Words cluster phonetically beyond phonotactic regularities

Isabelle Dautriche*[1,2], Kyle Mahowald*[3], Edward Gibson[3], Anne Christophe[1] and Steven T. Piantadosi[4]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University, Paris, France
[2]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom
[3]Department of Brain and Cognitive Science, MIT
[4]Department of Brain and Cognitive Sciences, University of Rochester

*These authors contributed equally to this work. For correspondence, e-mail isabelle.dautriche@gmail.com or kylemaho@mit.edu

Abstract

Recent evidence suggests that cognitive pressures associated with language acquisition and use could affect the organization of the lexicon. On one hand, consistent with noisy channel models of language (e.g., Levy 2008), the phonological distance between wordforms should be maximized to avoid perceptual confusability (a pressure for *dispersion*). On the other hand, a lexicon with high phonological regularity would be simpler to learn, remember and produce (e.g., Monaghan et al., 2011) (a pressure for *clumpiness*). Here we investigate wordform similarity in the lexicon, using measures of word distance (e.g., phonological neighborhood density) to ask whether there is evidence for dispersion or clumpiness of wordforms in the lexicon. We develop a novel method to compare lexicons to phonotactically-controlled baselines that provide a null hypothesis for how clumpy or sparse wordforms would be as the result of only phonotactics. Results for four languages, Dutch, English, German and French, show that the space of monomorphemic wordforms is clumpier than what would be expected by the best chance model according to a wide variety of measures: minimal pairs, average Levenshtein distance and several network properties. This suggests a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonologically distinct as possible.

Keywords: linguistics, lexical design, communication, phonotactics,

1 Introduction

de Saussure (1916) famously posited that the links between wordforms and their meanings are arbitrary. As Hockett (1960) stated: "The word 'salt' is not salty, 'dog' is not canine, 'whale' is a small word for a large object; 'microorganism' is the reverse." Despite evidence for non-arbitrary structure in the lexicon in terms of semantic and syntactic categories (Bloomfield, 1933; Monaghan et al., 2014), the fact remains that here is no systematic reason why we call a dog a 'dog' and a cat a 'cat' instead of the other way around, or instead of 'chien' and 'chat.' In fact, our ability to manipulate such arbitrary symbolic representations is one of the hallmarks of human language and makes language richly communicative, since it permits reference to arbitrary entities, not just those that have iconic representations (Hockett, 1960).

Because of this arbitrariness, languages have many degrees of freedom in what wordforms they choose and in how they carve up semantic space to assign these forms to meanings. Although the mapping between forms and meanings is arbitrary, the particular sets of form-meaning mappings chosen by any given language may be constrained by a number of competing pressures and biases associated with learnability and communicative efficiency. For example, imagine a language that uses the word 'feb' to refer to the concept HOT, and that the language now needs a word for the concept warm. If the language used the word 'fep' for WARM, it would be easy to confuse with 'feb' (HOT) since the two words differ only in the voicing of the final consonant and would often occur in similar contexts (i.e. when talking about temperature). However, the similarity of 'feb' and 'fep' could make it easier for a language learner to learn that those sound sequences are both associated with temperature, and the learner would not have to spend much time learning to articulate new sound sequences since 'feb' and 'fep' share most of their phonological structure. On the other hand, if the language used the word 'sooz' for the concept WARM, it is unlikely to be phonetically confused with 'feb' (HOT), but the learner might have to learn to articulate a new set of sounds and would need to remember two quite different sound sequences that refer to similar concepts.

Here, we investigate how communicative efficiency and learnability trade off in the large-scale structure of natural languages. We have developed a set of statistical tools to characterize the large-scale statistical properties of the lexicons. Our analysis focuses on testing and distinguishing two pressures in natural lexicons: a *pressure for dispersion* (improved discriminability) versus a *pressure for clumpiness* (re-use of sound sequences). Below, we discuss each in more detail.

A pressure for dispersion of wordforms

Under the noisy channel model of communication (Gibson et al., 2013; Levy, 2008; Shannon, 1948), there is always some chance that the linguistic signal will be misperceived as a result of errors in production, errors in comprehension, inherent ambiguity, and other sources of uncertainty for the perceiver. A lexicon is maximally robust to noise when the expected phonetic distance among words is maximized (Flemming, 2004; Graff, 2012), an idea used in coding theory (Shannon, 1948). Such dispersion has been observed in phonological inventories (Flemming, 2002; Hockett & Voegelin, 1955; Liljencrants & Lindblom, 1972) in a way that is sensitive to phonetic context (Steriade, 1997, 2001). The length and clarity of speakers' pronunciations are also sensitive to context predictability and frequency (e.g., Aylett & Turk, 2004; Bell et al., 2003; Cohen Priva, 2008; Pluymaekers et al., 2005; Raymond et al., 2006; Van Son & Van Santen, 2005), such that potentially confusable words have been claimed to be pronounced more slowly and more carefully. Applying this idea to the set of wordforms in a lexicon, one would expect wordforms to be maximally dissimilar from each other, within the bounds of conciseness and the constraints on what can be easily and efficiently produced by

the articulatory system. Indeed, a large number of phonological neighbors (i.e., words that are one edit apart like 'cat' and 'bat') can impede spoken word recognition (Luce, 1986; Luce & Pisoni, 1998), and the presence of lexical competitors can affect reading times (Magnuson et al., 2007). Phonological competition may also be a problem in early stages of word learning: young toddlers fail to use a single-feature phonological distinction to assign a novel meaning to a wordform that sounds similar to a very familiar one (e.g., learning a novel word such as "tog" when having "dog" in their lexicon, Dautriche, Swingley, & Christophe 2015; Swingley & Aslin 2007).

A pressure for clumpiness of wordforms

Dispersion of wordforms in the lexicon may be functionally advantageous. Yet, it is easy to see that a language with a hard constraint for dispersion of wordforms will have many long, therefore complex, words (as words need to be distinctive). A well designed lexicon must also be composed of simple signals that are easily memorized, produced, processed and transmitted over generations of learners. In the extreme case, one could imagine a language with only one wordform. Learning the entire lexicon would be as simple as learning to remember and pronounce one word. While this example is absurd, there are several cognitive advantages for processing words that are similar to other words in the mental lexicon. Words that overlap phonologically with familiar words are considered to be easier to process because they receive support from stored phonological representations. There is evidence that words that have many similar sounding words in the lexicon are easier to remember than words that are more phonologically distinct (Vitevitch et al., 2012) and facilitate production as evidenced by lower speech error rates (Stemberger, 2004; Vitevitch & Sommers, 2003). They also may have shorter naming latencies (Vitevitch & Sommers, 2003) (but see Sadat et al. 2014 for a review of the sometimes conflicting literature on the effect of neighborhood density on lexical production). Additionally, words with many phonological neighbors tend to be phonetically reduced (shortened in duration and produced with more centralized vowels) in conversational speech (Gahl, 2015; Gahl et al., 2012).This result is expected if faster lexical retrieval in production is associated with greater phonetic reduction in conversational speech as it is assumed for highly predictable words and highly frequent words (Aylett & Turk, 2006; Bell et al., 2003). In sum, while words that partially overlap with other words in the lexicon may be difficult to recognize (Luce, 1986; Luce & Pisoni, 1998), they seem to have an advantage for memory and lexical retrieval.

One source of wordform regularity in the lexicon comes from a correspondence between phonology and semantics and/or syntactic factors. Words of the same syntactic category tend to share phonological features, such that nouns sound like nouns, verbs like verbs, and so on (Kelly et al., 1992). Similarly, phonologically similar words tend to be more semantically similar within a language, across a wide variety of languages (Dautriche et al., submitted; Monaghan et al., 2014). The presence of these natural clusters in semantic and syntactic space therefore results in the presence of clusters in phonological space. Imagine, for instance, that all words having to do with sight or seeing had to rhyme with 'look'. A cluster of '-ook' words would develop, and they would all be neighbors and share semantic meaning. One byproduct of these semantic and syntactic clusters would be an apparent lack of sparsity among wordforms in the large-scale structure of the lexicon. There is evidence that children and adults have a bias towards learning words for which the relationship between their semantics and phonology is not arbitrary (Imai & Kita, 2014; Imai et al., 2008; Monaghan et al., 2011, 2014; Nielsen & Rendall, 2012; Nygaard et al., 2009). However such correspondences between phonology and semantic may affect some aspects of the production system: speech production errors that are semantically and phonologically close to the target (e.g., substituting 'cat' by 'rat') are much more likely to occur than errors than are purely semantic (e.g., substituting 'cat' by 'dog') or purely

phonological (e.g., substituting 'cat' by 'mat') in spontaneous speech (the *mixed error effect*, e.g., Dell & Reich, 1981; Goldrick & Rapp, 2002; Schwartz et al., 2006).

