. the number of words in $L$), $i$ and $j$ are words and $LD(i,j)$ is the Levenshtein distance between two words. That is, we calculate the edit distance between every pair of words in the lexicon, and then take the mean of these distances.

Because we generate the seed words randomly — so that all characters are equally likely to appear in all positions — words in the initial lexicon are always very different from each other: across 1000 randomly generated lexicons, average pairwise edit distance had a mean value of 7.54 ($SD = 0.05$). In other words, in the initial lexicon, any two randomly selected words will usually differ at every position. If words are becoming more similar to each other over time, this would be reflected by a *decrease* in average pairwise edit distance.

[2] The model could equally have been implemented in a Bayesian framework, with a compression-based prior (Kirby et al., 2015) that would favour lexicons with fewer unique sounds and sound combinations.

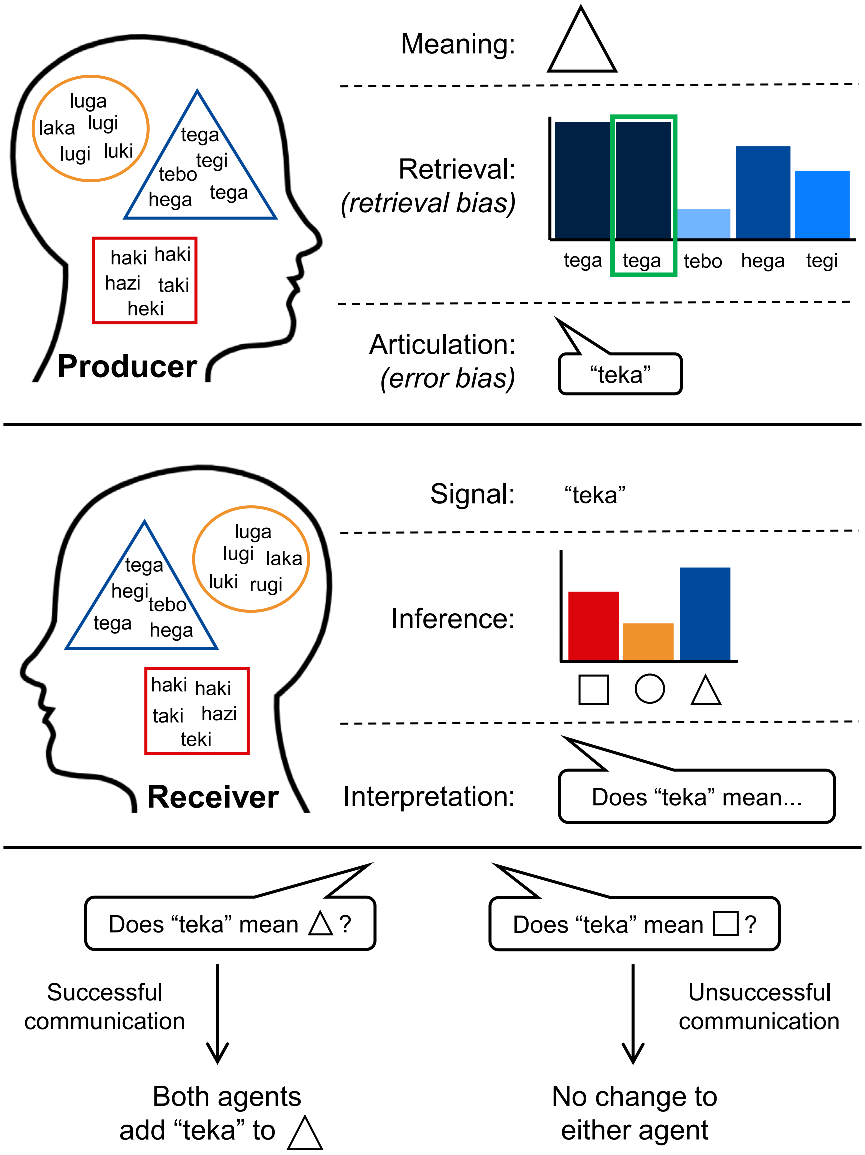

**Fig. 2.** Overview of the model architecture for a single communication episode. Both agents maintain an independent internal representation of the lexicon in the form of meaning categories (shapes) and associated signals (exemplars, which need not be unique: see Section 2.1.1). The Producer sends a signal to their partner to communicate about a target meaning, with two sources of similarity bias in this process. First, exemplars within the target meaning category are activated to different degrees depending on their phonotactic probability, meaning that exemplars that are more similar to others in the lexicon are more likely to be retrieved. Second, once an exemplar has been retrieved, there is some probability of an error being introduced into it during production; when an error is made, segments that are less frequent across the lexicon tend to be replaced by those that are more frequent. The Receiver compares the received signal to their stored exemplars to calculate a probability distribution over possible meanings, from which they sample a response; more distinctive signals give higher weight on the target meaning category relative to all other categories and are therefore more likely to result in successful communication, while signals that are more ambiguous between categories give a more uniform distribution over meanings and are therefore more likely to be misinterpreted. If the Receiver correctly infers the Producer's target meaning, both agents store the signal that was just sent as a new exemplar in that meaning category.

### *2.1.3. Communication*

In each communication round, agents take turns as Producer and Receiver for all meanings. The Producer's task is to transmit a signal given a target meaning; the Receiver's task is to decode the intended meaning given a received signal. Whenever the Receiver successfully recovers the meaning of a signal, both agents store that signal as a new exemplar in the relevant meaning category. Due to the memory limit described in Section 2.1.1, exemplars that are either not used or do not result in successful communication will tend to drop out of the agents' internal representations over time.

*Production.* Production consists of two stages: retrieval and articulation. In both of these stages, we build in observations from the psycholinguistic literature about how word similarity benefits word production. To summarise, exemplars that are more similar to others in the agent's internal representation are retrieved more easily (Chen & Mirman, 2012; Goldrick & Larson, 2008; Vitevitch, 2002; Vitevitch et al., 2004), and errors in the pronunciation of a target exemplar tend to replace lower frequency segments with higher frequency ones (Dell, 1986; Goldrick & Rapp, 2007; Levitt & Healy, 1985; Motley & Baars, 1975; Munson, 2001), thus creating sequences with higher phonotactic probability.

More specifically, production begins with the random choice of an exemplar from the target meaning category, where the probability of a particular choice depends on its phonotactic probability (average bigram positional probabilities across the string); exemplars with higher phonotactic probability are more strongly activated (the *retrieval bias* parameter). Before the exemplar is transmitted to the Receiver, an error is introduced into it with probability $E$.[3] All errors involve the substitution of a single segment in a randomly chosen position. The new segment is sampled from the set of segments in the language, where the probability of selecting a particular segment depends on the frequency with which it occurs in the same context as the original segment across all exemplars in the agent's internal representation (the *error bias* parameter). By default, we only consider a single preceding segment when calculating conditional segment frequencies; in this way, errors tend to create high-probability bigrams. We use Laplace smoothing with parameter 0.01 to assign non-zero probability to segments that were present in the initial lexicon but have dropped out entirely, or segments that don't appear in a particular bigram. We also allow "substitution" to replace a segment with itself, which can happen when the segment targeted for error is very high-frequency in the given position; in this way, exemplars with high phonotactic probability in the language become less likely to be mispronounced.

*Reception.* The final signal created by the Producer, including any error, is transmitted to the Receiver along with a context (list of possible meanings) which they have to choose from. The nature of this context is controlled by a *context size* parameter, which can take one of three values: maximal (the default: all meanings in the lexicon), random ($n$ randomly selected meanings, where $1 \leq n \leq 20$), or minimal ($= 1$).[4]

When the Receiver hears a signal, they must infer its meaning by comparing it to all their stored exemplars for each meaning category in the current context. If the context contains only one meaning, the Receiver automatically assigns the signal to that meaning category. Otherwise, the probability of recovering the intended meaning is calculated using the Generalised Context Model (Nosofsky, 1986, 2011),[5] which states that the probability of classifying stimulus $i$ into category $c_n$ is given by:

$$P(c_n|i) = \frac{\left[\sum_{j\in c_n} N_j \cdot \eta_{ij}\right]^{\gamma}}{\sum_{c\in C}\left[\sum_{k\in c} N_k \cdot \eta_{ik}\right]^{\gamma}} \tag{2}$$

where $\eta_{ij}$ denotes the similarity between exemplars $i$ and $j$ and $N_j$ is the frequency of exemplar $j$. The numerator is therefore simply the summed similarity score for the meaning category under consideration, and the denominator is the sum of all similarity scores for all meaning categories. $\gamma$ is a response-scaling parameter which controls the Receiver's sampling behaviour: when $\gamma = 1$, the Receiver responds by sampling directly from the distribution of relative summed similarities over all categories (i.e. probability matching), whereas for higher values of $\gamma$, the Receiver responds more deterministically with the category that yields the largest summed similarity. Similarity between exemplars $i$ and $j$ is itself operationalised as the complement of the Levenshtein distance $LD$ between the two strings, normalised by dividing by $M$, the length of the longer string[6]:

$$\eta_{ij} = 1 - \frac{LD(i,j)}{M} \tag{3}$$

The Receiver samples a meaning from the context using the relative similarity scores given by Eq. (2) as weights. The effect of this reception mechanism is that more distinctive signals will be more likely to result in successful communication, since they will give higher weight on the target meaning category relative to all other categories. On the other hand, signals that are similar to exemplars in multiple categories will give a more uniform distribution over possible meanings, and are therefore more likely to be misinterpreted.

#### 2.1.4. Iteration

At the end of every communication round, we extract the current state of the lexicon from one of the agents (randomly chosen) and calculate its average pairwise edit distance, $D(L)$. A new communication round then starts; each run of the model consists of 4000 such rounds. Note that there is no transmission of the language to naive individuals between communication rounds (cf. Kirby et al., 2015); the same pair of agents continue to communicate with each other throughout the simulation. Since there are no learning biases in this model, the only purpose of including naive agents would be to introduce a source of random drift, which is already provided by limiting our agents' memory capacity (Spike et al., 2013).[7]

### 2.2. Simulations

We use the model to run simulations in three conditions:

- **Production pressures only:** Both the *retrieval bias* and *error bias* parameters are switched on, but *context size* is set to minimal, such that there is no inference on the Receiver's part and communication is always successful.
- **Comprehension pressures only:** *Context size* is set to maximal, requiring the Receiver to compare received signals to exemplars in all possible meaning categories to determine the Producer's intended meaning. However, both the *retrieval bias* and *error bias* parameters are switched off: all exemplars have equal probability of being retrieved for production, and errors simply replace one random segment with another random segment.
- **Competing pressures:** Both the *retrieval bias* and *error bias* parameters are switched on, and *context size* is set to maximal.

For the latter two conditions, we also test a range of different values for the Receiver's $\gamma$ parameter (which influences how deterministically they choose the meaning category that best fits the received signal). For each configuration of parameter settings, we run 10 simulations — each with a different random input lexicon and set of starting exemplars.

[3] In the simulations presented below, we use an unrealistically high $E$ of 0.5, which would imply that language users mispronounce words around half the time. Using a larger $E$ does not qualitatively change the results compared to a smaller $E$, but does allow effects to be seen in fewer time steps, which improves runtime. In any case, the function of the error mechanism is to introduce variation that can provide the fodder for lexical evolution; similar mechanisms in related models often apply to *every* production (e.g. Flego, 2022; Wedel, 2012; Wedel & Fatkullin, 2017).

[4] Using the minimal context size removes comprehension pressures from the equation entirely, since the Receiver has access to full information about the Producer's intended meaning, rendering their task trivial. A real-life analogue would be an utterance that takes place in a situation where there is only one salient possible interpretation. In our case, where communication is essentially just a process of object labelling, it could also be thought of as a Producer pointing at their intended referent.

[5] We exclude the category bias term used in the Generalised Context Model, since we want all categories to be equally likely *a priori*.

[6] $M$ is a constant in this case, since all words in our model are the same length.

[7] Here by "random drift", we mean random fluctuations in the pool of available exemplars. In our model, random deletion of an older exemplar upon storage of a new one means that the relative proportions of competing word variants will fluctuate throughout a simulation; variants that are represented by very few exemplars can even be lost entirely. In the case of vertical transmission to naive agents, any acquisition process which includes an element of random sampling can similarly lead some variants to be lost. Spike et al. (2013) showed that these two approaches are functionally equivalent.

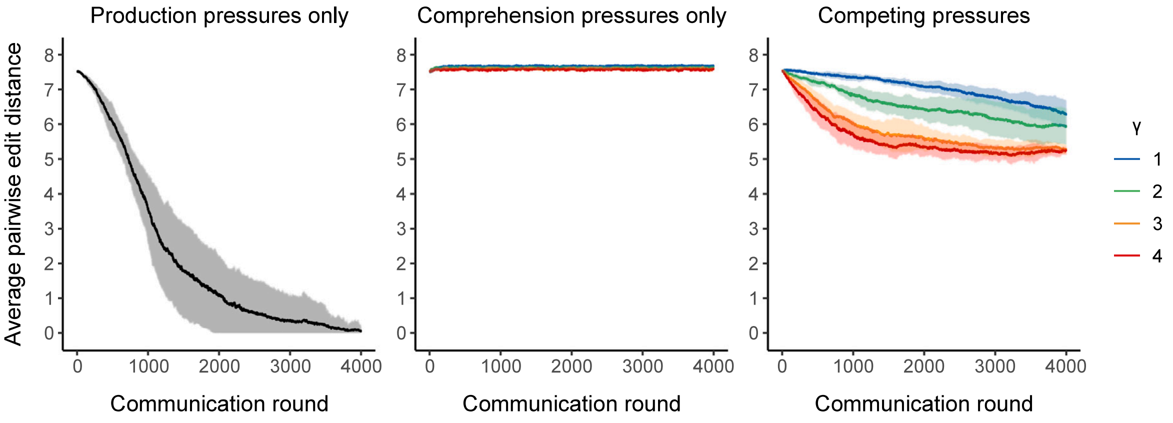

**Fig. 3.** Average pairwise edit distance over 4000 communication rounds in three conditions; lower numbers mean that words are more similar to each other. Bold lines represent the mean across 10 runs; shaded areas around these lines represent ±1 standard deviation. Colours in the two right-hand plots represent different values of the Receiver's $\gamma$ parameter, which controls the strength of the comprehension-side pressure for distinctiveness; higher values correspond to a weaker distinctiveness pressure. With production pressures alone, lexicons rapidly degenerate. With comprehension pressures alone, lexicons remain in their starting state, where words are all very different from each other. Only with competition between production and comprehension pressures does an intermediate state emerge, in which lexicons become somewhat more clustered but ultimately stabilise.

### 2.3. Results

Recall that the measure of similarity we use here is *average pairwise edit distance*, $D(L)$. When average pairwise edit distance is lower, it mean that words are more similar to each other. Fig. 3 shows the change in average pairwise edit distance over time in three conditions. When only production pressures are present, the Producer's similarity biases completely take over: lexicons become rapidly more clustered, often to the point of *degeneracy* (Kirby et al., 2015), where there is just one word for every meaning ($D(L) = 0$). Conversely, when comprehensibility is the only pressure on the language, lexicons remain very disperse over time.

When there is competition between similarity biases in production and the pressure for distinctiveness arising from communication, the result is a more balanced lexicon: words are somewhat more clustered together, but not to such an extreme degree (i.e. degeneracy) as in the production-only condition. The speed with which clustering increases depends on the strength of the comprehension-side pressure for distinctiveness, controlled by the Receiver's $\gamma$ parameter: when $\gamma$ is higher, the pressure for distinctiveness is weaker, which allows lexicons to change more rapidly. However, the curve eventually flattens out; this plateau can be thought of as the state in which words are as similar to each other as they can be whilst still allowing the Receiver to tell them apart with a reasonable level of accuracy.

Overall then, when we allow lexicons to be shaped by only one aspect of communication, the results are extreme and bear little resemblance to natural languages. Words either become so similar that they cannot be distinguished at all (production-only), or they remain totally dispersed (comprehension-only). It is only when both pressures are present — as they are in real communication — that a middle ground emerges.

#### 2.3.1. Adding frequency effects

As described in Section 1, the degree of clustering is not the same across all parts of natural language lexicons: more frequent words tend to be more similar to each other, while lower frequency words tend to be more distinctive (Frauenfelder et al., 1993; King & Wedel, 2020; Landauer & Streeter, 1973; Mahowald et al., 2018; Meylan & Griffiths, 2024). In the model results described above this effect is of course not observable, since all meanings were equally frequent. Next, we incorporate a simple notion of frequency to test whether the effect of frequency emerges from the model. Specifically, we assign 5 meanings to a high-frequency group, and the other 15 to a low-frequency group. During each round, agents communicate about the high-frequency meanings three times as often as the low-frequency meanings (three trials per agent per high-frequency meaning, versus one for the low-frequency meanings). Additionally, we increase agents' memory limit for high-frequency meanings to 30 (the memory limit for low-frequency meanings stays at 10) to capture the fact that high-frequency lexical items have stronger mental representations than their low-frequency counterparts (Alexandrov et al., 2011; Popov & Reder, 2020; see also the multiple-trace hypothesis: Hintzman & Block, 1971). The rest of the model architecture is identical.

Fig. 4 shows the change in average pairwise edit distance over time in the same three conditions as above, now additionally split by frequency. The results for the first two configurations look very similar as in Fig. 3, with no difference between frequent and infrequent words: lexicons remain in their starting state in the comprehension-only condition, and rapidly degenerate in the production-only condition. However, crucially, when production and comprehension pressures are in competition, there is a very subtle effect of frequency. Specifically, clustering increases slightly more on average in the high-frequency component of the lexicon, but only when the Receiver's $\gamma$ parameter is low; this suggests that the benefits conferred by increased frequency (due to having a stronger mental representation for higher frequency items) are washed out when the Receiver is already very proficient at telling words apart.

The effect of frequency becomes more apparent if we make two further modifications to the model architecture. First, we can modulate the strength of the producer biases such that they are stronger for higher frequency words. For example, in the case of word length, there is good evidence that speakers preferentially shorten high-frequency words (e.g. Bybee, 2002; Kanwal et al., 2017; Mahowald et al., 2013; Pierrehumbert, 2001). We can encode a similar preference to maximise ease-of-production for high-frequency items in our model by raising the activation values given by the Producer's *retrieval bias* parameter (described in Section 2.1.3) to the power of 2 when they are labelling a high-frequency meaning. This has the effect of exaggerating the preference for exemplars with high phonotactic probability. Second, we can treat high-frequency words as requiring less inference by the Receiver. The logic here is that high-frequency meanings will be weighted more highly *a priori*, so if a received signal is a good fit to a high-frequency category, the Receiver might not consider as many alternatives (note also that high-frequency words attract more attention early in processing: Dahan et al., 2001). We can operationalise this intuition by manipulating the *context size* parameter (described in Section 2.1.3): for high-frequency items, the Receiver only has to choose between 5 candidate meanings, while for low-frequency items, there are 15 candidate

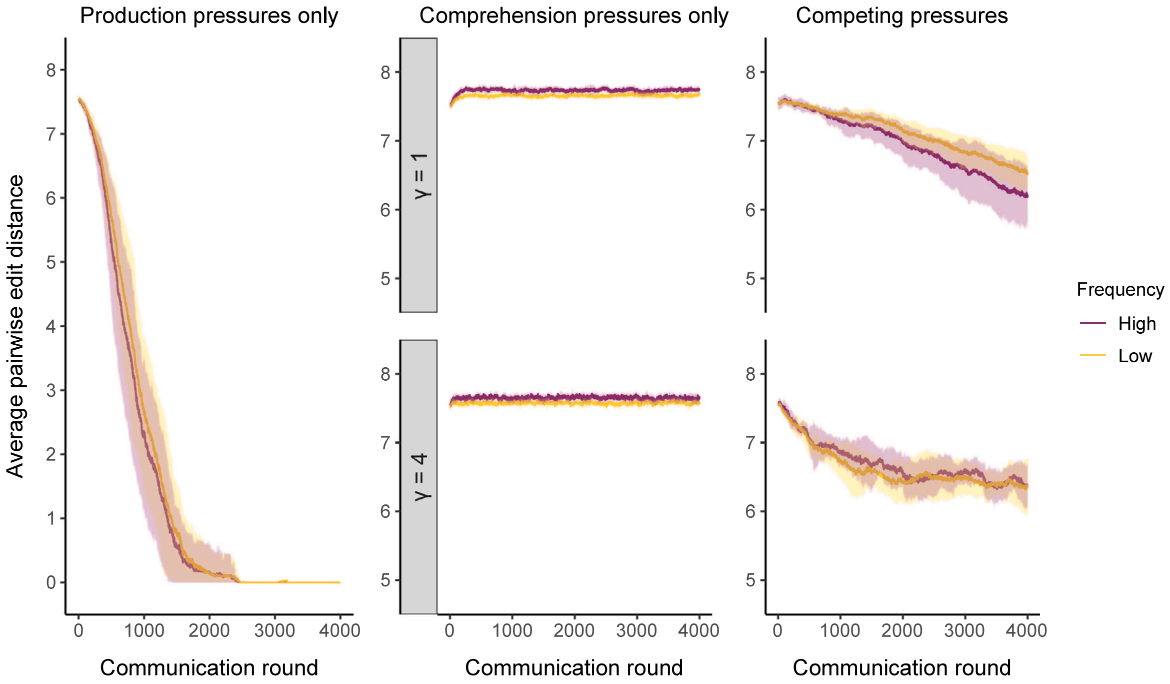

**Fig. 4.** Average pairwise edit distance for the high and low-frequency components of the lexicon over 4000 communication rounds. With only production pressures, lexicons rapidly degenerate, with no difference between frequent and infrequent words. With only comprehension pressures, both high and low-frequency words remain very distinct over time. When both production and comprehension pressures are present, a very subtle effect of frequency emerges: the high-frequency component of the lexicon becomes slightly more clustered than the low-frequency component, but only when the Receiver's $\gamma$ parameter is low (top).

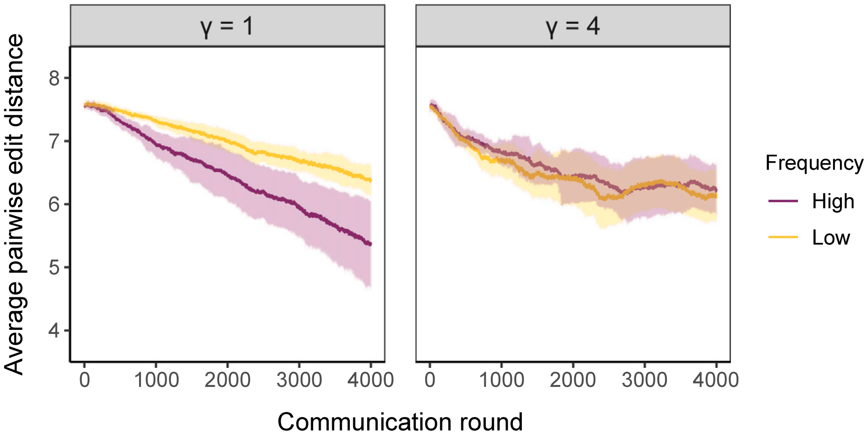

**Fig. 5.** Average pairwise edit distance for the high and low-frequency components of the lexicon when production and comprehension pressures are in competition, with two additional modifications to the model architecture: (1) Producer biases are stronger for high-frequency items, and (2) high-frequency items are more predictable for the Receiver. In this configuration, an effect of frequency is evident when the Receiver's $\gamma$ parameter is low (left), but still does not emerge for higher values of $\gamma$ (right).

meanings. Fig. 5 shows the results of this model configuration when production and comprehension pressures are in competition.[8] Here, the effect of frequency is much clearer: the high-frequency component of the lexicon becomes more clustered more quickly than the low-frequency component. However, again, this effect is only observable for lower values of the Receiver's $\gamma$ parameter.

### 2.4. Model discussion

Our model shows that phonetic clustering — a robust property of natural language lexicons — can emerge from initially random languages during repeated episodes of communication. Specifically, moderately-clustered lexicons emerge when there is competition between production pressures (which favour greater similarity between words) on the one hand, and comprehension pressures (which favour greater distinctiveness) on the other. With just one or other of these pressures, lexicons tend to fall within an extreme region of the possible design space: under the influence of production pressures alone, lexicons degenerate to the point of being communicatively useless, while when comprehension is the only pressure, lexicons remain in their initial, maximally disperse state.

Although models are always a simplification of the system they are designed to study, it is worth revisiting the specific simplifying assumptions we have made here. Firstly, as described in Section 2.1.2, we do not use a feature-based representation of the segments within a word, unlike in some similar models (e.g. Wedel, 2012). Such a model

[8] We only show this condition here since we have already established that there is no effect of frequency in the other two conditions.

architecture would probably improve the Receiver's performance, by allowing them to make more sophisticated comparisons between a received signal and their stored exemplars. However, since such fine-grained patterns of similarity do not feature in the calculations of phonotactic probability and bigram frequency that drive the Producer's behaviour, we do not think there would be significant downstream consequences for the eventual outcome of the model. Rather, clustering would likely just emerge *faster* since greater success on the Receiver's part results in more frequent storage of new exemplars and quicker turnover of old exemplars. In any case, corpus analysis suggests that a feature-based representation is unnecessary to explain the degree of clustering in natural language lexicons (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017), which is the basis on which we made this simplification.

Furthermore, whilst successful communication changes the agents' internal representation, there is no such feedback loop from unsuccessful communication in the model. This is a common feature of exemplar models in this tradition (e.g. Wedel, 2012; Wedel & Fatkullin, 2017), since there is no penalty on unsuccessful signals (beyond not being stored in the target category) encoded within the Generalised Context Model of signal reception (Nosofsky, 1986, 2011). However, other frameworks exist that could capture the intuition that language users might try not to use variants that they do not believe to be communicatively useful. For example, various types of models employ some kind of negative feedback after unsuccessful interactions, either deletion or inhibition as in reinforcement models (e.g. Barrett, 2006; Franke & Jäger, 2012; Skyrms, 2010) or weakening associations as in the Naming Game (Steels, 2012; Steels & Loetzsch, 2012); for further discussion of these mechanisms, see Spike et al. (2017). However, the decision about how to implement such mechanisms is not straightforward, especially in the case of signals containing errors whereby there is no exactly matching exemplar in either agents' internal representation that could be targeted. An alternative to penalising signals after communication has failed is to downweight signals that are more likely to result in failure *before* an interaction takes place, as in the Rational Speech Act (Frank & Goodman, 2014; Goodman & Frank, 2016); in such models, a pragmatic speaker reasons about how likely a listener would be to recover the intended meaning from the different utterances available to them. The downside of this kind of mechanism is that it requires a significant amount of computation in every communication episode, dramatically increasing the runtime of the models. Listener-oriented approaches have also been criticised as teleological (e.g. Wedel, 2006). In any case, we would argue that either of these approaches adds unnecessary complication to the model; selection of successful signals works by itself, it simply takes slightly longer to turn over less useful signals.