Another important source of phonological regularity in the lexicon is *phonotactics*, the complex set of constraints that govern the set of sounds and sound combinations allowed in a language (Hayes & Wilson, 2008; Vitevitch & Luce, 1998). For instance, the word 'blick' is not a word in English but plausibly could be, whereas the word 'bnick' is much less likely due to its implausible onset *bn-* (Chomsky & Halle, 1965).[1] These constraints interact with the human articulatory system: easy-to-pronounce strings like 'ma' and 'ba' are words in many human languages, whereas some strings, such as the last name of Superman's nemesis *Mister Mxyzptlk*, seem unpronounceable in any language.[2] Nevertheless, the phonotactic constraints of a language are often highly language-specific. While English does not allow words to begin with *mb*, Swahili and Fijian do. Phonotactic constraints provide an important source of regularity that aids production, lexical access, memory and learning. For instance, words that are phonotactically probable in a given language (i.e., that make use of frequent transitions between phonemes) are recognized more quickly than less probable sequences (Vitevitch, 1999). Furthermore, infants and young children seem to learn phonotactically probable words before learning less probable words (Coady & Aslin, 2004; Storkel, 2004, 2009; Storkel & Hoover, 2010) and infants prefer listening to high-probability sequences of sounds compared to lower probability sequences (Jusczyk & Luce, 1994; Ngon et al., 2013).[3]

The upshot of this regularity for the large-scale structure of the lexicon is to *constrain* the lexical space. For instance, imagine a language called *Clumpish* in which the only allowed syllables were those that consist of a nasal consonant (like *m* or *n*) followed by the vowel *a*. Almost surely, that language would have the words 'ma', 'na', 'mama', 'mana', and so on since there are just not that many possible words to choose from. The lexical space would be highly constrained because most possible sound sequences are forbidden. From a communicative perspective, such a lexicon would be disadvantageous since all the words would sound alike. The result would be very different from the lexicon of a hypothetical language called *Sparsese* in which there were no phonotactic or articulatory constraints at all and in which any phoneme was allowed. In a language like that, lexical neighbors would be few and far between since the word 'Mxyzptlk' would be just as good as 'ma'.

Assessing lexical structure

In this work, we ask whether the lexicon tends toward clumpiness or sparseness. But, because of phonotactics and constraints on the human articulatory system, a naive approach would quickly conclude that the lexicon is clumpy. Natural languages look more like *Clumpish* than they do like *Sparsese* since any given language uses only a small portion of the phonological space available to human language users.[4] We therefore focus on the question of whether lexicons show evidence for clumpiness

[1] There are many existing models that attempt to capture these language-specific rules. A simple model is an n-gram model over phones, whereby each sound in a word is conditioned on the previous n-1 sounds in that word. Such models can be extended to capture longer distance dependencies that arise within words (Gafos, 2014) as well as feature-based constraints such as a preference for sonorant consonants to come after less sonorant consonants (Albright, 2009; Goldsmith & Riggle, 2012; Hayes, 2012; Hayes & Wilson, 2008).

[2] Though as a anonymous reviewer pointed out, some have succeeded in doing so (`https://en.wikipedia.org/wiki/Mister_Mxyzptlk#Pronunciation`)

[3] Note that wordform similarity seems to have a different influence on word learning: phonological probability helps learning but neighborhood density makes it difficult to attend to and encode novel words (Storkel et al., 2006).

[4] As an illustration, English has 44 phonemes so the number of possible unique 2-phone words is $44^2 = 1936$, yet there are only 225 unique 2-phone word forms in English (among all the word forms appearing in CELEX (Baayen et al., 1993), thus only 11% of the space of possible two-phone words is actually used in English (in the absence of any phonotactic

or dispersion above and beyond phonotactics in the *overall* (aggregate) structure of the lexicon.

The basic challenge with assessing whether a pressure for dispersion or clumpiness drives the organization of wordform similarity in the lexicon is that it is difficult to know what statistical properties a lexicon should have in their absence. If we believe, for instance, that the wordforms chosen by English are clumpy, we must be able to quantify clumpiness compared to some baseline. Such a baseline would reflect the *null hypothesis* about how language may be structured in the absence of cognitive forces. Indeed, our methods follow the logic of standard statistical hypothesis testing: we create a sample of null lexicons according to a statistical baseline with no pressure for either clumpiness nor dispersion. We then compute a test measure (e.g., string edit distance) and assess whether real lexicons have test measures that are far from what would be expected under the null lexicons. We present a novel method to compare natural lexicons to phonotactically-controlled baselines that provide a null hypothesis for how clumpy or scattered wordforms would be as the result of only phonotactics. Across a variety of measures, we find that natural lexicons have the tendency to be clumpier than expected by chance (even when controlling for phonotactics). This reveals a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonologically distinct as possible.

2 Method

Assessing the extent to which the lexicons of natural languages are clumpy or sparse requires a model of what wordforms should be expected in a lexicon in the absence of either force.

This idea of developing models to simulate the properties of language has antecedents in the domain of phonology: Previous research developed quantitative models of contrast selection in vowel inventories that are based on maximization of distinctiveness and minimization of stored information (e.g., Liljencrants & Lindblom, 1972). Prior studies looking at the statistics of the lexicon—in particular Zipf's law (Mandelbrot, 1958; Miller, 1957)—have made use of a *random typing* model in which sub-linguistic units are generated at random, occasionally leading to a word boundary when a "space" character is emitted (see e.g., Ferrer-i Cancho & Moscoso del Prado Martín, 2011).

Another line of research (including the present one), goes beyond prior studies in that it takes into account phonotactic constraints, which previous studies did not. By assuming that the sounds composing words are not generated randomly but follow complex constraints (Baayen, 1991; Hayes, 2012), these studies aim at modeling the true generative processes of language (Howes, 1968; Piantadosi et al., 2013). Baayen (1991, 2001) studied wordform similarity in relation to words' frequencies by simulating the lexicon of Dutch through ecologically valid models of language. In particular, Baayen (1991) implemented a model combining a Markov string generator (see Mandelbrot 1958) with a re-use model (see Simon 1955) to generate words. Such a model qualitatively approximates the frequency distribution of words, and importantly for our purpose the neighborhood density of words. However, model selection in Baayen (1991) was performed by evaluating the model's ability to reproduce the properties of the lexicon (i.e., frequency, wordform similarity), thus mixing the properties that may arise by chance – the Markov model can be viewed as a phonotactic model – and the properties that may exist for cognitive reasons – the Simon model can be viewed as an implementation of factors related to language usage (see Baayen (1991); p5).

Here we propose a fundamentally different approach, as we do not select our model based on its ability to reproduce the pattern of wordform similarity in the lexicon as a whole, but rather on its ability to generate candidate words that are scored as having high probability. As such, because our model selection is done independently from the property we are interested in, we can analyze whether

rules).

the properties of the set of words that we obtain through simulation differ from what we observe in the real lexicon, and in what direction.

To accurately capture the phonotactic processes at play in real language, we built several generative models of lexicons: n-grams over phones, n-grams over syllables, and a PCFG over syllables. After training, we evaluated each model on a held-out dataset to determine which most accurately captured each language. The best model was used as the statistical baseline with which real lexicons are compared. We studied monomorphemes of Dutch, English, German and French. Because our baseline models capture effects of phonotactics, we are able to assess pressures for clumpiness or dispersion over and above phonotactic and morphological regularities. Our data is publicly available on GitHub (`https://github.com/SbllDtrch/NullLexicons`).

2.1 Real Lexicons

We used the lexicons of languages for which we could obtain reliably marked morphological parses (i.e., whether a word is morphologically simple like 'glad' or complex like 'dis-interest-ed-ness'). For Dutch, English and German we used CELEX pronunciations (Baayen et al., 1993) and restricted the lexicon to all lemmas which CELEX tags as monomorphemic. The monomorphemic words in CELEX were compiled by linguistic students and include all words that were judged to be nondecomposed. [5] For French, we used Lexique (New et al., 2004), and I.D. (a native French speaker) identified monomorphemic words by hand. (Note that, for Dutch, French and German, these monomorphemic lemmas include infinitival verb endings (*-er* in French, *-en* or *-n* in German and Dutch).)[6] Because we wanted to remove polysemous words (which are morphologically related), we included a phonemic form only once when two words with different spellings shared the same phonemic wordform (e.g., English 'pair' and 'pear' are both pronounced /per/). We did this to be conservative, because it is not clear how to separate homophones (which might be morphologically unrelated) from polysemy. This exclusion accounted for 236 words in Dutch, 646 words in English and 193 words in German. Note that by discarding these words, we already exclude a source of clumpiness in the lexicon.

In order to focus on the most used parts of the lexicon and not on words that are not actually ever used by speakers, we used only those words that were assigned non-zero frequency in CELEX or Lexique. Including these words in the simulation, however, does not change the observed results. All three CELEX dictionaries were transformed to turn diphthongs into 2-character strings in order to capture internal similarity among diphthongs and their component vowels. In each lexicon, we removed a small set of words containing foreign characters and removed stress marks. Note that since we removed all the stress marks in the lexicons, noun-verb pairs that differ in the position of stress were counted as a single wordform in our lexicon (e.g., in English the wordform 'desert' is a noun when the stress in on the first vowel 'désert' but is a verb when the stress is on the last vowel 'desért' but we use only the wordform /desert/ once). These exclusions resulted in a lexicon of 5343 words for Dutch, 6196 words for English, 4121 words for German and 6728 words for French.

2.2 Generative models of Lexicons

In order to evaluate each real lexicon against a plausible baseline, we defined a number of lexical models. These models are all generative and commonly used in natural language processing applications

[5] Note, however, that although we use monomorphemic words, the lexicon may include word pairs that once shared a common morpheme but are no longer analyzed as such.

[6] Removing these verb endings and running the same analysis on the roots did not change the results observed for these 3 languages (but see section 4.2 for an analysis where verb endings matter).

in computer science. The advantage of using generative models is that we can use the set of words of real lexicons to construct a probability distribution over some predefined segments (phones, syllables, etc.) that can be then used to generate words, thus capturing phonotactic regularities.[7] These models are all lexical models, that is, their probability distributions are calculated using word types as opposed to word tokens, so that the phonemes or the syllables from a frequent word like *the* are not weighted any more strongly than those from a less frequent word.[8] We defined three categories of models:

- **n-phone models**: For n from 1 to 6, we trained a language model over n phones. Like an n-gram model over words, the n-phone model lets us calculate the probability of generating a given phoneme after having just seen the previous *n-1* phonemes: $P(x_i|x_{i-(n-1)},...,x_{i-1})$. The word probability is thus defined as the product of the transitional probabilities between the phonemes composing the word, including symbols for the beginning and end of a word. For example, the word 'guitar' is represented as ► g ɪ t ɑː r ◄ in the lexicon where ► and ◄ are the start and the end symbols. The probability of *guitar* considering a bigram model is therefore:

  $$P(\text{g}|\blacktriangleright) \times P(\text{ɪ}|\text{g}) \times P(\text{t}|\text{ɪ}) \times P(\text{ɑː}|\text{t}) \times P(\text{r}|\text{ɑː}) \times P(\blacktriangleleft|\text{r})$$

  These probabilities are estimated from the lexicon directly. For example $P(\text{ɑː}|\text{t})$ is the frequency of tɑː divided by the frequency of t.

- **n-syll models**: For n from 1 to 2, we trained a language model over syllables. Taking the same example as above, 'guitar' is represented as ► gɪ tɑːr ◄ and its probability from a bigram novel over syllables is:

  $$P(\text{gɪ}|\blacktriangleright) \times P(\text{tɑːr}|\text{gɪ}) \times P(\blacktriangleleft|\text{tɑːr})$$

  In order to account for out-of-vocabulary syllables in the final log probabilities, we gave them the same probability as the syllables appearing one time in the training set.

- **Probabilistic Context Free Grammar** (PCFG; Manning & Schutze (1999)): Words are represented by a set of rules of the form $X \rightarrow \alpha$ where X is a non-terminal symbol (e.g., Word, Syllable, Coda) and α is a sequence of symbols (non-terminal and phones). We defined a word as composed of syllables differentiated by whether they are initial, medial, final or both initial and final.

$$\begin{aligned}
\textit{Word} &\rightarrow \textit{SyllableI}\ (\textit{Syllable})^{+}\ \textit{SyllableF} \\
\textit{Word} &\rightarrow \textit{SyllableIF} \\
\textit{Syllable} &\rightarrow (\textit{Onset})\ \textit{Rhyme} \\
\textit{Rhyme} &\rightarrow \textit{Nucleus}\ (\textit{Coda}) \\
\textit{Onset} &\rightarrow \textit{Consonant}^{+} \\
\textit{Nucleus} &\rightarrow \textit{Vowel}^{+} \\
\textit{Coda} &\rightarrow \textit{Consonant}^{+}
\end{aligned}$$

[7] Fine-grained models of phonotactics exist for English (e.g., Hayes (2012)) yet adapting them to other languages is not straightforward and there is no common measure that will allow us to compare their performances.

[8] Using token-based probability estimates instead of type-based probability estimates to capture phonotactic regularities does not change the pattern of results for the 4 languages.

These rules define the possible structures for words in the real lexicon.[9] They are sufficiently general to be adapted to the four languages we are studying, given the set of phonemes for each language. Each rule has a probability that determines the likelihood of a given word. The probabilities are constrained such that for every non-terminal symbol X, the probabilities of all rules with X on the left-hand side sum to 1: $\sum P(X \rightarrow \alpha) = 1$. The likelihood of a given word is thus the product of the probability of each rule used in its derivation. For example, the likelihood of 'guitar' is calculated as the product of all probabilities used in the derivation of the best parse (consonant and vowel structures are not shown for simplification):

```
Word → SyllableI(Onset(g) Rhyme(Nucleus(ɪ)))
SyllableF(Onset(t) Rhyme(Nucleus(ɑː) Coda(r)))
```

The probabilities for the rules are inferred from the real lexicon using the Gibbs sampler used in Johnson et al. (2007) and the parse trees for each word of the held-out set are recovered using the CYK algorithm (Younger, 1967).

2.3 Selection of the best model

To evaluate the ability of each model to capture the structure of the real lexicon, we trained each model on 75% of the lexicon (the training set) and evaluated the probability of generating the remaining 25% of the lexicon (the validation set). This process was repeated over 30 random splits of the dataset into training and validation sets. For each model type, we smoothed the probability distribution by assigning non-zero probability to unseen n-grams or rules in the case of the PCFG. This was to allow us to derive a likelihood for unseen but possible sequences of phonemes in the held-out set. Various smoothing techniques exist, but we focus on Witten-Bell smoothing and Laplace smoothing which are straightforward to implement in our case.[10] All smoothing techniques were combined with a backoff procedure (though not for the PCFG), such that if the context AB of a unit U has never been observed ($p(U|AB) = 0$) then we can use the distribution of the lower context ($p(U|B)$). The smoothing parameter was set by doing a sweep over possible parameters and choosing the one that maximized the probability of the held-out set. The optimal smoothing was obtained with Laplace smoothing with parameter .01 and was used in all models described.

In order to compare models, we summed the log probability over all words in the held-out set. The model that gives the highest log probability on the held-out data set is the best model, in that it provides a "best guess" for generating random lexicons that respect the phonotactics of the language.

[9] Because of space considerations, we do not present the rules for *SyllableI*, *SyllableF* and *SyllableIF*. They follow the same pattern as the non-terminal *Syllable*.

[10] Other smoothing techniques such as Good Turing or Kneser-Ney cannot be implemented easily as they rely on the number of units for which frequency is equal to one, which is not available in every model we tested.

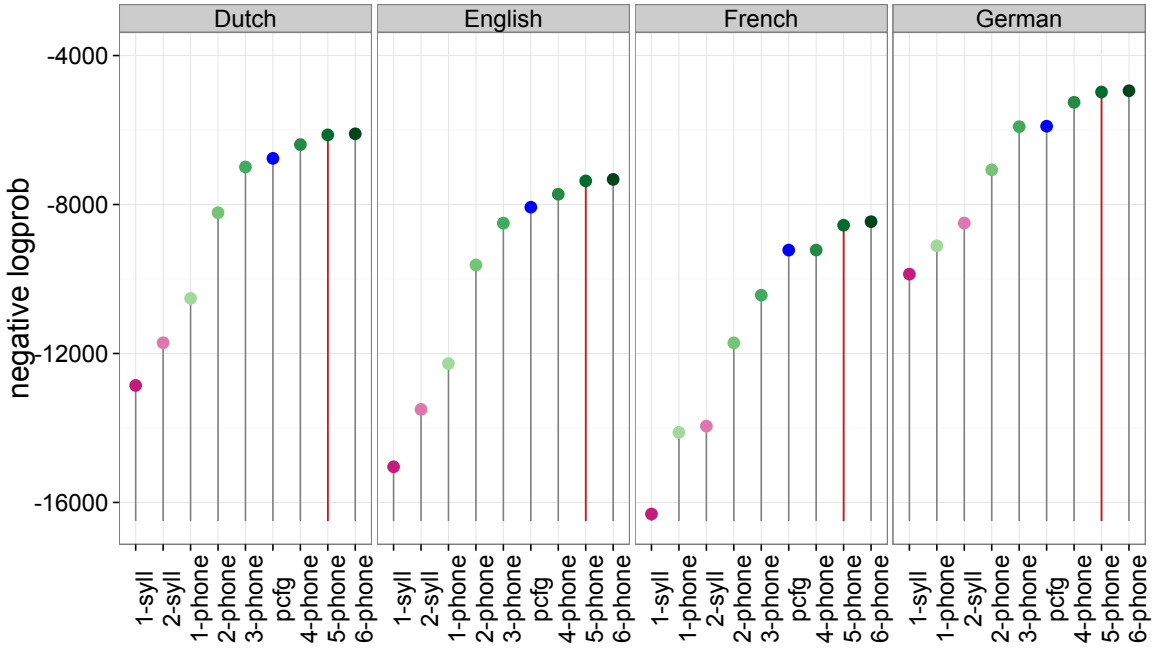

Figure 1: Each point represents the mean log probability of one model to predict the held-out data set. The n-phone models are represented in green, the n-syll models in pink and the PCFG in blue. The 5-phone model has the highest log probability (indicated by a red segment) for all languages. The standard deviation of the mean is presented in each, but is too small to be visible at this scale.

As shown in Figure 1, the 5-phone model gives the best result for all lexicons. In all cases, the 6-phone was the next best model, and the 4-phone was close behind, implying that n-phone models in general provide an accurate model of words. The syllable-based models performed particularly poorly. Thus, we focus our attention on the 5-phone model in the remainder of the results, treating this as our best guess about the null structure of the lexicon (see the Supplemental material, for a robustness check of our results across the 3 best models according to our evaluation).

2.4 Building a baseline with no pressure for clumpiness or dispersion

We use the 5-phone model to generate simulated null lexicons—ones without any pressure for clumpiness or dispersion other than the 5-phone generating process—and study the position of the real lexicon with respect to the simulated ones. For each language, we trained the 5-phone model on the entire real lexicon and used the resulting language model to generate words for 30 simulated lexicons. It is simplest to visualize how word generation works for the 1-phone case. In such case, all the phones of a given language cover the entire probability space from 0 to 1, each phone covering an interval proportional to its frequency in the real lexicon. We pick a random number between 0 and 1 and and select the phone that corresponds to that value. Phones are generated until the we randomly generate the end-symbol. For the 5-phone model, the same technique is applied except that each phone generation is constrained by the last 4-phones of the word: We first generate a random 5-phone sequence starting with 4 start-symbols, then we generate the next 5-phone sequence to follow given the last 4 phones of

the word according to the sequence probability, and so on until the end-symbol is encountered.