Finally, it is true that comprehension does not straightforwardly favour word dissimilarity, as suggested by our model of reception: specifically, increases in phonotactic probability have been found to facilitate word recognition (Vitevitch & Luce, 1998). However, pure recognition — in terms of deciding whether a received stimulus is familiar (word) or unfamiliar (non-word) — is very different from the categorisation task faced by our agents, a task where competition between multiple activated referents is known to inhibit processing (Luce & Pisoni, 1998). Indeed, Vitevitch and Luce (1998) describe the effect of phonotactic probability as facilitative for sub-lexical processing (for example, segmenting the speech stream, or processing novel sound sequences) and inhibitory for lexical processing (for example, determining the intended meaning of a received signal, as in our model). Wedel (2012) also points out that the general behaviour of these exemplars models is the same whether similarity biases are encoded once (in production) or twice (in production and perception).

Returning to the frequency effects discussed in Section 2.3.1, our results suggest that frequency may modulate the rate of lexical evolution, with the effect depending to some extent on the assumptions we make about the processing consequences of frequency. In the most basic version of our frequency manipulation, we implicitly assume that production biases are underlyingly frequency-*in*dependent. In other words, the model architecture is such that producers want to maximise production ease across the board; frequency-dependent lexical evolution emerges simply because they can get away with doing so more for high-frequency items. The fact that frequency effects are so subtle under this assumption makes sense when we examine how frequency actually impacts the two participants in a conversation. From the comprehender's side, a frequency advantage is baked into the reception mechanism (Eq. (2)): the stronger mental representation of high-frequency items (due to their larger memory limit) increases the Receiver's certainty that a received signal maps onto a target category. However, from the producer's side, any selection which may be acting to change a word's form is competing against the fact that the representation of the word's existing form is very strong; this may also be why, for example, high-frequency irregular items tend to resist regularisation (e.g. Bybee, 1995; Cuskley et al., 2014; Sims-Williams, 2022; Smith et al., 2023; Wu et al., 2019). Therefore, while comprehension may permit greater clustering for high-frequency items, the production process may be slower to generate the variation required for selection to act upon for these items. A stronger effect of frequency can emerge from the model under certain conditions, but of course, it may not be desirable to make the additional assumptions required to generate this result (Marquet et al., 2014). Future work could expand upon the frequency aspect of our model, for example, by using a more realistic distribution of word frequencies (i.e. following a power law) rather than treating frequency as a binary value.

Overall though, our model predicts that production or comprehension pressures in isolation will give rise to lexicons at one extreme of clustering or the other. An intermediate state, with levels of clustering more similar to those found in natural language lexicons, should emerge when these pressures are in competition. In the next section, we simulate these same pressures in a communication experiment with human participants, focusing more specifically on the interaction between clustering and frequency.

## 3. Communication experiment

We use an artificial language learning paradigm to investigate how production and comprehension pressures trade-off against each other to influence language users' lexical choices during communication. The experiment is inspired by Kanwal et al. (2017), who showed that Zipf's Law of Abbreviation (Zipf, 1949) emerges from precisely such a trade-off. Specifically, in their experiment, participants were trained on a miniature lexicon in which two objects that differed in frequency were labelled with either a unique, long label ("zopudon" or "zopekil") or a shared (and therefore ambiguous) short label, "zop". Kanwal et al. found that participants favoured the ambiguous short label (which was quicker to produce) under time pressure, and the unambiguous long labels under pressure for accuracy. When both of these pressures were present, participants converged on an optimal solution, whereby the short label was consistently mapped to the high-frequency object and the long label to the low-frequency object, consistent with the Law of Abbreviation. By simulating the pressures inherent to real communication, this method provides a convenient way to disentangle the individual effects of opposing pressures, and to show that key structural properties of natural languages can emerge from their confluence.

Following Kanwal et al., rather than relying on participants to introduce changes to the lexicon themselves — i.e. make errors in production — we designed a lexicon incorporating lexical variation. However, the competitors in our experiment are words from different phonological neighbourhoods, rather than words of different lengths. Specifically, each object was labelled by two different words: one from a high-density neighbourhood (highly confusible with words belonging to other meanings), and one from a low-density neighbourhood (highly dissimilar from all other words in the language). As in Kanwal et al., participants were trained on the different names for two objects that differed in frequency, and were then paired up to play a communication game, during which we manipulated the presence or absence of a production-side pressure for similarity (Stemberger, 2004; Vitevitch & Luce, 2005; Vitevitch & Sommers, 2003) and a comprehension-side pressure for distinctiveness (Chan & Vitevitch, 2009; Luce & Pisoni, 1998). We predicted that natural-language-like properties would arise only when both these pressures were present.

## 3.1. Methods

The study was approved by the PPLS Ethics Committee at the University of Edinburgh and was pre-registered with the Open Science Foundation (https://osf.io/jucn6).

### 3.1.1. Materials

The meaning space consisted of two objects — a compass and a lightbulb — represented by drawings from the MultiPic databank (Duñabeitia et al., 2018). The two drawings score very similarly for visual complexity (2.65 and 2.41 respectively, on a scale from 1 to 5). To investigate the role of frequency on clustering, one object (randomly chosen for each participant) appeared three times more frequently than the other throughout the experiment. The language consisted of four artificial CVC words: “zun” [zʌn] and “zan” [zæn] (the *high neighbourhood density* words; henceforth, HND) and “mig” [mɪg] and “tep” [tɛp] (the *low neighbourhood density* words; henceforth, LND). The artificial words are matched for neighbourhood density in English (56 ± 1) according to the CELEX corpus (Baayen et al., 1995) and have average positional phoneme probability ranging between 0.0498 and 0.0583 according to the Irvine Phonotactic Online Dictionary (Vaden et al., 2009). We designed the words in this way to ensure that any preference for either HND or LND words would be driven only by their status within the artificial language, not by their relationship to participants' native English. Audio files for each word were synthesised using an online IPA to Speech tool (https://www.antvaset.com/ipa-to-speech). For each participant, each object was randomly assigned two names: one from each neighbourhood. Unlike in Kanwal et al. (2017), the competitor labels for an object were therefore not variants of a single word (e.g. “zopudon” → “zop”), but two completely different words. We designed the lexicon in this way to maximise the distance between the LND words: any words that were more clearly derived from the HND words would necessarily also be quite similar to each other, reducing their distinctiveness. Finally, to ensure that every possible object–word mapping was phonotactically equivalent, the language was designed such that neighbourhood membership precisely determines a word's relationship with every other word in the lexicon. Concretely, HND words have one neighbour each (the other HND word) and average edit distance of 2.33 from any other word (edit distance between the two HND words is 1; edit distance between either HND word and either LND word is 3). LND words, meanwhile, have *no* neighbours, and edit distance of exactly 3 from every other word.

### 3.1.2. Procedure

The experiment was written in JavaScript using the jsPsych library (de Leeuw et al., 2023). The design is based on the paradigm developed by Kanwal et al. (2017). A schematic of the experimental design and procedure is given in Fig. 6. Participants completed the following phases, in the order shown below.

*Training.* On each training trial, an object was presented on screen alone for 1000 ms while the audio file of the appropriate word played once. The orthographic form of the word then appeared below the image in the English frame ‘This is a …‘. After another 1500 ms, a ‘next’ button appeared to let participants advance to the next trial. Participants completed 24 training trials: 18 for the frequent object, and 6 for the infrequent object. Each object appeared half the time with its HND word and half the time with its LND word. The order of training trials was randomised for each participant.

*Pre-test.* After the training phase, participants were tested on their knowledge of the language. On each trial, participants were presented with a word from the artificial language in the English frame ‘Which of these is the …?’ and asked to choose between the two objects. They received full feedback on their response. Again, participants completed 24 trials, with the same distribution over frequent/infrequent meanings and HND/LND words as in training. The order of trials was randomised for each participant. Participants were required to reach at least 83% accuracy (i.e. ≥ 20 trials correct) to proceed to the interaction phase. Additionally, two attention checks were randomly interspersed within this phase. On these trials, participants saw a familiar English word in the same ‘Which of these is the …?’ frame, along with two previously unseen pictures. They received no feedback on their response to these trials. Participants were required to pass at least one of these attention checks to proceed to the interaction phase.

*Interaction.* The interaction phase of the experiment was managed via a Python WebSockets server (based on code from https://kennysmithed.github.io/oels2023/).[9] At the start of the interaction phase, participants were put into a virtual waiting room ready to be paired with the next participant who completed the pre-test. An on-screen timer kept participants informed of how long they had been waiting. If participants were not paired with a partner within 5 min, they were removed from the waiting room and paid for their time.

Once participants were paired, they played a communication game. Participants were instructed that they had two goals: to score as many points as possible (i.e. the *accuracy* pressure in Kanwal et al., 2017) and to complete the game as quickly as possible (i.e. the *time* pressure in Kanwal et al., 2017).

On each trial, one participant acted as the Director and the other as the Matcher; roles alternated between every trial. The Director was shown an object and asked to name it for their partner. An on-screen stopwatch tracked how long the Director took to complete this task (to reinforce the pressure for speed). The Director was always given both object names as options, but the method of producing a word differed between conditions, as outlined below. The Matcher was shown the word sent by the Director (with or without noise depending on condition; see below) and asked to choose which object they thought their partner was describing. Both participants received feedback as to whether the Matcher chose the correct object (to reiterate the pressure for accuracy). Participants completed 24 trials as Director and 24 as Matcher, with the same distribution over frequent/infrequent meanings as in training. The order of each participant's Director trials was randomised. At the end of the interaction phase, both participants were shown their pair's final score and overall completion time.

To avoid having to ensure that participants were trained on the same version of the input language (since the assignment of objects to frequencies and words to objects was randomised for each participant), participants' responses were translated via a shared underlying representation before being transmitted, following a similar method to that used by Smith et al., 2025. Specifically, if the object being labelled by the Director was the high-frequency object in their training set, then the target object (i.e. correct answer) for the Matcher would be whichever object was the high-frequency object in *their* training set. Similarly, if the Director sent the HND word that they were trained on for their target object, then the Matcher would see the HND word that *they* were trained on for *their* target object (i.e. the object of the same frequency as the object seen by the Director). This procedure is illustrated in Fig. 7. Importantly, because of the way the artificial language was designed, this translation procedure never resulted in the Matcher seeing a word that had different phonotactic properties (with respect to the rest of the lexicon) than the one the Director had sent: see Section 3.1.1.

Each pair was randomly assigned to one of the three experimental conditions. There were two different versions of the Director and

[9] Full code for the experiment is available at https://osf.io/vsy6z/.

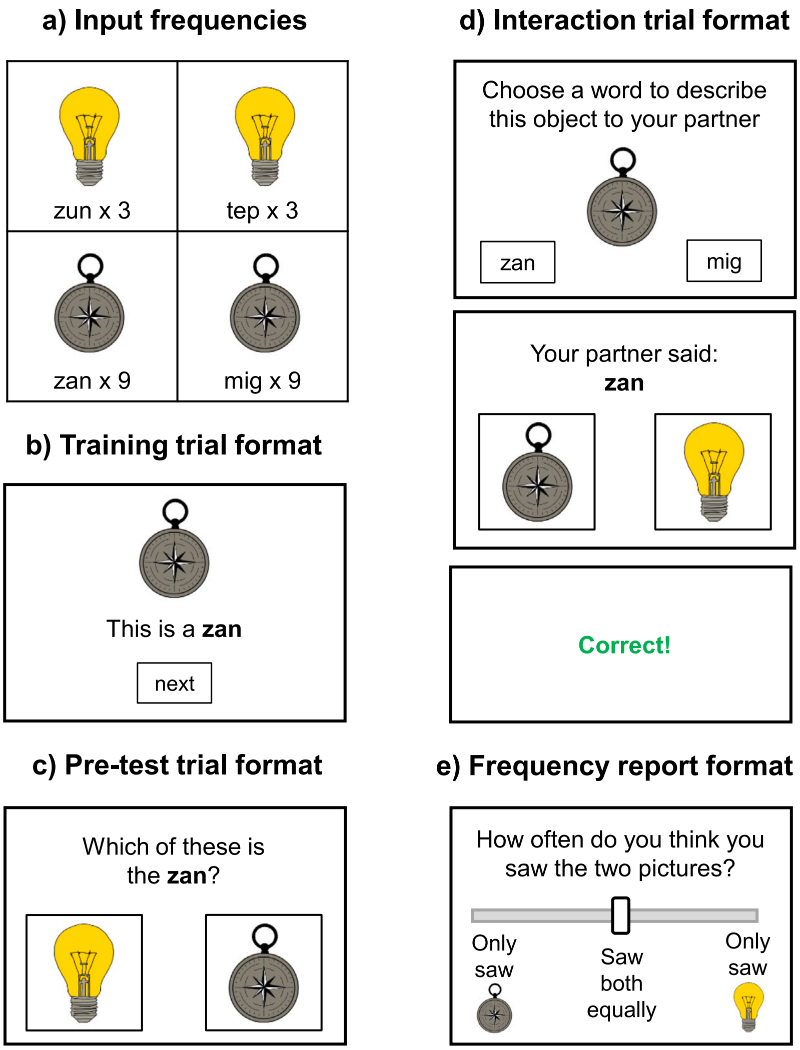

**Fig. 6.** Schematic of the experimental design and procedure. (a) Example training set (the exact permutation of objects and labels was randomised for each participant) showing the 75/25 frequency distribution over the two objects (rows) and 50/50 distribution over HND and LND words (columns). (b) Example training trial. (c) Example pre-test trial. (d) Example interaction trial, proceeding from a Director trial (top) to a Matcher trial (middle) and then feedback to both participants (bottom). (e) Example frequency report trial.