The number of words generated for each simulated lexicon was matched to the number of words in the corresponding real lexicon. We additionally constrained the generation to ensure that the distribution of word lengths in each simulated lexicon matches the distribution of word lengths in the real lexicon and that, similarly to real lexicons, the simulated lexicon contained no homophones. Practically, it means that we discarded a word every time we generated a word did not match the distribution of word lengths of the real lexicon (either because all words of that length have already been generated, or because that length did not exist in the real lexicon) or a word that already existed in the simulated lexicon.

On average our best lexicon model generated 52% real words for Dutch, 53% for English, 47% for French, and 41% for German. Note that it is not surprising that the best lexicon model generates *only* about 50% of real words since the smoothing parameter allowed the generation of non-words likely to be attested in the language.

3 Results: Overall similarity in the lexicon

To compare real and simulated lexicons, it is necessary to define a number of test statistics that can be computed on each lexicon to assess how it uses its phonetic space. As in null hypothesis testing, we compute a z-score using the mean and standard deviation estimated from the 30 lexicons generated by our best lexicon model. We then ask whether the real lexicon value falls outside the range of values that could be expected by chance under the null model. The p-value reflects the probability that the real lexicon value could have arisen by chance under our chosen 5-phone null model.[11]

We present results separately for a number of different measures of wordform similarity: minimal pairs, Levenshtein distance, and several network measures.

3.1 Minimal pairs

We first considered the number of minimal pairs present in each lexicon. A minimal pair is a pair of words of the same length for which a single sound differs (e.g., 'cat' and 'rat'). If real lexicons are clumpier than expected by chance, then the real lexicons should have more minimal pairs than their simulated counterparts. If they are more dispersed, the real lexicons will have fewer minimal pairs.

Figure 2 summarizes this hypothesis test, showing how the various simulated lexicons compare to the real lexicons in terms of number of minimal pairs for each language. Each histogram represents a distribution of minimal pair counts broken up by language across the 30 simulated lexicons. The red dot represents the real lexicon value and the dotted lines represent the 95% confidence interval. All histograms fall to the left of the red dot, which suggests that the real lexicon has more minimal pairs than any of the simulated ones in all four languages (all $ps < .001$; see Table 1). This pattern suggests that lexicons are clumpier than expected by chance.

[11] Note that while doing so we assume that the different test statistics we are measuring (different measures of wordform similarity), are distributed normally (which is reasonably the case, see Figures below). The advantage of computing a z-score over doing a permutation test is that we can work with a reasonable number of random lexicons. Indeed, in a permutation test, the p-value is calculated as the proportion of random lexicons where a given measure of wordform similarity would be greater or less than the actual value we found in the real lexicon and require thus substantially more random lexicons (typical permutations analyses use 1,000 or 10,000 permutations).

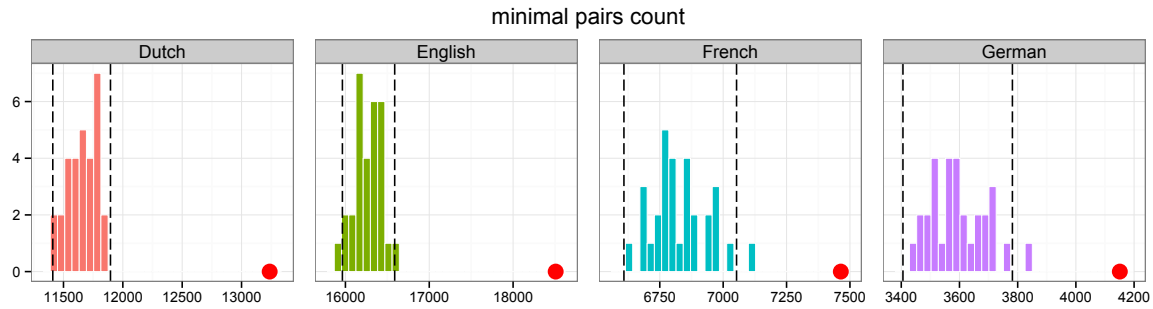

Figure 2: Comparison of the total number of minimal pairs for each language (red dot) to the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. For all four languages, the real lexicon has significantly more minimal pairs than predicted by our baseline.

	Dutch	English	French	German
real	13,237	18,508	7,464	4,151
μ (simulated)	11,653	16,276	6,830	3,594
σ (simulated)	124	159	113	96
z	12.77	14.03	5.61	5.80
p	$<$**.001**	$<$**.001**	$<$**.001**	$<$**.001**

Table 1: z-statistics comparing the total number of minimal pairs in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of minimal pairs counts in the 30 simulated lexicon for each language.

To see whether this effect is driven by words of specific lengths, we looked at the number of minimal pairs for each length. We concentrated on words of length 2 to 7 which represent more than 90% of all words in each language. As shown in Figure 3, the real lexicon has more minimal pairs than the simulated ones consistently across words of any length. For all languages, the effect is larger for words of smaller length (length 3 to 4; 30% to 50% of all words in each language) where most minimal pairs are observed. The smaller effect for longer words (especially words of length 7 and above) is likely due to a floor effect since longer words are far less likely to have minimal pairs than short words. Note that, for words of length 2, we see a somewhat degenerate case since there are relatively few possible 2-phoneme words, yet for at least 3 languages it appears that there are more minimal pairs of length 2 than what would be expected by chance. This is explained by the smoothing parameter of the model that allows the generation of unseen sequences of sounds (recall that we smoothed the probability distribution to account for rare sequences of sounds that may be unseen in the lexicon of monomorphemes). As a result, the model does not reproduce all the 2-phoneme words of the languages.[12]

[12]Inspection of these 2-phoneme words reveals that most of these words are actual wordforms present in the language (hence attested forms, e.g. "is" in English) but are not counted as distinct monomorphemic lemmas and thus are not included in our real lexicons.

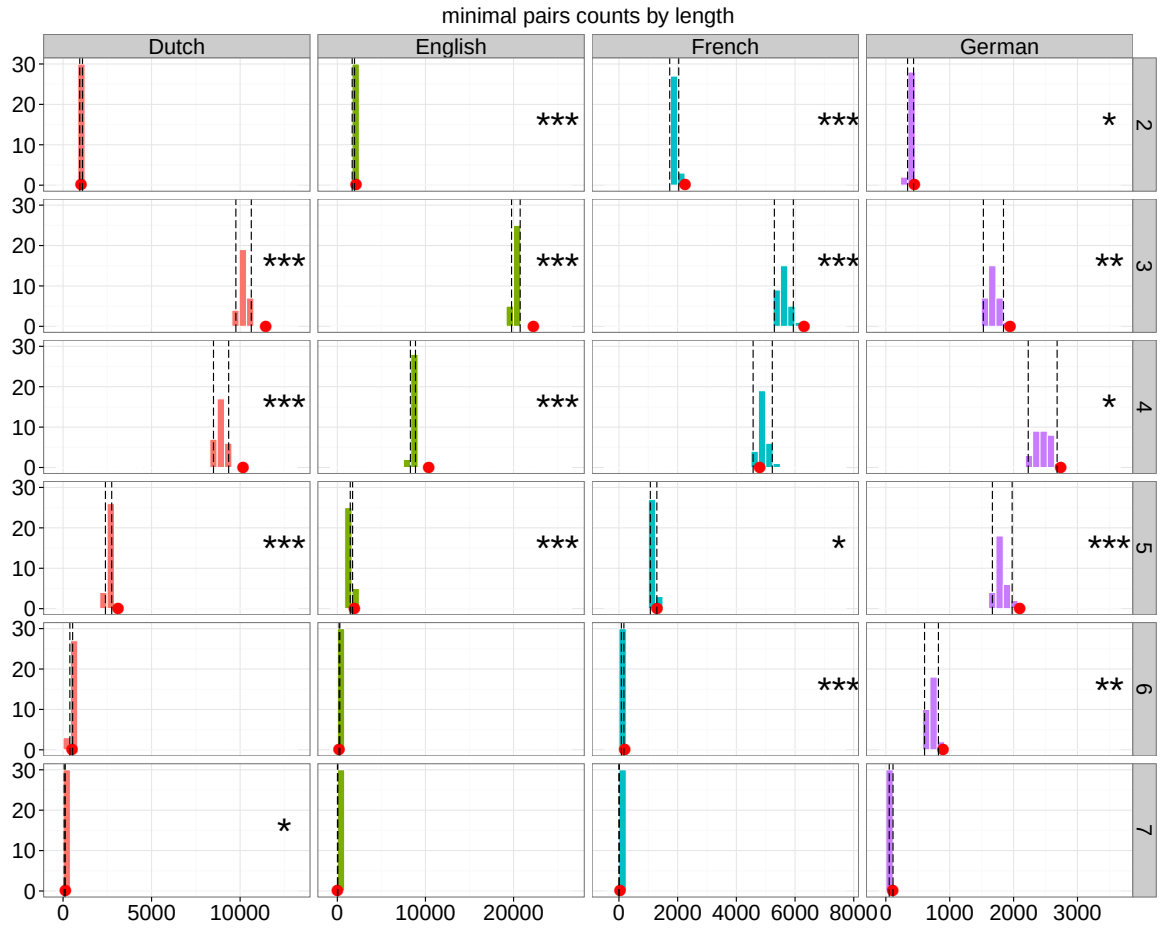

Figure 3: Comparison of the number of minimal pairs by word length (2-7) for each language (red dots) to the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. $* = p < .05$, $** = p < .01$, $*** = p < .001$.