Matcher trials — an easy version, and a more difficult version — depending on condition. In the Production condition, Director trials were difficult but Matcher trials were easy. In the Comprehension condition, it was the other way around: Matcher trials were difficult but Director trials were easy. In the critical Combined condition, both tasks were difficult. Specifically, the manipulations were as follows (also illustrated in Fig. 8):

- **Easy Director trials:** The Director was presented with both word options for the target object (in a random order) and simply asked to click on the word they wished to send.
- **Difficult Director trials:** The Director was presented with both word options for the target object (in a random order) and asked to use a 3 × 6 on-screen keyboard to type one of the words. They were only able to transmit one of the valid words; if they submitted a word that didn't exist in the artificial language, or that referred to the other object, they were asked to try again.[10] The letters required to make an HND word ("z", "u", "a" and "n") always appeared in the same positions in the centre of the keyboard. The letters required to make an LND word ("t", "e", "p", "m", "i" and "g"), along with six other distractor letters that were not used in the artificial language, appeared around the outside of the keyboard and changed positions on every trial. Additionally, the central four buttons were three times as large (both in area and in font size) as the outer buttons. In this way, HND words were easier to produce than LND words. This design was intended to simulate the idea that, in spoken word production, frequently-used phonemes are pronounced more quickly and accurately, while less frequently-used phonemes present more of a moving target for pronunciation (Goldrick & Larson, 2008; Goldrick & Rapp, 2007; Munson, 2001; Vitevitch et al., 2004).
- **Easy Matcher trials:** Transmission was clean, and the Matcher was presented with the full word sent by the Director (after any necessary translation; see above).

[10] We included this restriction for two reasons. Firstly, the translation procedure illustrated in Fig. 7 would only work if it was possible to definitively map participants' responses to categories from the input language. And secondly, the Matcher in the Comprehension condition would always see a valid word since the Director had no freedom to invent new forms, so we wanted to ensure that this aspect was parallel across conditions.

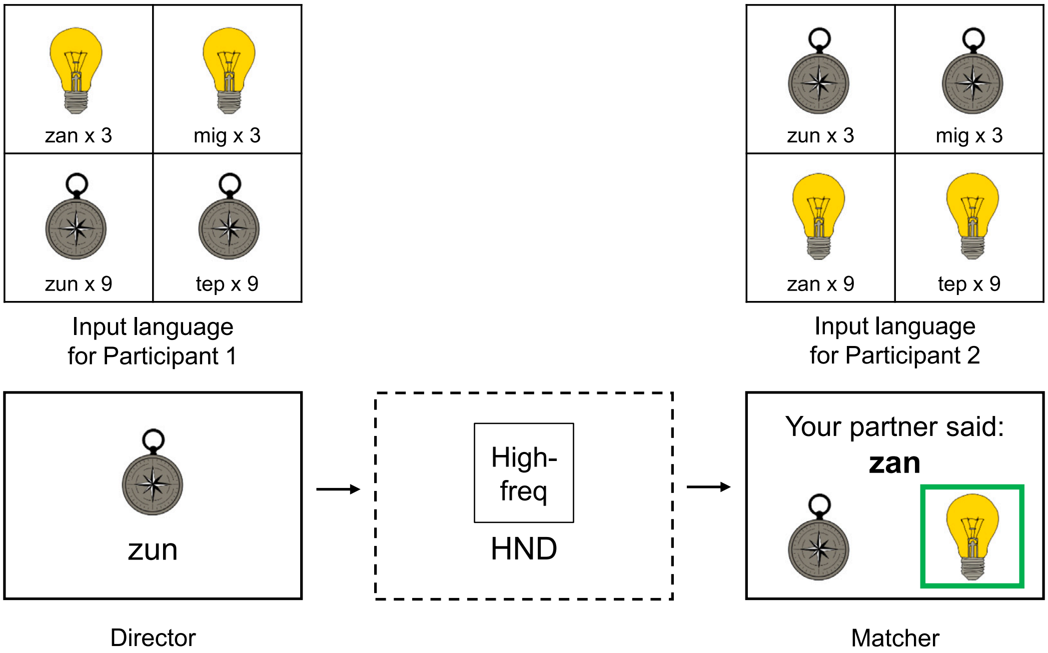

**Fig. 7.** Example of the procedure for transmitting responses in the interaction phase between two participants who were trained on a different random permutation of the input language. The Director sees the compass (which was the high-frequency object in their training set) and sends the word "zun". This is first translated into an underlying representation whereby objects are represented by their frequency and words by their neighbourhood, rather than either being associated with specific forms. This underlying representation is then used to determine which word form to show the Matcher and which object should be the target; in this case, the lightbulb is the target object since this was the high-frequency object in the Matcher's training set, and its associated HND word is "zan".

- **Difficult Matcher trials:** Transmission was noisy, and the Matcher was presented with only the first letter of the word sent by the Director (after any necessary translation; see above). One letter provided enough information to distinguish between the LND words, but this information loss rendered the HND words identical and therefore ambiguous between the two objects. This design was intended to simulate the idea that, in spoken word perception, words with many neighbours activate many candidate meanings, and are thus more likely to be misinterpreted, while more distinctive words are more likely to activate only the target meaning (Chan & Vitevitch, 2009; Luce & Pisoni, 1998).

*Frequency report.* Once participants completed the interaction phase, they were asked to complete one final task individually. This task was included as a sense check that participants had noticed the frequency imbalance between the two objects. Participants were presented with a continuous slider over percentages and asked "How often do you think you saw the two pictures? Did you see one more than the other?". The slider was accompanied by three labels: "Only saw *Object 1*" at one end, "Saw both objects equally often" in the middle, and "Only saw *Object 2*" at the other end. Which object appeared at which end of the slider was randomised for every participant.

### 3.1.3. Participants and exclusions

We used Prolific to recruit 220 adults resident in the UK who self-reported that their first language was English and that they had no known language disorders. They were provided with a downloadable information sheet and gave informed consent to participate. The experiment took around 20 min to complete in full (median time = 17:46), for which participants were paid £3.50 (above UK National Minimum Wage at the time of running the experiment). Seven participants were prevented from proceeding to the communication game due to low accuracy on the pre-test[11]; these participants were paid a reduced rate of £1.75. 27 participants started but failed to complete the interaction phase (either due to technical difficulties during the communication game or because they timed-out of the waiting room before being paired with a partner); these participants were paid a variable rate depending on how far they had got through the experiment. Six participants (one pair in each condition) completed the communication game and were paid the full rate, but their data was excluded from analysis because their completion time was more than 3 standard deviations above the median in that condition. We also pre-registered that we would exclude data from participants who admitted to taking written notes in a debrief questionnaire; no participants were excluded on this criterion. After all exclusions and dropouts, we were left with 30 pairs in each condition: a total of 180 individual participants.

### 3.1.4. Predictions

We predicted that participants in the Production condition, where HND words were easier to produce than LND words, would tend to use the HND word for both objects, regardless of frequency. By contrast, we predicted that participants in the Comprehension condition, where noisy transmission meant that HND words (but not LND words) became indistinguishable, would tend to use the LND word for both objects, regardless of frequency. We predicted that we would observe a natural-language-like frequency trade-off in the critical Combined condition, where both these pressures were present, such that participants would consistently map the frequent object to the HND word and the infrequent object to the LND word. This is the optimal strategy by which to minimise production effort (and therefore complete the game as quickly as possible) but still maintain an unambiguous one-to-one form-meaning mapping (and therefore score as many points as possible).

## 3.2. Results

### 3.2.1. Confirmatory analysis

Fig. 9 shows the proportion of trials on which each pair used the HND word on Director trials, split by object frequency and condition.

[11] All participants passed both attention checks, so these exclusions were all due to low accuracy on critical trials.

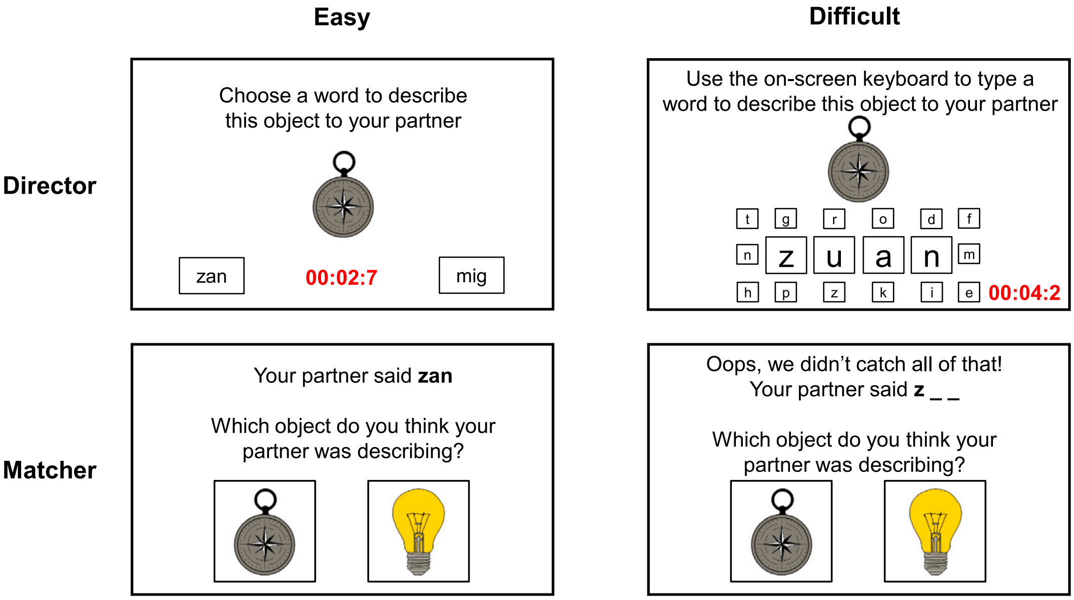
Easy
Difficult
Director
Matcher
Choose a word to describe this object to your partner
zan
00:02:7
mig
Use the on-screen keyboard to type a word to describe this object to your partner
t g r o d f
n z u a n m
h p z k i e
00:04:2
Your partner said **zan**
Which object do you think your partner was describing?
Oops, we didn't catch all of that! Your partner said **z _ _**
Which object do you think your partner was describing?

**Fig. 8.** Easy (left) and more difficult (right) versions of the Director (top) and Matcher (bottom) tasks. When the tasks are easy, HND and LND words are similarly easy to produce and comprehend. When the tasks are difficult, there is a production-side pressure in favour of HND words, which are made up of more accessible segments, and a comprehension-side pressure in favour of LND words, which are able to overcome the noise on transmission.

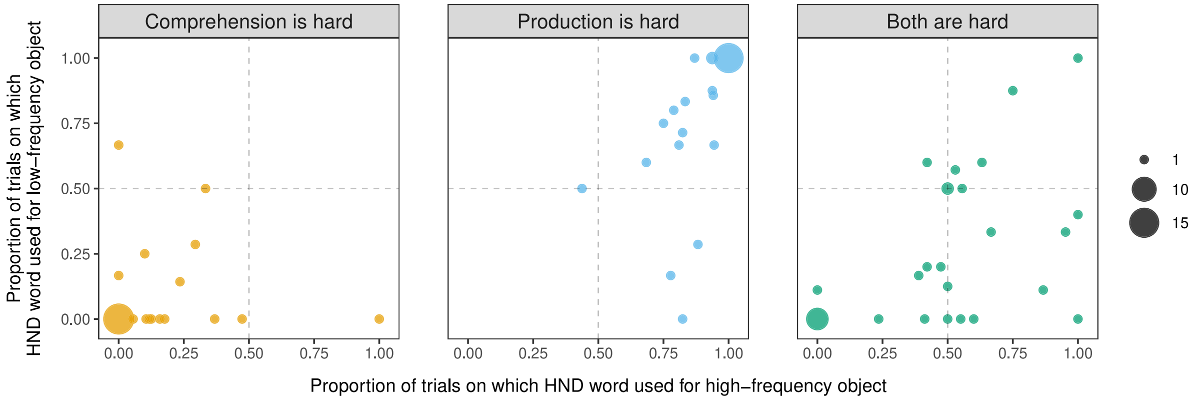

**Fig. 9.** Proportion of trials on which the HND word was used for the high-frequency object vs. the proportion of trials on which it was used for the low-frequency object. Each data point combines a pair of communicating players, representing the sum of their Director trial productions. As in Kanwal et al. (2017), only data from the second half of each pair's interaction trials is shown, as participants were more likely to have converged on a stable mapping by this time. Data points in the bottom left quadrant indicate pairs who are mostly using the LND words for both objects; participants are clustered in this quadrant in the Comprehension condition (left), where only the LND words are reliably distinguishable and there is no countervailing pressure from production in favour of the HND words. Data points in the top right quadrant indicate pairs who are mostly using the HND words for both objects; participants are clustered in this quadrant in the Production condition (middle), where HND words are considerably easier to produce than LND words and there is no countervailing pressure from comprehension in favour of the LND words. Data points in the bottom right quadrant indicate pairs who are mostly using the HND word for the frequent object and the LND word for the infrequent object. This behaviour, consistent with the frequency trade-off seen in natural languages, is numerically most common in the critical Combined condition (right), where both production and comprehension pressures are at play, but a range of other behaviours are also represented in this condition.

As predicted, most participants in the Comprehension condition used the LND word for both objects, while in the Production condition, most participants used the HND word for both objects. In the critical Combined condition, where the HND words were considerably easier to produce for the Director but functionally ambiguous for the Matcher, participants adopted a range of strategies. Some arrived at the optimal strategy described in Section 3.1.4. However, many were willing to expend extra time and effort to use the LND words for both objects and thus ensure accurate communication, while others opted to use the HND words for both objects and thus minimise transmission time at the expense of perfect accuracy.