3.2 Levenshtein distance

We can evaluate clustering using more global measures by considering the average string edit distance (*Levenshtein distance*) between words (Levenshtein, 1966). The Levenshtein distance between two sound strings is simply the number of insertions, deletions and replacements required to get from one string to another. For instance, the Levenshtein distance between 'cat' and 'cast' is 1 (insert an 's'), and it is 2 between 'cat' and 'bag' (c $\rightarrow$ b, t $\rightarrow$g). To derive a measure of Levenshtein distance that summarizes the whole lexicon, we compute the *average Levenshtein distance* between words in the lexicon by simply computing the distances between every pair of words in the lexicon and then averaging these distances.[13] If the lexicon is clumpier than expected by chance, words will tend to be more similar to one another and we expect to observe a smaller average Levenshtein distance. In

[13] A possible objection to using Levenshtein distances is that there is little apparent difference in phonological confusability between a pair like 'cats' and 'bird', which has a Levenshtein distance of 4, and a pair like 'cats' and 'pita,' which has a Levenshtein distance of only 3 but which is arguably even more different since it differs in syllable structure. Ultimately, neither pair is especially confusable: the effects of phonological confusability tail off after 1 or 2 edits.

contrast, a larger average Levenshtein distance in the real lexicons relative to the simulated lexicons would suggest that the lexicon is more dispersed than expected by chance.

As shown in Figure 4, the average Levenshtein distance between words is significantly smaller for the real lexicon than in the simulated lexicons for all four languages (see Table 2). The difference is numerically small, but that is to be expected because minimal pairs are statistically unlikely. That is, the edit distance between two words is largely a product of their lengths. For example, on average, the edit distance between two 5-letter words is close to 5. Nonetheless, the Levenshtein metric provides us with an additional piece of evidence that words in the real lexicons are more similar to each other than what would be expected by chance.

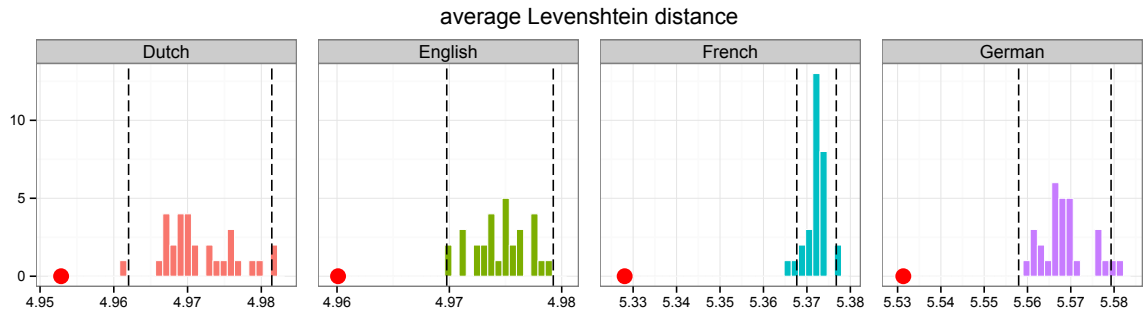

Figure 4: Distribution of average Levenshtein distances for each of the 30 simulated lexicons. The red dot represents the real lexicon's value, and the dotted lines are 95% confidence intervals.

	Dutch	English	French	German
real	4.95	4.96	5.32	5.53
μ (simulated)	4.97	4.97	5.34	5.57
σ (simulated)	0.005	0.002	0.002	0.005
z	-3.80	-6.0	-6.2	-6.9
p	$<$**.001**	$<$**.001**	$<$**.001**	$<$**.001**

Table 2: z- statistics comparing the average Levenshtein distance in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of average Levenshtein distance in the 30 simulated lexicon for each language.

Similarly, to see whether this effect is driven by words of specific lengths, we looked at the average Levenshtein distance for words of length 2 to 7. As shown in Figure 5, the real lexicon has a smaller average Levenshtein distance than the simulated ones consistently across words of most lengths in all languages.

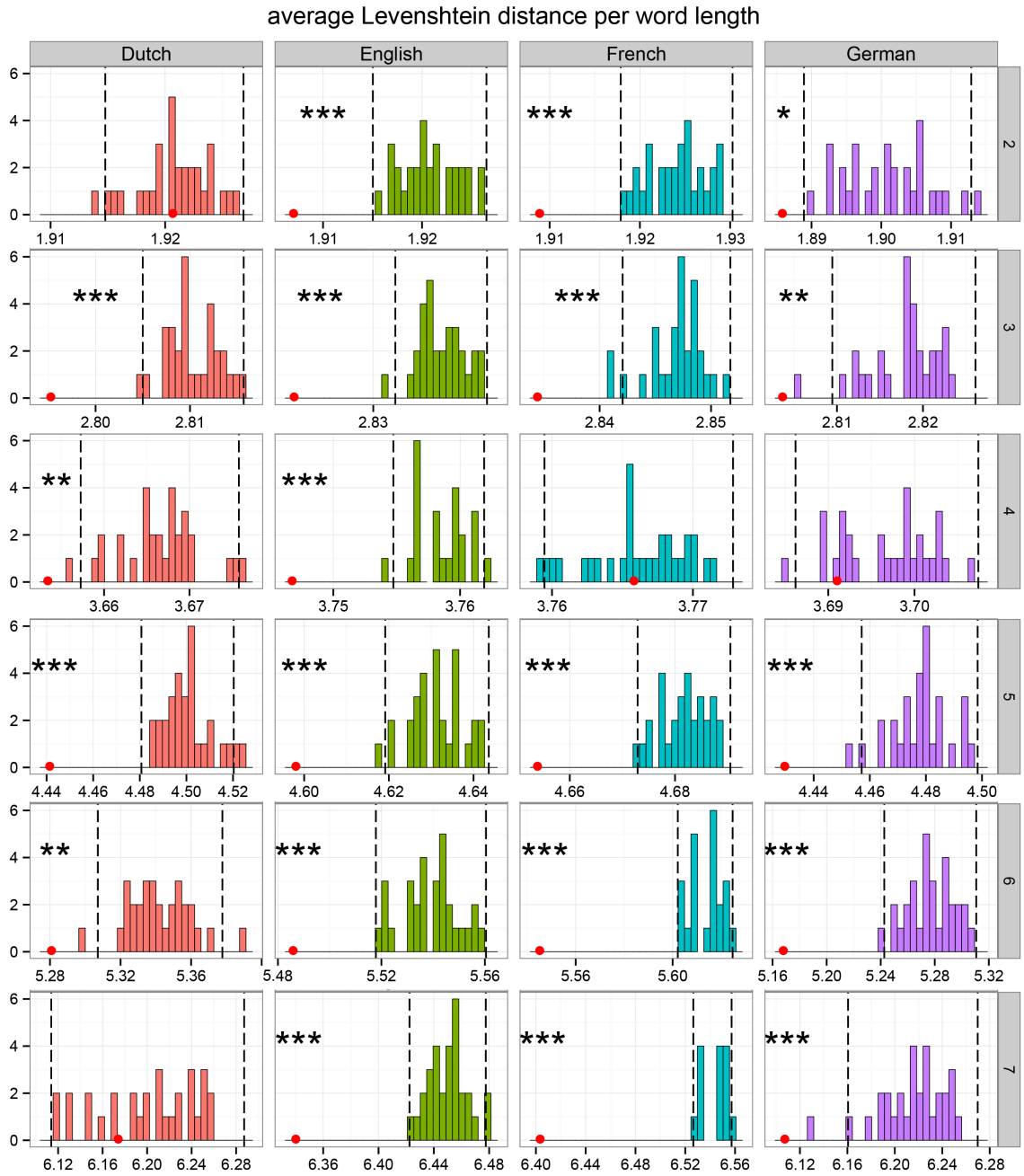

Figure 5: Average Levenshtein distance by word length (2-7) for each language (red dots) compared to the distribution of average Levenshtein distance obtained across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. $* = p < .05$, $** = p < .01$, $*** = p < .001$.

3.3 Network measures

Simply calculating phonological neighbors, however, does not tell us everything about how wordforms are distributed across a lexicon. Perhaps some words have many neighbors while others have few. Or

it could be the case that neighbor pairs tend to be more uniformly distributed across the lexicon. To answer these questions, we constructed a phonological neighborhood network as in Arbesman et al. (2010), whereby we built a graph in which each word is a node and all phonological neighbors are connected by an edge, as in the toy example in Figure 6, which represents a lexicon of 14 words.

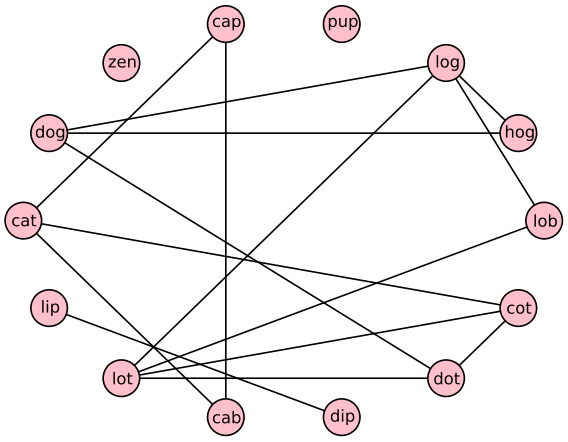

Figure 6: Example phonological network. Each word is a node, and any words that are 1 edit apart are connected by an edge.