We used the `lme4` package (Bates et al., 2015) in R (R. Core Team, 2024) to fit a logistic mixed effects model to the data, with a binary dependent variable of HND word use (as contrasted with LND word use, i.e. 1 if the participant produced the HND word, 0 if they produced the LND word). The model included fixed effects of experimental condition (treatment-coded with the Comprehension condition as the reference level), object frequency (treatment-coded with low-frequency as the reference level) and their interaction, and nested by-participant and by-pair random intercepts and random slopes for object frequency.[12] As in Kanwal et al. (2017), only data from the second half of each participant's Director trials was included in the model, as pairs were more likely to have converged on a stable mapping by this time. The model reveals that participants in the Comprehension

[12] Model formula: `HND word ~ condition + frequency + condition:frequency + (frequency | pair/participant)`.

**Table 1**
Summary of fixed effects for a logistic mixed effects model with HND word use as the binary dependent variable, and nested by-participant and by-pair random effects for object frequency. The predicted effects are shown in bold. Coefficient estimates are on the log-odds scale.

| | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| **intercept (object = infrequent, condition = Comprehension)** | **−8.075** | **1.590** | **−5.078** | **<0.001** |
| object = frequent | 0.807 | 1.707 | 0.473 | 0.636 |
| **condition = Production** | **14.024** | **2.526** | **5.553** | **<0.001** |
| condition = Combined | 3.893 | 1.434 | 2.714 | <0.01 |
| object = frequent & condition = Production | 0.582 | 2.787 | 0.209 | 0.835 |
| **object = frequent & condition = Combined** | **1.689** | **1.458** | **1.158** | **0.247** |

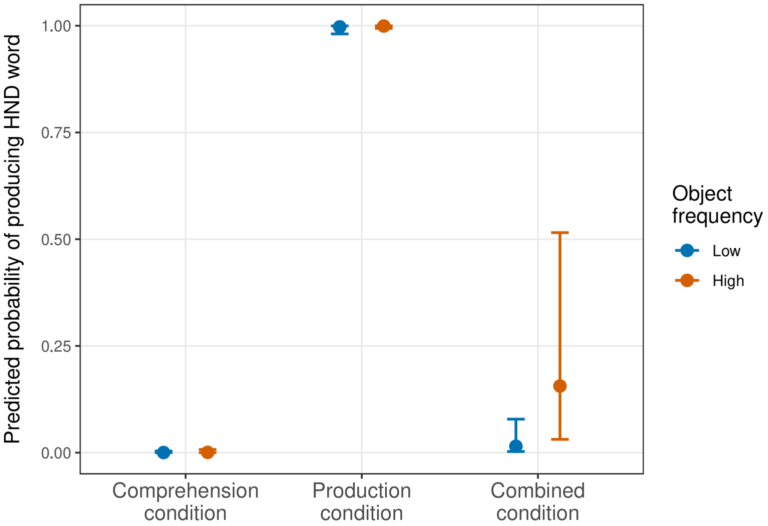

**Fig. 10.** Model predictions for each combination of condition and object frequency, generated using the `ggeffects` package (Lüdecke, 2018). Points represent the predicted probability of producing an HND word; error bars represent the 95% confidence interval around this value. Although the model predicts that participants in the critical Combined condition were numerically more likely to produce an HND word for the high-frequency object than the low-frequency object, this interaction between condition and frequency was not statistically significant (see Table 1).

condition were very unlikely to use the HND words for either object, while participants in the Production condition were very likely to use the HND words for both objects. The predicted interaction between condition and frequency was not statistically significant, meaning that there is insufficient evidence to conclude that participants in the critical Combined condition were displaying a frequency trade-off in their use of HND vs. LND words. However, there was a significant main effect of condition, such that participants in the Combined condition were more likely *overall* to use the HND words than participants in the Comprehension condition. A full summary of model coefficients is given in Table 1. The model's predictions for each combination of condition and object frequency are shown in Fig. 10.

### 3.2.2. Exploratory analysis

Fig. 9 suggests that when only one aspect of the communicative task was difficult, most participants took the same approach to mitigating this difficulty: data points are strongly clustered in the bottom-left and top-right corners in the Comprehension and Production conditions respectively. By contrast, when both aspects of the task were difficult, it is less clear that participants were converging on a single optimal solution: data points are more widely scattered around the plot in the Combined condition. In particular, there are a number of points towards the centre of the plot (on at least one axis) in this condition, representing pairs who appear to be probability matching to the input by using the HND and LND words approximately 50% of the time each (for at least one object). However, this method of visualisation disguises some underlying differences between the two members of the pair. Specifically, while it is possible that a pair at the centre of this plot could consist of two participants probability matching to the input, it is equally possible that these points represent pairs where one participant is *only* using the HND words and the other is *only* using the LND words. Indeed, if we plot individual participants instead of collapsing across pairs, we can see that the data tends to move away from the centre and towards the corners (Fig. 11).

To further explore this trend, we calculated a convergence score for each pair by comparing the languages produced by each member of the pair. Each participant's output language can be fully described by a 2-dimensional vector (*HF, LF*) where *HF* is the proportion of trials on which the participant used the HND word for the high-frequency object and *LF* is the proportion of trials on which they used the HND word for the low-frequency object. For example, the vector $(1,0)$ captures a language showing the expected frequency trade-off (i.e. in the bottom-right corner of the plot). The *divergence* between two members of a pair is given by the Euclidean distance $e$ between their output languages. The maximum possible Euclidean distance between two $n$-dimensional vectors is equal to $\sqrt{n}$ when the input values are bounded between 0 and 1. Therefore, the *convergence* between two members of a pair is given by $\sqrt{2} - e$. Fig. 12 shows the distribution of convergence scores by condition. We fit a linear regression model to this data, predicting convergence score as a function of experimental condition (treatment-coded with the Comprehension condition as the reference level). The model reveals that within-pair convergence was significantly lower in the Combined condition ($\beta = -0.407$, $SE = 0.107$, $t = -3.804$, $p < 0.001$), while there was no significant difference between the Comprehension and Production conditions ($\beta = -0.073$, $SE = 0.107$, $t = -0.682$, $p = 0.497$).

Since pairs in the Combined condition are often failing to converge on a shared language, we might also expect accuracy on Matcher trials to be lower in this condition. Fig. 13 shows how often the Matcher successfully selected the target object in each condition, depending on the object's frequency and the word used to label it. We fit a logistic

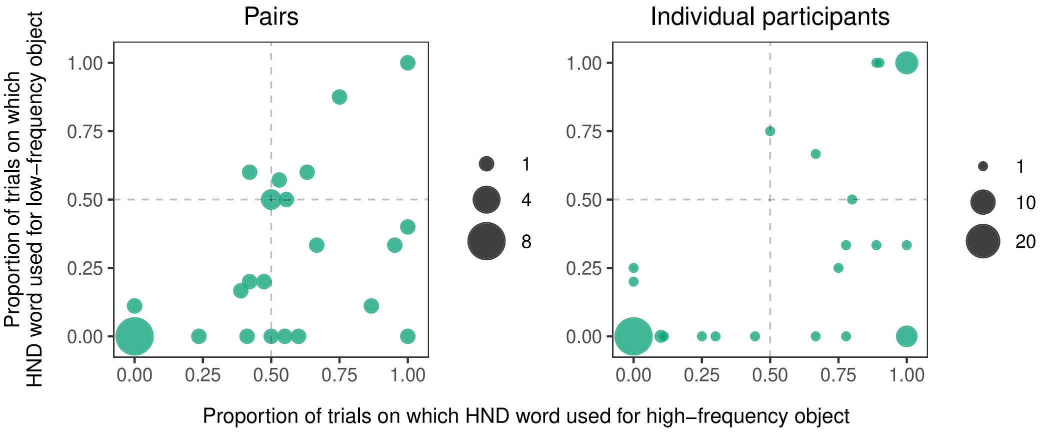

**Fig. 11.** By-pair (left) vs. by-participant (right) data for the Combined condition. Although it appears that a number of pairs are producing HND and LND words with roughly equal frequency, it is clear that individual participants are at least somewhat consistent in their choice of word. This suggests that pairs towards the centre of the left-hand panel have not converged on a shared language; rather, these pairs probably consist of one participant who is mostly using the HND words for both objects and one who is mostly using the LND words for both objects.

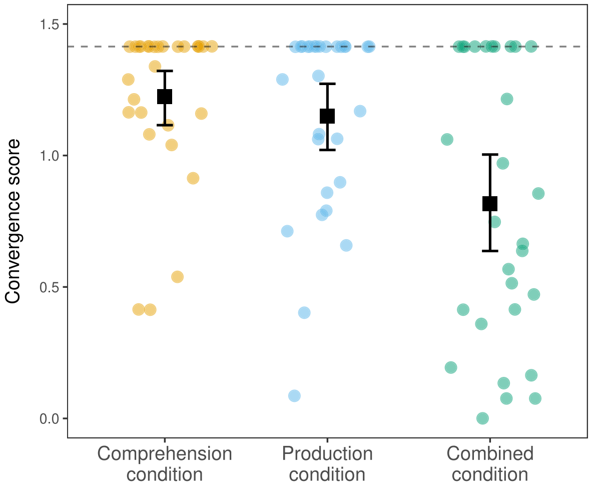

**Fig. 12.** Convergence scores by condition. The dashed line indicates the maximum possible score, which is achieved when both members of a pair produce exactly the same output language. Each coloured point represents an individual pair. Black points represent the mean over all pairs in that condition; error bars represent bootstrapped 95% confidence intervals over the mean. Convergence scores are similarly high in the Comprehension and Production conditions, but significantly lower in the Combined condition.

mixed effects model to this data, predicting accuracy as a function of experimental condition (treatment-coded with the Comprehension condition as the reference level), word type (treatment-coded with LND as the reference level), object frequency (treatment-coded with low-frequency as the reference level), and all two-way and three-way interactions between them. The model also included by-participant random intercepts, but failed to converge with random slopes for object frequency or nested random intercepts by-participant and by-pair. There was no main effect of being in the Combined condition ($\beta = -0.389$, $SE = 1.120$, $t = -0.347$, $p = 0.728$). However, the model yielded a significant three-way interaction between condition, frequency and word type, such that the probability of a correct response was higher in the Combined condition when the target object was high-frequency and labelled with the HND word ($\beta = 4.136$, $SE = 1.607$, $t = 2.574$, $p < 0.05$).

This three-way interaction could indicate that participants had some expectations of a natural-language-like frequency trade-off in comprehension (even if this was not borne out in their productions). Specifically, participants were relatively successful at inferring their partner's intended meaning when an HND word was used to label the high-frequency object, even though the information provided by the word form alone could equally point to either object. Conversely, participants were very unlikely to infer that their partner was referring to the low-frequency object when they used an HND word. However, it is difficult to determine whether this discrepancy only arises in the Combined condition because participants in this condition understand that there are pressures in favour of both HND and LND words and therefore form different expectations about how their partner might be behaving, or because this is the only condition where both word types are used frequently enough to observe a difference between them. In other words, it may be that accuracy for HND words only appears to be similar across the two object frequencies in the Comprehension condition because these words are hardly ever used for either object.[13] If this is the case, then accuracy for HND words in the Combined condition may simply reflect a strategy of guessing meanings proportional to their frequency when the signal is ambiguous (i.e. guess the high-frequency meaning 75% of the time and the low-frequency meaning 25% of the time).

### 3.3. Experiment discussion

In our experiment, we found that language users were easily able to adapt their lexical choices for efficient communication when *only* production was difficult or *only* comprehension was difficult. However, the picture was less clear when both of these pressures were present. Some participants converged on the efficient natural-language-like solution: mapping easy-to-produce but potentially ambiguous words to frequent objects and harder-to-produce but easily distinguishable words to infrequent objects. However, other participants apparently prioritised one pressure over the other, either by using only the unambiguous LND words despite their cost in production, or by using only the easily accessible HND words despite their cost in comprehension. Nonetheless, as in our model, the lexicons that emerged when production and comprehension pressures were in competition represented an intermediate state between the extreme outcomes observed when only one of these pressures was at play, at least in terms of the *overall* likelihood of producing an HND word.

[13] Accuracy in the Production condition is, unsurprisingly, at ceiling across the board, since the clean transmission channel in this condition ensures that all words are unambiguous.

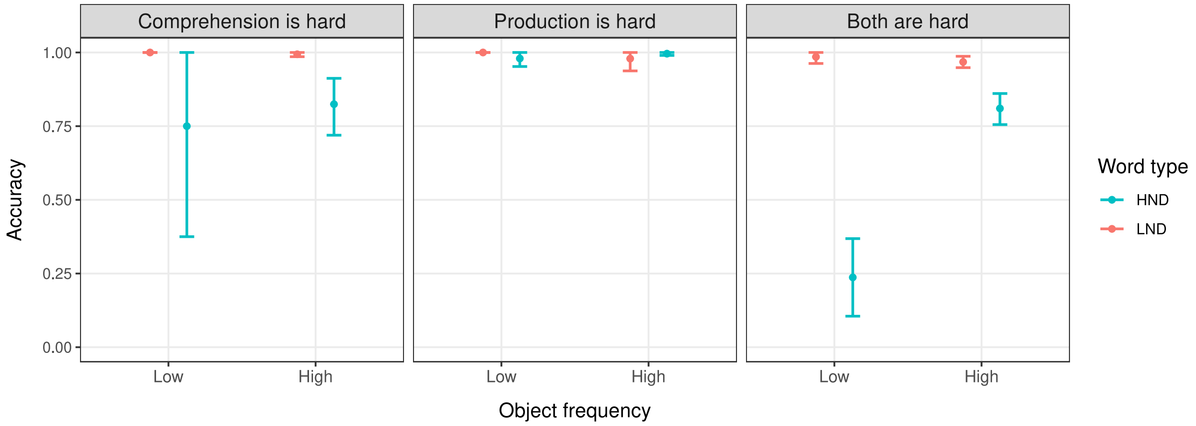

**Fig. 13.** Accuracy on Matcher trials by condition, object frequency and word type. Accuracy is high across the board for LND words, which are always unambiguous. Accuracy for HND words depends both on condition and object frequency: participants in the Combined condition are significantly more likely to successfully infer the intended meaning of these words when they are used to label the high-frequency object than when they are used to label the low-frequency object, suggesting that participants in this condition may have some expectations of a natural-language-like frequency trade-off when interpreting ambiguous signals.