Figure 7 shows examples of such networks for English 4-phone words, where each word is a node, with an edge drawn between any two words that are phonological neighbors (1 edit away). Words with no or few neighbors tend to be clustered on the outside. (The ring of points around the perimeter of the circle represent the isolates–words with no neighbors.) Words with many neighbors are, in general, plotted more centrally.

We compared the shape of lexicons generated by different models to the real lexicon. As can be seen in Figure 7, the 5-phone model most closely resembles the real lexicon. Substantially more clustering is observed in the more restrictive generative models: the 5-phone, 2-syll and PCFG models have many more connected neighbors than a 1-phone model. This corresponds to the fact that many more words are possible in the 1-phone model (e.g. 'cktw' is a possible word), than in a more constrained model that respects phonotactics. Therefore the space is largest in the 1-phone model, and the probability of generating a word that is a neighbor of a previously generated word is correspondingly lower. Crucially, however, the real lexicon seems even clumpier overall than the lexicons produced by any of the generative models.

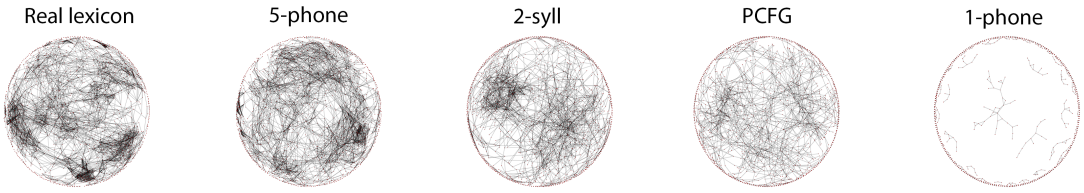

Figure 7: Sampling of phonological neighbor network from the different generative models applied on all 4-phone wordforms of the English lexicon. Each point is a word, and any two connected words are phonological neighbors. The simulated lexicons from less constrained generative models are less clustered and have more isolates (words with no neighbors, plotted on the outside ring).

Using techniques from network analysis that have been fruitfully applied to describe social networks and other complex systems (Barabási & Albert, 1999; Wasserman & Faust, 1994; Watts & Strogatz, 1998), we can quantitatively characterize the clustering behavior of the lexicon. We computed the *transitivity*, *average clustering coefficient*, and the proportion of nodes in the *giant component*. All three of these measures can be used to evaluate how tightly clustered the words in the lexicon are. A graph's *transitivity* is the ratio of the number of triangles (a set of 3 nodes in which each node in the set is connected to both other nodes in the set) to the number of triads (a set of 3 nodes in which at least two of the nodes are connected). Transitivity therefore quantifies how likely it is that A is connected to C, given that A is connected to B and B is connected to C. The *clustering coefficient* of a node is a closely related measure is defined as the fraction of possible triangles that *could* go through a node that actually do go through that node. We can then compare the average clustering coefficient across networks. Both transitivity and average clustering coefficient measure the extent to which nodes cluster together. The largest cluster in a network is known as the *giant component*. A network with many isolated nodes will have a relatively small giant component, whereas one in which nodes are tightly clustered will have a large giant component. These measures give us some insight into the internal structure of the lexicon, over and above those obtained by looking at more global measures such as the number of minimal pairs and the average Levenshtein distance.

Previous studies (Arbesman et al., 2010; Vitevitch, 2008) showed that phonological networks across many languages (i.e., English, Spanish, Mandarin, Hawaiian, and Basque) exhibit several interesting properties, notably that these networks display a higher average clustering coefficient and higher transitivity than *random* networks. Random networks in these studies are networks having the same number of nodes and the same average number of connection per nodes but whose connections between the nodes have been randomly placed. However, such random networks do not reflect what the lexical graph would be like if there were no cognitive pressures on the lexicon as they do not take into account the processes by which languages are generated (e.g., phonotactics).

Here, we provide an answer to this concern by comparing the lexicon of natural languages to a statistical model (our 5-phone model), our chance baseline, that reflects the phonotactic processes of the language. If the real lexicon is clumpier than expected by chance, we predict that, relative to the simulated lexicons, the real lexicons will show higher transitivity, higher average clustering coefficients, and a larger proportion of words in the giant component.

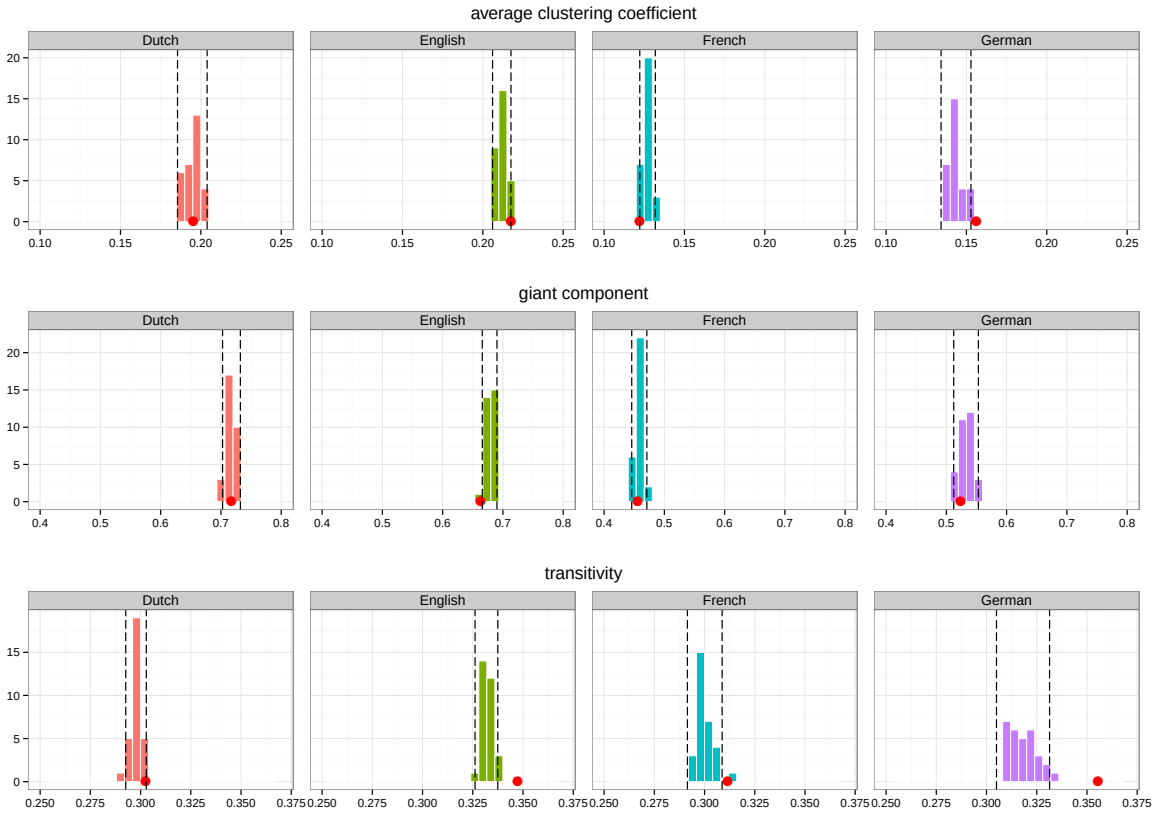

Figure 8: Distributions of our best generative model (the histograms) compared to the real lexicon (the red dot) in terms of network measures for lexical networks (where each node is a word and any 2 nodes that are minimal pairs are joined in the network): the average clustering coefficient, the proportion of nodes in the giant component, and transitivity.

		Dutch	English	French	German
Average Clustering coefficient	real	0.2	0.22	0.12	0.16
	μ (simulated)	0.19	0.21	0.13	0.14
	σ (simulated)	0.005	0.003	0.002	0.005
	z	0.1	2	-2	2.7
	p	0.9	.05	.05	$<$**.01**
Giant component	real	0.72	0.66	0.46	0.52
	μ (simulated)	0.72	0.68	0.46	0.53
	σ (simulated)	0.008	0.006	0.006	0.01
	z	-0.1	-2.4	-0.4	-0.9
	p	0.9	$<$.05	0.7	0.4
Transitivity	real	0.3	0.35	0.31	0.36
	μ (simulated)	0.3	0.33	0.3	0.32
	σ (simulated)	0.003	0.003	0.004	0.007
	z	1.8	5.4	2.6	5.5
	p	0.07	$<$**.001**	$<$**.05**	$<$**.001**

Table 3: z- statistics comparing various network measures (Average clustering coefficient, proportion of words in the giant component, transitivity) in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of these measures in the 30 simulated lexicon for each language.

As observed in Figure 8, there is no systematic difference between the real lexicon and the simulated ones regarding the average clustering coefficient measures and the proportion of nodes in the giant component. Yet there is a significant effect of transitivity (see Table 3). The reason that average clustering coefficient shows less of an effect than transitivity is likely due to the fact that the average clustering coefficient is more dependent on low-degree nodes, like the many isolates that exist for longer words in lexical networks. The transitivity measure avoids this problem and can be viewed as a normalized clustering coefficient (Newman, 2003). The lack of effect for the giant component measure may simply be because the proportion of words in the giant component is not a particularly robust measure since it can be dramatically shifted by the addition or deletion of one or two key neighbors. The higher transitivity, however, suggests that in addition to having more overall neighbors in the real lexicons, the neighborhoods themselves are more well-connected than the neighborhoods in simulated lexicons are. That is, if two words A and B are both neighbors of word C, A and B are themselves more likely to be neighbors in the real lexicon than they are in the simulated lexicons.

4 Results: Finer-grained patterns of similarity in the lexicon

Across a variety of measures, we found that wordforms tend to be more similar than expected by chance across all languages under study. Yet, while wordform similarity might be explained by a variety of cognitive advantages (see Introduction), it does not necessarily follow that the lexicon is not subject to communicative pressure favoring wordform distinctiveness. A possibility is that the similarity among wordforms may not be uniformly distributed across the real lexicon but may be constrained by other dimensions that maximize their distinctiveness in the course of lexical processing,

such as:

1. **phonological distinctiveness:** Not every pair of phonemes is equally confusable. For instance, a minimal pair like 'cap' and 'map' are unlikely to be confused since /k/ and /m/ are quite distinct. But 'cap' and 'gap' differ by only the voicing of the first consonant and are thus much more confusable (see e.g., Gahl & Strand, 2016, for evidence that perceptual phonological neighborhood density produces an effect on spoken word recognition over and above phonological neighborhood density based on segment difference). This more subtle contrast is potentially much more troublesome for communication and is therefore more likely to be avoided. So even though the number of minimal pairs is higher than expected by chance in natural lexicons, this might not be problematic for communication as long as they are not based on confusable contrasts.

2. **grammatical categories:** Not every pair of words is equally confusable. For instance, nouns (e.g. 'berry') are more likely to be confused with other nouns (e.g., 'cherry') than words from another grammatical category (e.g., the intensifier 'very') because they appear in a noun syntactic context which constrains listeners to expect a noun in this position (see e.g., Strand et al., 2014; Viebahn et al., 2015) for evidence of phonological competition among words of the same grammatical category). Therefore, from a communicative point of view, there should be more minimal pairs distributed across syntactic categories than within the same syntactic category to minimize the risk of miscommunication.

In the following we test how the simulated lexicons compare to the real lexicons along these two dimensions.

4.1 Wordform distinctiveness in minimal pairs

The accurate recognition of a word depends on the distinctiveness of the phonological contrasts distinguishing words. For instance, it is easier to confuse 'cap' (/kap/) with 'gap' (/gap/) than with 'map' (/map/) simply because /k/ is more similar to /g/ than to /m/. If lexicons aim to minimize confusability, they should prefer distinctive contrast minimal pairs (e.g., 'cap'/'map') as opposed to confusable ones (e.g., 'cap'/'gap'). Thus, it is possible that lexicons can have the learning benefit of having many minimal pairs, as long as they are not based on confusable contrasts.

To evaluate this hypothesis, we looked at the 5% most frequent minimal pair contrasts and derived a measure of confusability for these contrasts. Phonemes can be characterized by their phonological features: For consonants, place of articulation (e.g., labial, dental, palatal), manner of articulation (e.g., stop, fricative, glides) and voice (voiced, unvoiced); For vowels, height (low, mid, high), backness (front to back) and roundness. For each of the 5% most frequent pairs of contrasts, we calculated the difference in phonological features between each member of the pair. For example the pair /k/ and /m/ has 3 features that differ: place, manner and voicing. Bailey & Hahn (2005) found that the number of non-matching major articulatory features between two sounds is a good measure of phonemic similarity between those sounds. Thus, the test statistic that we use here is the average number of features that differ in a minimal pair. This measure ranges from 1 (highly confusable) to 3 (highly distinguishable).[14]

Figure 9 shows the average number of features that differ in the 5% most frequent minimal pair contrasts in the real lexicon and across all simulated lexicons for each language. The minimal pairs

[14]For French we added nasalization as a vowel feature. The measure for French vowel contrasts therefore ranged from 1 to 4.

contrasts in the real lexicon are no more distinguishable in phonetic space than are the minimal pairs in the chance lexicon. This indicates that minimal pairs do not rely on more perceptible contrasts for distinctiveness than what is expected by phonotactics alone.

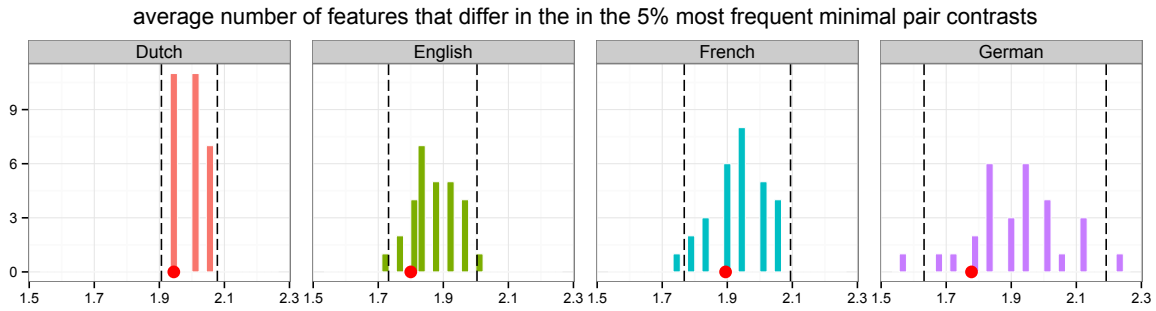

Figure 9: Distributions of the average number of feature difference for the 5% most frequent minimal pair contrasts in the simulated lexicon compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. There is no evidence that these frequent contrasts are more perceptible than expected by chance (all $ps > .30$).

The previous measure showed that frequent minimal pair contrasts are not more perceptible than expected by chance alone. However, although the average number of different phonological features between two minimally different words is a good predictor of their acoustic similarity (Bailey & Hahn, 2005), there is evidence that the *nature* of phonological features differing between two words, a finer grained measure, is also an important factor for words confusion (e.g., Cole et al., 1978; Martin & Peperkamp, 2015; Miller & Nicely, 1955). In addition, we looked only at the most used contrasts and not at the whole range of contrasts available. Thus, it could still be the case that a more perceptual and language-specific measure of phoneme confusability–looking at a broader range of possible contrasts–would be a better predictor of clumpiness. If the lexicons prefer minimal pairs to be distinctive then we should observe more minimal pairs with easily perceptible contrasts than with confusable contrasts. In order to investigate this possibility, we looked at minimal pairs in English, for which confusability data between phonemes are readily available (Miller & Nicely, 1955). We computed the distance between the mean number of minimal pairs in our simulated lexicons and the number of minimal pairs in the real lexicon for each of the 120 contrasts present in the Miller and Nicely dataset. The distance is simply the difference between a) the mean number of minimal pairs in the simulated lexicons and b) the number of minimal pairs in the real lexicon, divided by the standard deviation of the value across the 30 simulated lexicons. In effect, this acts as a z-score that tells us how far the real lexicon value falls from what we expect under a null model.

Figure 10 shows the z-score obtained for each phonemic contrast as a function of its confusability (the higher the more confusable). As can be observed, the z-score is uniformly above 0 which is in line with previous results showing that there are more minimal pairs in lexicons than expected by chance. Yet, crucially there is no effect of confusability on the z-score ($p > 0.5$). That is, there is no evidence that the English lexicon is more clumpy around highly distinctive contrasts than around highly confusable contrasts.

One obvious limitation of this analysis is that we have only looked at a single language. In addition, the original study of Miller & Nicely (1955) examined the confusion between consonants

in pre-lexical perception by playing a noisy version of the VCV (Vowel-Consonant-Vowel) materials to participants. Thus there are two reasons why the confusion matrix may be not appropriate for our case. First, the spectral noise used in this study may interfere with the recognition process, so that this confusion matrix may reflect the recognition of sounds in noise, rather than the general perception of sounds. Second, it may be the case that it is not pre-lexical confusion between phonemes that matters but rather lexical confusion, as shown by e.g., Ernestus & Mak 2004; Miller & Nicely 1955.