Notably, this experiment was designed as a relatively close replication of Kanwal et al. (2017). Although the exact production and comprehension pressures we simulate are not identical, the net effect of these pressures was very similar: LND words (like long words in Kanwal et al.) took longer to produce, and HND words (like short words in Kanwal et al.) were ambiguous in communication. Despite these parallels, we do not replicate the frequency trade-off that arose in Kanwal et al.'s Combined condition. In considering why our findings did not robustly bear out our predictions, it is worth laying out what might have led to this discrepancy.

Certainly, the two experiments do differ in a number of important ways. Firstly, the input languages are quite unalike. The two objects in Kanwal et al.'s experiment shared a short name (“zop”) which was derived by clipping their unique long names (“zopekil” and “zopudon”). In this way, there was a clear relationship between an object's alternative names, and the ambiguity of the short name was a property of the lexicon that was evident throughout the experiment, including during training. Conversely, the two names for each object in our experiment were clearly unrelated, and while the HND words were very similar to each other, there was no outright ambiguity in the lexicon: the ambiguity only arose during communication as a side-effect of noisy transmission. It may therefore be the case that participants in Kanwal et al. were starting to form ideas about how they would deal with the ambiguity earlier in the experiment, whereas participants in our experiment had insufficient time to explore different strategies once they realised that the HND words were functionally ambiguous. In fact, it is possible that participants in our experiment didn't even realise that the HND words *were* ambiguous for their partner; anecdotally, a handful of participants reported on the debrief questionnaire that their partner was only sending one-letter responses, suggesting that not all participants understood that the noisy transmission was symmetrical and their partner had the same kind of comprehension difficulty as themselves. This is an inherently different situation from the one in Kanwal et al., where participants knew exactly how much information the different labels provided for their partner; future work should look to ensure that participants are aware of what has actually been transmitted to their partner on Director trials. Furthermore, it is likely that participants have more explicit awareness and experience of abbreviating frequent words (e.g. “information” → “info”) than they do of preferentially selecting between synonyms to maximise ease of pronunciation, and may be bringing this experience to bear when considering how to solve the task.

Secondly, the manipulation of production effort in Kanwal et al. was perhaps more transparent than our keyboard task: the time for which participants had to click and hold to send a longer word in the former was effectively dead time, whereas participants in our experiment were still engaged in the task whilst forming LND words, even if it did take longer. Although our manipulation clearly works in the sense that participants in the Production condition strongly favoured the easier-to-form HND words, it could still be the case that it is too subtle when a competing pressure is present. This may also be exacerbated by the fact that the pressure for accuracy probably feels inherently stronger for participants than the pressure for speed: Prolific participants are highly motivated to complete tasks “correctly” to avoid having their submissions rejected. We tried another version of the experiment which attempted to address these first two points (reported in the Appendix), but the effect of frequency was not obviously stronger in this follow-up; the most notable change in participants' behaviour was simply an increased preference in favour of the HND words *overall*.

Finally, long words in Kanwal et al. remained consistently arduous throughout the experiment, since they always took a fixed number of seconds to transmit. On the other hand, participants in our experiment may have been able to improve at the keyboard task, thereby reducing the cost to produce LND words over time (relative to the cost for their partner by *not* producing them). However, we think this is unlikely to account for much of the variance between the two experiments since the letters required to form LND words changed position on every trial, so the only thing participants could really learn that would help them produce these words on subsequent trials is that they could ignore the centre of the keyboard (which should have become obvious almost immediately).

Leaving aside the differences between our experiment and that of Kanwal et al., it is also possible that we do not reproduce their results because we are missing some key mechanism that drives phonetic clustering in the real world; in other words, it may be that processes other than the trade-off between production and comprehension pressures are at play. An obvious candidate is iconicity: in natural language lexicons, words that sound similar also tend to have similar meanings (Dautriche, Mahowald, Gibson, & Piantadosi, 2017). It is worth noting that iconicity could reasonably have been expected to play a role in Kanwal et al.'s experiment also, since there is some evidence that word length tends to correlate with conceptual complexity (Lewis & Frank, 2016). Nonetheless, future work could certainly look to control for iconicity between forms and meanings in the design of experimental stimuli; such a study could help to adjudicate between competing hypotheses around the origin of phonetic clustering, or indeed provide evidence that this phenomenon arises from multiple sources.

In summary, although we do not find evidence for a frequency trade-off, our experiment does provide further evidence that neither production pressures nor comprehension pressures *alone* give rise to the kind of organisational structure we see in real lexicons, in line with Kanwal et al.'s results regarding Zipf's Law of Abbreviation and with the results of our computational model when it comes to word similarity. Furthermore, to the extent that there are subtle tendencies towards a natural-language-like frequency trade-off when both pressures are present, we would expect these to be amplified through transmission to successive generations of participants (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Thompson et al., 2016).

## 4. General discussion

In this paper, we investigated how pressures operating during individual episodes of communication might give rise to an emergent structural property of language, whereby lexicons tend to be more phonetically clustered than required by their phonotactics, especially for high-frequency items.

In an exemplar-based computational model, we showed that clustering emerges under competition between production-side pressures for word similarity and comprehension-side pressures for discriminability. The lexicons that arise from this competition are neither as clustered nor as disperse as they possibly could be, although there is some variance in the exact details of how the two pressures are balanced depending on the strength of the comprehender-side pressure for distinctiveness and, to a lesser extent, frequency. With only one communicative pressure at work, the resulting lexicons very clearly fall at one extreme or the other. Specifically, when producibility is the only pressure, the outcome of repeated communication is a lexicon that is extremely easy to produce but communicatively degenerate, in that all words sound almost exactly the same. On the other hand, when comprehensibility is the only pressure, lexicons are maximally expressive in that all words are very distinct, but arduous from a production perspective due to the lack of shared sound sequences across words.

In a communication experiment using an artificial language, we showed that, when ease of production is the only pressure shaping participant behaviour, a strong preference emerges in favour of words from a high-density neighbourhood, while when ease of comprehension is the only pressure, the opposite preference (in favour of words from low-density neighbourhood) emerges. Extrapolating these preferences to an imagined wider lexicon, it is clear that our experiment makes the same predictions as our model: production pressures alone would be expected to give rise to a highly clustered lexicon, while comprehension pressures alone would lead to a highly disperse lexicon. As in the model, an intermediate state emerges when these pressures are in competition. Specifically, one neighbourhood does not completely win out over the other in this scenario; rather, words from both neighbourhoods have their place. However, it is not clear that selection between words from the different neighbourhoods is modulated by frequency.

Naturally, our model and experiment differ in some respects. Perhaps most notably, the model relies on a much larger meaning space (20) than the experiment (2). This discrepancy stems chiefly from practical considerations: computational models allow us to explore phenomena at a scale that would be infeasible with human participants in a short experiment. Similarly, while computational models are still simplifications of the real-world systems they are used to study, they do afford greater complexity than human experiments in some aspects. An example in our study is the way we manipulate production and comprehension difficulty. At a high-level, this manipulation is equivalent between the model and the experiment: words that are more similar to other words can be made easier to produce, harder to interpret, both, or neither. However, it is true that this is essentially a binary distinction in the experiment, while the model treats word similarity (and accordingly, its effects on production and comprehension) as a probabilistic gradient. This divergence is, we would contend, in line with the affordances of the two methodologies: while models benefit from more nuanced representations, the level of control we exercise in experiments arguably yields more precise and easily interpretable results.

It is also worth noting that we do not see the discrepancies between the model and the experiment as problematic, because our intention in this study was not to design a model *of* the experiment (unlike e.g. Brochhagen, 2021, who uses cross-validation and model comparison to find a set of parameters that best fits the experimental data from Kanwal et al., 2017). Instead, we see our model as a "pragmatic cognitive tool" (Pesonen, 2024), and its relationship to the experiment as one of conceptual rather than formal similarity. The reason to adopt such a combination of methodologies is simply to provide converging evidence — in general terms — of the conditions under which artificially evolved languages come to resemble natural languages (as in e.g. Kirby et al., 2015). Of course, both the model and the experiment are simplifications of the real world: they provide a qualitative understanding of the phenomenon of interest, but neither is designed to make detailed quantitative predictions about the evolution of any individual natural language (see Boyd & Richerson, 1985 for similar commentary on this scientific approach). Rather, the key contribution of both elements is one advanced by Woodward (2003)'s theory of causal explanation: "to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways" (p. 11).

In any case, while the model and the experiment may not be directly (i.e. quantitatively) comparable, it is certainly possible to draw some general conclusions across the two pieces. Overall, our results demonstrate that mechanisms operating during individual episodes of communication can shape the structure of the lexicon. Crucially, we show that evolving lexicons balance the influence of competing pressures that pull in different directions. However, with respect to the role of frequency, our results are less clear: frequency effects were subtle in our model, and do not emerge robustly in our experiment. Clearly, it is not possible to make precise predictions from natural language data about what effect sizes we would expect in such highly simplified, simulated lexicons. However, it is worth noting that the relationship between frequency and clustering in real languages is not necessarily a strong one; in fact, it is specifically described as a "weak tendency" by Frauenfelder et al. (1993). Correlations between frequency and different measures of clustering in Mahowald et al. (2018) were generally small, with Pearson's $r$ values deemed as statistically significant starting at 0.08 and rarely exceeding 0.3. The relationship between frequency and clustering may also be stronger for word beginnings than endings (King & Wedel, 2020), or for content words over function words (Frauenfelder et al., 1993), factors not considered here. Therefore, we would suggest that the subtlety of the frequency effect across our model and experiment may be exactly as expected.

One criticism that might be levelled at our study is that the extreme outcomes that emerge under the influence of a single communicative pressure paint a highly unrealistic picture of the cognitive biases that shape language. As pointed out by Wasow et al. (2005), if our notion of "production effort" includes the effort required to clarify what was intended for a confused receiver, then effort would clearly not be minimised by a degenerate language (with only one word for every meaning). However, in the limit, a bias to re-use sound sequences across words points to exactly such a language, and we would argue that, all else being equal, producers would want their language to conform to this bias. It is exactly because producers have communicative goals that all else is *not* equal, and a compromise position has to emerge. Similarly, it is clearly true that, as comprehenders, we can happily cope with some amount of noise in the linguistic signal, because there are plenty of other ways to extract an interlocutor's intended meaning — from contextual cues in the environment to the many multimodal features of language like co-speech gesture and facial expression. Even so, if all language users cared about was maximising comprehensibility, there would certainly be no harm in having lexicons be as disperse as their phonotactics would allow. It is precisely because comprehensibility is *not* the only thing language users need to worry about that we do not see such lexicons in the real world. Whilst acknowledging that these counterfactual either-or situations do not represent real language use, it is still useful to examine their consequences in isolation; by doing so, we can verify that the phenomena we are trying to explain do in fact result from a trade-off between competing pressures, and cannot be more simply explained by one pressure or the other.

Natural language lexicons, as in the critical conditions of our model and experiment, are under pressure to adapt to several competing forces. The way in which they achieve an optimal balance between these pressures is clearly not simple, and depends on several factors. For example, biases can vary in strength: in our model, one source of variation was captured by the Receiver's $\gamma$ parameter (Section 2.1.3), but there are no doubt others in the real world, such as differences in articulatory or auditory apparatus that might make certain sound sequences more or less difficult to pronounce for certain individuals (e.g. Franken et al., 2017). In our experiment, a variety of individual differences may have pushed different participants to arrive at different solutions to the task; for example, more risk averse participants may have been less willing to sacrifice accuracy for the sake of speed (Carver & White, 1994). Nonetheless, the lexicons that emerge under competing pressures are, in some sense, *efficient* (Gibson et al., 2019; Jaeger & Tily, 2011): words are just distinctive "enough" whilst still being as easy to produce "as possible" (where "enough" and "as possible" are defined with reference to a specific communicative or cognitive context). Optimising for producibility inevitably means introducing some ambiguity, but as pointed out by Piantadosi et al. (2012), ambiguity is actually a hallmark of an efficient communication system since it allows for the reuse of words and sounds that are more easily produced, and doesn't impede communication as long as there are other ways for the comprehender to overcome the ambiguity. In our experiment, for example, participants could overcome the ambiguity of the HND words during Matcher trials either by adopting a very simple heuristic of probability matching their guesses to the relative frequencies of meanings in the world (since words are, *a priori*, more likely to refer to things we talk about more), or by establishing a shared code with their partner that would allow them to use probabilistic information from previous interactions to inform future ones.

While our study provides further evidence for the role of competing communicative pressures in driving language efficiency, our simulation of the pressures acting on language is undoubtedly a simplification in a number of ways. Mostly notably, our experiment *simulates* the pressures involved in language use, rather than relying on them to emerge at scale in the lab. Most obviously, typing is not language production in the usual sense, and naturalistic comprehension is not the same as image selection. Replicating this study in a more ecologically valid setting (i.e. with oral production and auditory comprehension tasks) is a logical next step for a few reasons. First, allowing pressures to emerge naturally could, in principle, provide more compelling evidence for a causal link between individual-level behaviour and population-level language trends like phonetic clustering. Second, there may be specific aspects of production effort that are not well-simulated by anything other than oral production. However, it seems likely that the difficulty associated with these tasks would still need to be artificially inflated — for example, through the use of highly phonotactically complex words, or environmental noise on transmission — to observe, in a brief experiment, the kinds of effects that otherwise accumulate only over much larger timescales. The benefit of our design is that it allows us to easily manipulate task difficulty in a way that affects all participants roughly equally and does not depend on, for example, prior experience with pronouncing certain sounds, or auditory acuity. By doing so, we can get an idea of how small and potentially noisy effects at an individual-level might accumulate into large effects at a population-level (Kirby et al., 2007).

The present work also does not account for every possible mechanism that could play a role in shaping this aspect of lexicon structure. For example, it is possible that clustering emerges more strongly from new words entering the lexicon than from changes to or selection between existing words. Such a mechanism could also go some way to explaining the frequency effects we see in natural languages: if high-frequency words are a stronger attractor for the form of new words than low-frequency words, new coinages would tend to increase connectivity more in high-frequency components of the lexicon (see Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017 for a similar suggestion). Future work should investigate how different kinds of lexical evolution — from coinage to sound change and, ultimately, obsolescence — might differentially drive changes in the network properties of the lexicon.

Our model and experiment also do not account for the role of learning biases in shaping linguistic systems (Christiansen & Chater, 2008; Culbertson, 2012; Griffiths et al., 2008; Kalish et al., 2007; Kirby et al., 2008, 2014; Smith et al., 2003). There are several reasons to think that learning might play a role in driving increased clustering. For one, lexicons built from a smaller inventory of sound sequences are more compressible (Ferrer-i-Cancho et al., 2013), a property which reduces storage demands (Storkel & Maekawa, 2005) and allows languages to pass more easily through the bottleneck imposed by repeated transmission to naive individuals (Kirby et al., 2015). Moreover, infants and children show clear preferences for words composed of the highest-frequency sound sequences in their target language (Altvater-Mackensen & Mani, 2013; Jusczyk et al., 1994; Ngon et al., 2013) and generally acquire such words earlier (Coady & Aslin, 2004; Gonzalez-Gomez et al., 2013; Storkel, 2004). Since early-acquired words are also known to be more stably represented within a community's language (Monaghan, 2014), we might expect these developmental effects to show up in evolution. However, a learning-based account does not straightforwardly point to a clustering advantage (see e.g. Dautriche et al., 2015; Jones & Brandt, 2020; Storkel et al., 2006; Storkel & Lee, 2011; Swingley & Aslin, 2007).

Furthermore, lexicons are not, contrary to the dominant view of "design features" (Hockett, 1960), entirely arbitrary. Rather, languages are rife with sound symbolism and other systematic associations between form and meaning (Bergen, 2004; Blasi et al., 2016; Cuskley & Kirby, 2013; Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Dingemanse et al., 2015; Monaghan et al., 2007, 2014; Perlman & Woodin, 2021; Perniss et al., 2010; Tamariz, 2008; Winter & Perlman, 2021; Winter et al., 2022). A detailed account of the role of semantics is missing from our study, since there is no level of analysis below the atomic meaning (e.g. we do not consider the meaning "lightbulb" to have any features that might be shared across other meanings, such as being man-made or having to do with electricity). Although correlations between semantic similarity and wordform similarity are not necessarily strong, they are significantly higher than would be expected by chance, and can be observed consistently across a diverse sample of languages (Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Monaghan et al., 2014). It is therefore likely that iconicity is an important driver of phonetic clustering in natural language lexicons; future work should look to tease apart the relative strength of the communicative pressures we study here and other selective pressures acting on languages, such as those favouring iconic mappings between form and meaning. Another source of non-arbitrariness is shared etymology: words that come from the same historical root may consequently sound similar in their modern form (Klein, 1971). We do not take into account any such structure in our models since we use randomly-generated lexicons as the input to the agents. However, we would argue that if the phonetic clustering that resulted from shared etymology was detrimental for communication, it could be selected out through cultural evolution; the fact that natural language lexicons are observably more clustered than they could be suggests that this is not the case. Nonetheless, future work could look to incorporate notions of semantic and historic relatedness as a more conservative test of our hypotheses. Our model could also be adapted to test a variety of different starting conditions.

Finally, it is possible that our study overstates the role of functional pressures altogether (Caplan et al., 2020; Ceolin, 2020; Kauhanen, 2017; Labov, 1987, 1994), or that our conceptualisation of the functional pressures at work is too general. For example, production pressures which favour high phonotactic probability sequences may work in concert with a specific comprehension-side pressure to minimise outright homophony, thus creating clusters of similar but non-identical words (Trott & Bergen, 2022). This would lead to the appearance of dense neighbourhoods with no direct selection for neighbourhood properties *per se*. Our results do not speak to the role of an anti-homophony pressure, since we treat word similarity as a gradient rather than a binary classification: this means that homophony is seen simply as a high degree of similarity, not a distinct phenomenon. Future work could look to adjudicate between the more general principle we explore here — that comprehension broadly favours discriminability — and a more specialised pressure which targets homophony in particular.

## 5. Conclusion

Corpus data shows that natural language lexicons are more phonetically clustered than would be expected, even accounting for phonotactic rules, morphology and sound symbolism. This study provides the first evidence that this organisational property of the lexicon can arise as a result of mechanisms operating at the level of individual language users and individual communication episodes. Specifically, we show that emergent lexicon structure balances the influence of competing functional pressures: a pressure for distinctiveness arising from comprehension, and a pressure for reuse of forms arising from production. When only one of these pressures is present, the lexicons that emerge exhibit extreme levels of clustering or dispersion unlike those seen in natural languages. This study adds to a growing body of evidence showing that, through a process of cultural evolution, languages are optimised for efficient communication.

## CRediT authorship contribution statement

**Aislinn Keogh:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jennifer Culbertson:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Simon Kirby:** Writing – review & editing, Supervision, Methodology, Conceptualization.

## Acknowledgements

This project was supported by funding from the Economic and Social Research Council, United Kingdom (grant ref. ES/P000681/1, awarded to AK) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643, awarded to JC). We are grateful to Juan Guerrero Montero, Andy Wedel and Matt Spike for invaluable discussion on the model, to Kenny Smith for support with the interaction server, and to Annie Holtz for help designing the artificial language used in the experiment.

## Appendix. Follow-up experiment

As discussed in Section 3.3, there were a number of differences between the design of our experiment and the one it was modelled after (Kanwal et al., 2017). In particular, we felt that our manipulation of production effort may have been too subtle to push participants towards an efficient solution in the presence of a competing pressure for accuracy. We also wondered whether the unclear relationship between an object's two alternative names may have changed participants' representation of the language in a way that could influence their behaviour during communication. We therefore ran a follow-up experiment which attempted to address these two concerns, while maintaining the general design whereby words from the high-density neighbourhood were easier to produce but functionally ambiguous, while words from the low-density neighbourhood were harder to produce but easily distinguishable. The changes are summarised in Fig. A.1 and described below.

### A.1. Materials

The meaning space consisted of the same two objects in the same frequency distribution as in the first experiment. The language consisted of four artificial CVC words: "foz" [fɑz] and "faz" [fæz] (the HND words) and "poz" [pɑz] and "vaz" [væz] (the LND words). Each LND word in this lexicon has a corresponding HND word (with which it shares the final two phonemes) which is derived by a known process of sound change: /p/ → /f/ (e.g. Foulkes, 1997) and devoicing as in /v/ → /f/ (e.g. Velde et al., 1996).

### A.2. Procedure

The procedure was identical as in the first experiment, except for the design of the difficult Director trials. On these trials, as before, the Director was presented with both word options for the target object and asked to use an on-screen keyboard to type one of the words. However, the keyboard in this experiment contained only letters that were part of the artificial language, and all buttons were the same size and appeared in the same position from trial-to-trial (the configuration was randomised for each participant). Instead, the two keys required to make an LND word ("p" and "v") were wonky (a random angle of ±10, ±15 or ±20 degrees was chosen for each button on each trial), and had a cracked texture around the edge. At the start of each trial, a random integer between 1 and 3 was generated, representing the total number of times either of these keys would need to be pressed before the correct letter would appear; other times, a random letter that wasn't part of the artificial language would appear. Every time one of these keys produced an incorrect letter, participants would need to press an "undo" button to get rid of that letter before trying again. Participants were told that some of the buttons were faulty and might need to be pressed a few times. As before, this design was intended to simulate the observation that less frequently-used phonemes are more error prone; however, we hoped that this manipulation would make the LND words more costly from participants' perspective than in the first experiment.

### A.3. Participants and exclusions

Due to financial constraints, we were only able to run the critical Combined condition in this follow-up experiment. We used Prolific to recruit 72 participants who had not taken part in the first experiment. The experiment took around 25 min to complete in full (median time = 22:44) for which participants were paid £4.25. One participant was prevented from proceeding to the communication game due to low accuracy on the pre-test and paid a reduced rate of £2. 13 participants started but failed to complete the interaction phase and were paid a variable rate depending on how far they had got through the experiment. Two participants (one pair) completed the communication game and were paid the full rate, but their data was excluded from analysis because their completion time was more than 3 standard deviations above the median. After all exclusions and dropouts, we were left with 28 pairs: a total of 56 individual participants.

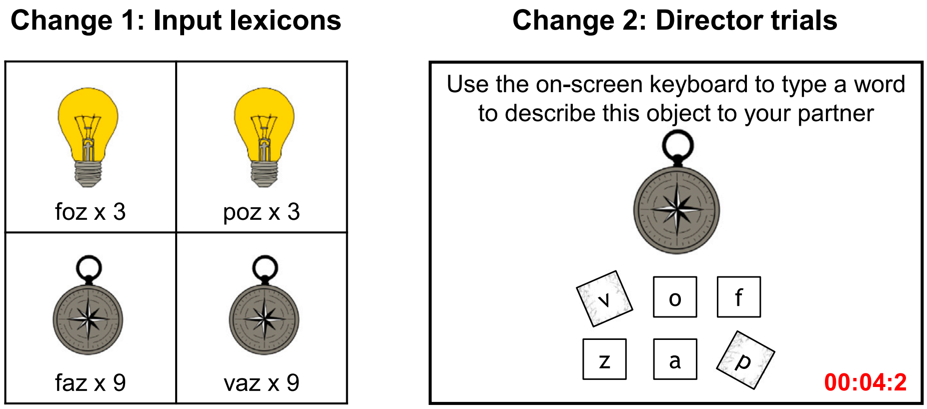
Change 1: Input lexicons
foz x 3
poz x 3
faz x 9
vaz x 9
Change 2: Director trials
Use the on-screen keyboard to type a word to describe this object to your partner
v
o
f
z
a
p
00:04:2

**Fig. A.1.** Summary of design changes in the follow-up experiment. Input lexicons were designed such that the HND words were clearly variants of the LND words, rather than completely different words (left). Director trials used an on-screen keyboard in which the keys required to form an LND word were faulty — indicated by their cracked texture and wonky placement — and sometimes produced an incorrect letter (right).

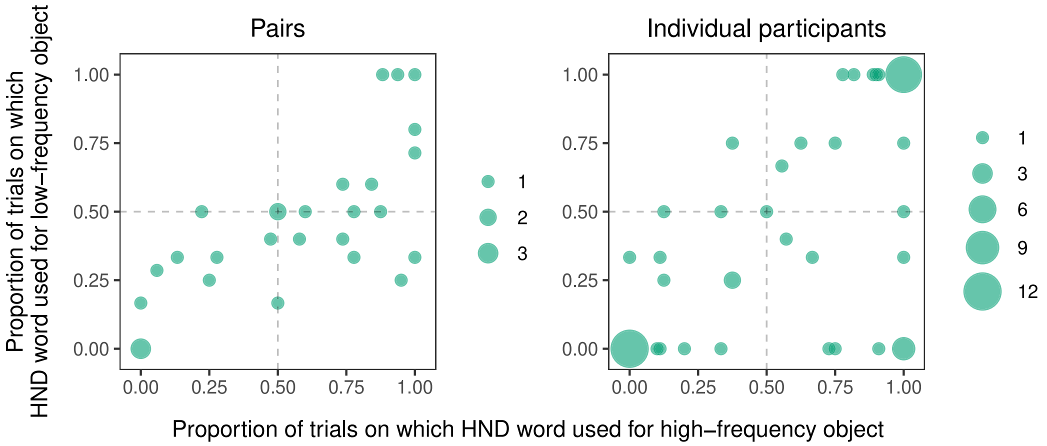

**Fig. A.2.** Proportion of trials on which the HND word was used for the high-frequency object vs. the proportion of trials on which it was used for the low-frequency object, by-pair (left) and by-participant (right). As in the first experiment, individual participants are more strongly clustered in the corners than pairs, suggesting that not all pairs are converging on the same language. As in the first experiment, a range of behaviours are represented, and it is not clear that a natural-language-like frequency trade-off (bottom right quadrant) is the most common strategy.

### A.4. Results

Fig. A.2 shows the proportion of Director trials on which the HND word was used for the high and low-frequency objects. As in the first experiment, a range of strategies are represented, and it is not clear that most participants are converging on the predicted frequency trade-off. We fit a reduced version of the model described in Section 3.2.1; since we only ran one condition in this follow-up experiment, there is no longer a fixed effect of condition, nor an interaction between condition and frequency. The model had by-participant random intercepts and random slopes for object frequency, but failed to converge with the nested by-pair random effects structure used in Section 3.2.1. Model predictions are shown in Fig. A.3. The model reveals a significant main effect of frequency, such that participants were more likely to use the HND word to label the high-frequency object ($\beta = 0.877$, $SE = 0.392$, $t = 2.237$, $p < 0.05$). This result follows straightforwardly from the fact that there are many more participants below than above the diagonal in Fig. A.2 i.e. for participants who showed *any* effect of frequency, it was generally the predicted one. In other words, very few participants adopted an anti-efficient strategy of using the difficult-to-produce LND word for the high-frequency object and the easy-to-produce HND word for the low-frequency object.

However, if we consider the two experiments as a whole, it seems that the key difference between them is not in the strength of the frequency effect. We pooled the data from the Combined condition of the

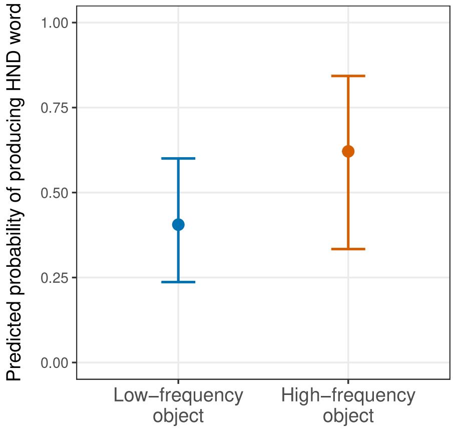

**Fig. A.3.** Model predictions generated using the `ggeffects` package (Lüdecke, 2018). The model predicts that participants were more likely to produce an HND word for the high-frequency object than for the low-frequency object.

**Table A.1**
Summary of fixed effects for a logistic mixed effects model with HND word use as the binary dependent variable and by-participant random effects for object frequency. The main experiment reported in Section 3 is labelled as 1a; the follow-up experiment is labelled as 1b. Coefficient estimates are on the log-odds scale.

| | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| intercept (object = infrequent, experiment = 1a) | −3.039 | 0.707 | −4.300 | <0.001 |
| object = frequent | 1.537 | 0.799 | 1.923 | 0.054 |
| experiment = 1b | 2.546 | 0.851 | 2.993 | <0.01 |
| object = frequent & experiment = 1b | −0.452 | 0.944 | −0.479 | 0.632 |

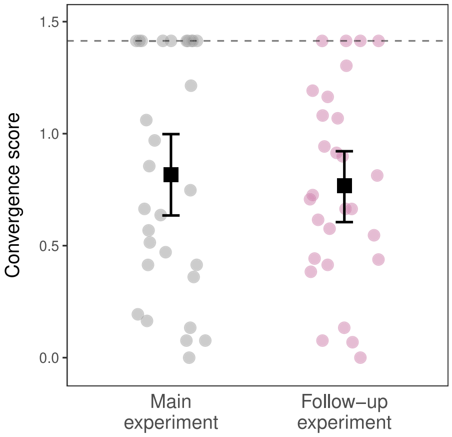

**Fig. A.4.** Convergence scores for the Combined condition of the main experiment (left) and the follow-up experiment (right). Convergence is very similar between the two experiments.

first experiment with the data from this follow-up experiment, and fit a mixed effects logistic regression model predicting HND word use as a function of object frequency, experiment, and their interaction. Again, the model had by-participant random intercepts and random slopes for object frequency, but failed to converge with a nested by-pair random effects structure. A full summary of model coefficients is given in Table A.1. The model reveals no overall effect of frequency, despite the significant effect of frequency when considering the follow-up experiment in isolation. However, there is also no interaction between frequency and experiment; that is, there is no evidence that either experiment showed a clearer effect of frequency. Crucially, the model does show a significant main effect of experiment, such that the *overall* probability of producing an HND word was higher in the follow-up experiment. In other words, our changes to the experimental design succeeded in making the LND words more costly for participants to produce, but not in such a way that made the predicted frequency trade-off emerge more robustly. Convergence between the two members of a pair (i.e. the extent to which they settled on a shared language) also did not improve in the follow-up experiment (Fig. A.4).

Overall, the results of this follow-up experiment provide further evidence that, insofar as there is a relationship between frequency and clustering, it may be more subtle than the relationship between frequency and word length probed by Kanwal et al. (2017)'s experiment.

## Data availability

All data and code is available at https://osf.io/vsy6z/.

## References

Alexandrov, A. A., Boricheva, D. O., Pulvermüller, F., & Shtyrov, Y. (2011). Strength of word-specific neural memory traces assessed electrophysiologically. *PLoS One*, *6*(8), Article e22999.

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6), 980–990.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (version 2).

Barrett, J. A. (2006). Numerical simulations of the lewis signaling game: Learning strategies, pooling equilibria, and the evolution of grammar. *Institute for Mathematical Behavioral Sciences, UC Irvine*.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using Lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311.

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

BNC Consortium (2007). British national corpus, XML edition.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Brochhagen, T. (2021). Brief at the risk of being misunderstood: Consolidating population- and individual-level tendencies. *Computational Brain & Behavior*, *4*(3), 305–317.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425–455.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261–290.

Ferrer-i Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565–1578.

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.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333.

Ceolin, A. (2020). Functionalism, lexical contrast and sound change.

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1934–1949.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: retroactive effects on multiple activation. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(3), 551–563.

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.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*(5), 310–329.

Cuskley, C., & Kirby, S. (2013). Synesthesia, cross-modality, and language evolution. In *Oxford library of psychology*, *the oxford handbook of synesthesia* (pp. 869–899). Oxford University Press.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of english. *PLoS One*, *9*(8), Article e102882.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

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

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

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology (2006)*, *71*(4), 808–816.

Flego, S. (2022). *The emergence of vowel quality mutation in germanic and dinka-nuer: modeling the role of information-theoretic factors using agent-based simulation* (Doctoral dissertation), Indiana University.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In B. Hayes, D. Steriade, & R. Kirchner (Eds.), *Phonetically based phonology* (pp. 232–276). Cambridge University Press.

Foulkes, P. (1997). Historical laboratory phonology—Investigating /p/>/f/>/h/ changes. *Language and Speech*, *40*(3), 249–276.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Franke, M., & Jäger, G. (2012). Bidirectional optimization from reasoning and learning in games. *Journal of Logic, Language and Information*, *21*(1), 117–139.

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *Journal of the Acoustical Society of America*, *142*(4), 2007.

Frauenfelder, U. H., Baayen, R. H., & Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, *32*(6), 781–804.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

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.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, *28*(5), 501–518.

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition*, *107*(3), 1155–1164.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, *102*(2), 219–260.

Gonzalez-Gomez, N., Poltrock, S., & Nazzi, T. (2013). A "bat" is easier to learn than a "tab": Effects of relative phonotactic frequency on infant word learning. *PLoS One*, *8*(3), Article e59601.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *363*(1509), 3503–3514.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*(3), 297–306.

Hockett, C. F. (1960). The origin of speech. *Scientific American*.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

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

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402–420.

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.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kauhanen, H. (2017). Neutral change. *Journal of Linguistics*, *53*(2), 327–358.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364.

King, A., & Wedel, A. B. (2020). Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind*, *4*, 1–12.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kirby, S., Griffiths, T. L., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Klein, E. (1971). *A comprehensive etymological dictionary of the english language: dealing with the origin of words and their sense development thus illustrating the history of civilization and culture*. Elsevier Publishing Company.

Krevitt, B., & Griffith, B. C. (1972). A comparison of several zipf-type distributions in their goodness of fit to language data. *Journal of the American Society for Information Science*, *23*(3), 220–221.

Labov, W. (1987). The overestimation of functionalism. In R. Dirven, & V. Fried (Eds.), *Linguistic and literary studies in eastern europe*, *Functionalism in linguistics* (p. 311). John Benjamins Publishing Company.

Labov, W. (1994). *Principles of linguistic change: internal factors*: *Vol. 1*, Blackwell Publishing.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*(2), 119–131.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken english: based on the British National Corpus* (1st ed.). Routledge.

Levitt, A. G., & Healy, A. F. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language*, *24*(6), 717–733.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing* (p. 234). Association for Computational Linguistics.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Lorch, M. P., & Meara, P. (1989). How people listen to languages they don't know. *Language Sciences*, *11*(4), 343–353.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.

Macklin-Cordes, J. L., & Round, E. R. (2020). Re-evaluating phoneme frequencies. *Frontiers in Psychology*, *11*.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156.

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

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Marks, E. A., Bond, Z., & Stockmal, V. (2003). Language experience and the representation of phonology in an unknown language. *Revista Española de Linguistica Aplicada*.

Marquet, P. A., Allen, A. P., Brown, J. H., Dunne, J. A., Enquist, B. J., Gillooly, J. F., Gowaty, P. A., Green, J. L., Harte, J., Hubbell, S. P., O'Dwyer, J., Okie, J. G., Ostling, A., Ritchie, M., Storch, D., & West, G. B. (2014). On theory in ecology. *BioScience*, *64*(8), 701–710.

Martindale, C., Gusein-Zade, S. M., McKenzie, D., & Borodovsky, M. Y. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes*. *Journal of Quantitative Linguistics*, *3*(2), 106–112.

Mayer, C., Kondur, A., & Sundara, M. (2022). UCI phonotactic calculator.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Meylan, S. C., & Griffiths, T. L. (2024). Word forms reflect trade-offs between speaker effort and robust listener recognition. *Cognitive Science*, *48*(7), Article e13478.

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. *Cognition*, *133*(3), 530–534.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259–305.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *369*(1651), Article 20130299.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Motley, M. T., & Baars, B. J. (1975). Encoding sensitivites to phonological markedness and transitional probability: Evidence from spoonerisms. *Human Communication Research*, *1*(4), 353–361.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research: JSLHR*, *44*(4), 778–792.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (2011). The generalized context model: an exemplar model of classification. In A. J. Wills, & E. M. Pothos (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge University Press.

Perlman, M., & Woodin, G. (2021). A complete real-world theory of language should explain how iconicity remains a stable property of linguistic systems. *Journal of Cognition*, *4*(1).

Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, *1*.

Pesonen, R. (2024). Highly idealized models of scientific inquiry as conceptual systems. *European Journal for Philosophy of Science*, *14*(3), 39.

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

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In *Typological studies in language*: *Vol. 45*, *Frequency and the emergence of linguistic structure* (pp. 137–157). John Benjamins Publishing Company.

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1–46.

R. Core Team (2024). R: a language and environment for statistical computing, vienna, austria.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(3), 394–410.

Sims-Williams, H. (2022). Token frequency as a determinant of morphological change. *Journal of Linguistics*, *58*(3), 571–607.

Skyrms, B. (2010). *Signals: evolution, learning, and information*. Oxford University Press.

Smith, K., Ashton, C., & Sims-Williams, H. (2023). The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Smith, K., Bowerman, J., & Smith, A. D. M. (2025). Semantic extension in a novel communication system is facilitated by salient shared associations. *Cognition*, *261*, 106129.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*, 371–386.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2013). Learning, feedback and information in self-organizing communication systems. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 3442–3447).

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*(3), 623–658.

Steels, L. (2012). Self-organization and selection in cultural language evolution. In L. Steels (Ed.), *Advances in interaction studies*, *Experiments in cultural language evolution* (pp. 1–37). John Benjamins Publishing Company.

Steels, L., & Loetzsch, M. (2012). The grounded naming game. In L. Steels (Ed.), *Advances in interaction studies*, *Experiments in cultural language evolution* (pp. 41–59). John Benjamins Publishing Company.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1), 413–422.

Stockmal, V., Muljani, D., & Bond, Z. (1996). Perceptual features of unknown foreign languages as revealed by multi-dimensional scaling. In *Proceeding of fourth international conference on spoken language processing*: *Vol. 3*, (pp. 1748–1751 vol.3).

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192.

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211.

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language*, *32*(4), 827–853.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132.

Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, *3*(2), 259–278.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*(16), 4530–4535.

Trott, S., & Bergen, B. (2022). Can a pressure against homophones explain phonological neighborhoods? In *Proceedings of the annual meeting of the cognitive science society*: *Vol. 44*.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine phonotactic online dictionary.

Velde, H. v. d., Gerritsen, M., & Hout, R. v. (1996). The devoicing of fricatives in standard dutch: A real-time study based on radio recordings. *Language Variation and Change*, *8*(2), 149–175.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(4), 735–747.

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), 514–529.

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. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, *52*(2), 193–204.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*(1), 47–62.

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), 306–311.

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 and Cognition*, *31*(4), 491–504.

Wasow, T., Perfors, A., & Beaver, D. I. (2005). The puzzle of ambiguity. In *Morphology and the web of grammar: essays in memory of Steven G. Lapointe*. CSLI Publications.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, *23*(3), 247–274.

Wedel, A. B. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, *4*(4), 319–355.

Wedel, A. B., & Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, *2*(1), 77–93.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967.

Winter, B., & Perlman, M. (2021). Size sound symbolism in the english lexicon. *Glossa: A Journal of General Linguistics*, *6*(1).

Winter, B., Sóskuthy, M., Perlman, M., & Dingemanse, M. (2022). Trilled /r/ is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, *12*(1), 1035.

Woodward, J. (2003). *Making things happen: a theory of causal explanation*. Oxford University Press.

Wu, S., Cotterell, R., & O'Donnell, T. (2019). Morphological irregularity correlates with frequency. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5117–5126). Association for Computational Linguistics.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.