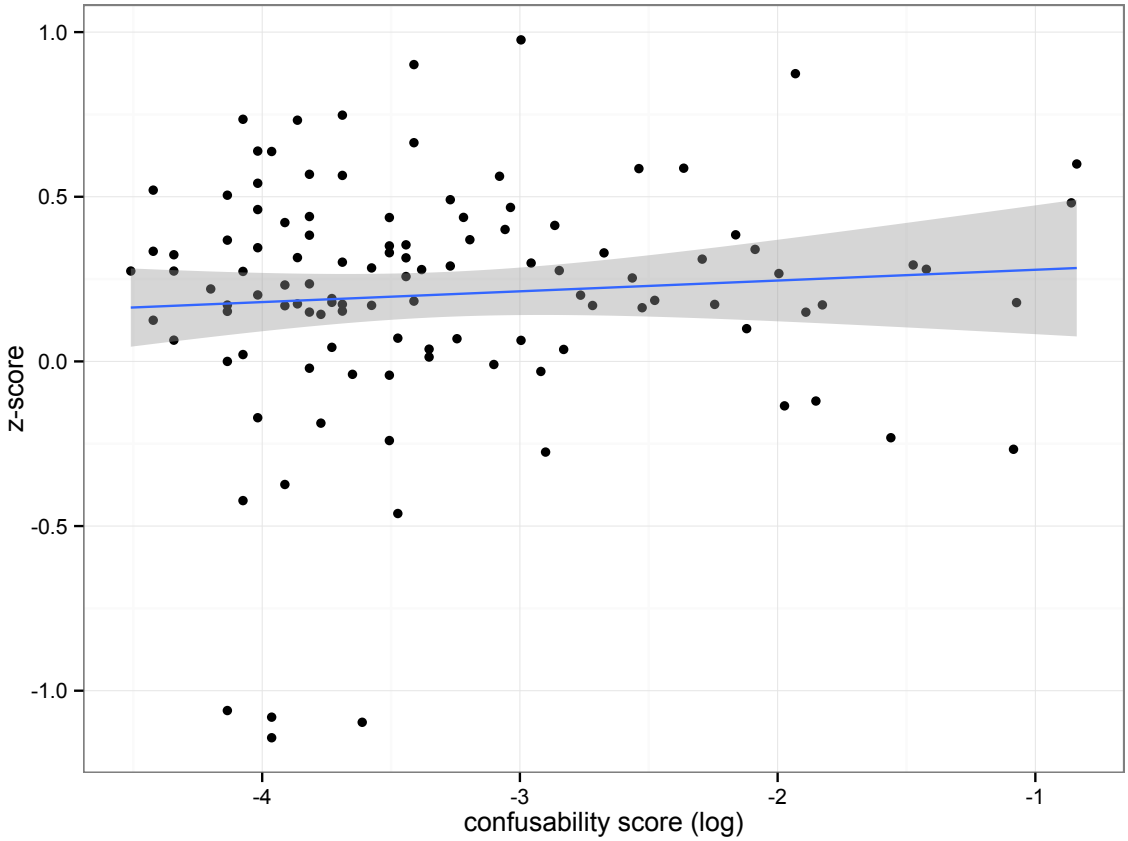

Figure 10: z-score obtained between the mean number of minimal pairs in the real lexicon and in the simulated lexicons for each of the minimal pair contrasts present in the Miller and Nicely's dataset as a function of their log confusability.

Despite the limitations of these two analyses taken separately, it appears that the clumpiness effect is driven not just by highly distinct sound sequences but is present even when considering highly confusable sounds. This points to a pressure for lexical clumpiness which may work against robust communication.

4.2 Wordform similarities within and across grammatical categories

Words do not usually appear in isolation but are embedded in richer linguistic contexts. A wealth of studies show that adults and children use the context of a sentence to constrain lexical access (e.g., Altmann & Kamide, 1999; Borovsky et al., 2012). Hence even if the lexicon is clumpy as a whole, the context might be sufficient to disambiguate between two similar wordforms. Several studies have

shown that the grammatical context of a word may constrain lexical access to grammatically appropriate competitors of the word (Strand et al., 2014; Viebahn et al., 2015). Thus, one obvious contextual disambiguation is the syntactic category of words. For example, consider the sentence "did you see my sock?" The chance that a native English speaker might confuse the word 'sock' with 'wok' in the context of following 'my' might be greater than confusing 'sock' with 'mock', because 'wok' is a noun–which is consistent with the syntactic context–whereas 'mock' is a verb, which is inconsistent with the syntactic context. Moreover, because children as young as 18-months have been shown to use function words to recognize verbs and nouns on-line (Cauvet et al., 2014; Dautriche, Fibla, & Christophe, 2015; de Carvlaho et al., 2015), these sorts of categorizing effects may be crucial to language acquisition.

As with the lexicon more broadly, there are two possible outcomes that could arise from comparing word forms within as opposed to across syntactic categories. On the one hand, because context is usually enough to distinguish among different parts of speech, confusability of words should be less of a problem across syntactic categories. That is, even though 'bee' and 'see' are minimal pairs, one is unlikely to misperceive "I was just stung by a *bee*" as "I was just stung by a *see*." This account predicts more similarity across syntactic categories than within syntactic categories. On the other hand, increased similarity between words of the same part of speech, i.e., having nouns that sound like other nouns and verbs that sound like other verbs, could convey a processing and a learning advantage (Monaghan et al., 2011). Under this account, we would expect more similarity within as opposed to between syntactic category.

For this evaluation, we used the Part Of Speech (POS) tags in CELEX for Dutch, English and German and in Lexique for French to count the number of minimal pairs within the same syntactic categories (e.g., 'wok' / 'sock') and across different syntactic categories (e.g., 'mock' / 'sock'). For each simulated lexicon, we randomly assigned the syntactic categories of real words of length n to generated words of length n and similarly counted the number of minimal pairs appearing within and across the same syntactic categories.[15] Note that for wordforms having several syntactic categories in the real lexicon (homophones, e.g.,'seam'/'seem' which are counted as a single wordform in our lexicons, /sim/), we chose the syntactic category of the most frequent item (e.g., because the most frequent meaning of /sim/ is 'seem' it will be categorized as a verb). Because there are more across-category word pairs than within-category word pairs across languages, we compared the probability of finding a minimal pair, across-categories or within category. These probabilities were obtained by dividing the number of minimal pairs appearing across and within categories by the number of across- and within-category word pairs respectively. The final measure is thus the probability of getting a minimal pair, across categories or within a category.

[15]This was to ensure that certain categories, such as pronouns, which are reserved for smaller words will not be assigned to longer words.

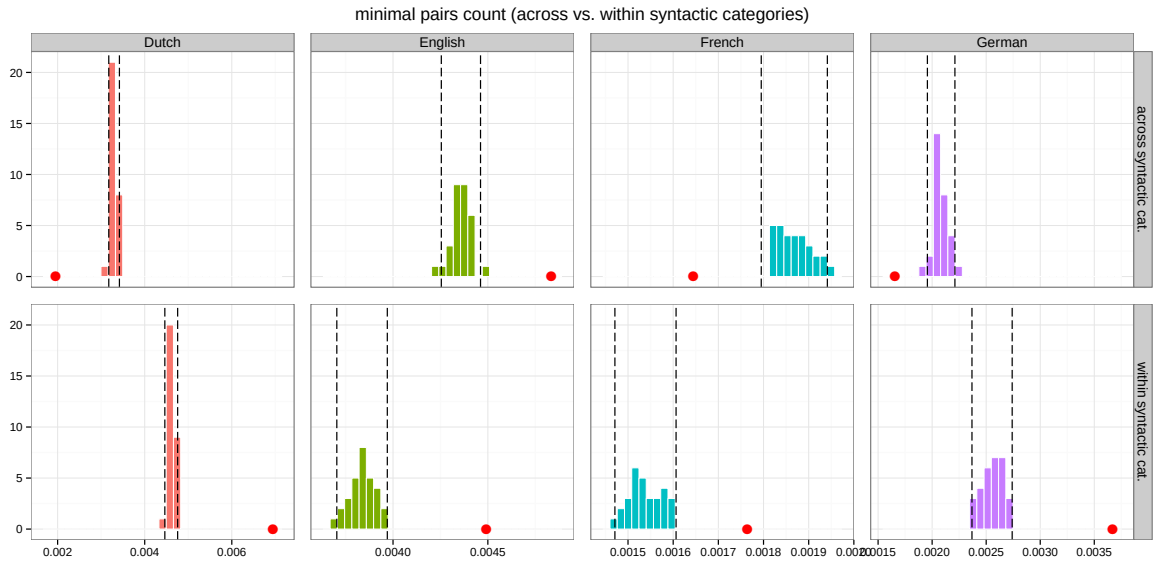

Figure 11: Distributions of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. All 4 languages are significantly more likely to have minimal pairs within categories than would be expected by chance.

		Dutch	English	French	German
across syntactic categories	real	0.002	0.0048	0.0016	0.0017
	μ (simulated)	0.0033	0.0044	0.0019	0.0021
	σ (simulated)	1e-04	1e-04	1e-05	1e-04
	z	-21.5	9	-6	-6.6
	p	**<.001**	**<.001**	**<.001**	**<.001**
within syntactic categories	real	0.0069	0.0045	0.0018	0.0037
	μ (simulated)	0.0046	0.0038	0.0015	0.0026
	σ (simulated)	1e-04	1e-04	1e-05	1e-04
	z	31.2	9.6	6.5	11.7
	p	**<.001**	**<.001**	**<.001**	**<.001**

Table 4: z- statistics comparing the probability of getting a minimal pair within and across syntactic categories in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of the probability of having a minimal pair in the 30 simulated lexicon for each language. The red p-values shows a significant effect of clumpiness and the blue ones a significant effect in the opposite direction.

As before, we compare the real lexicon to the simulated lexicons but break the measures down by similarity within syntactic category (only looking at the similarity of nouns to other nouns, verbs to other verbs, and so on) and between syntactic category (only looking at the similarity of nouns to non-nouns, verbs to non-verbs, etc.). As shown in Figure 11, we found that there are *more* minimal

pairs within the same syntactic category in the real lexicons than would be expected by chance for all 4 languages. That is, for within syntactic category analyses, all four languages are clumpier than expected under the null models. For the across-category analysis, the result is less clear. For French, German, Dutch, there are *fewer* minimal pairs across different syntactic categories than would be expected by chance. For English, there are more across-category minimal pairs than expected by chance.

A subsequent post-hoc analysis found that the unclear results for the across-category analysis can in part be explained by the infinitival affixes that appear on French, Dutch, and German verbs. When we remove these verb endings, the across-category differences look roughly like what one expects by chance (see Figure 12). This result is unsurprising since the presence of verb stems like *-er* means that any given verb is less likely to be a neighbor of a noun since most nouns do *not* end in *-er*. The within-category analysis is qualitatively unchanged by focusing on roots (in all cases the real lexicon is clumpier than expected by chance).

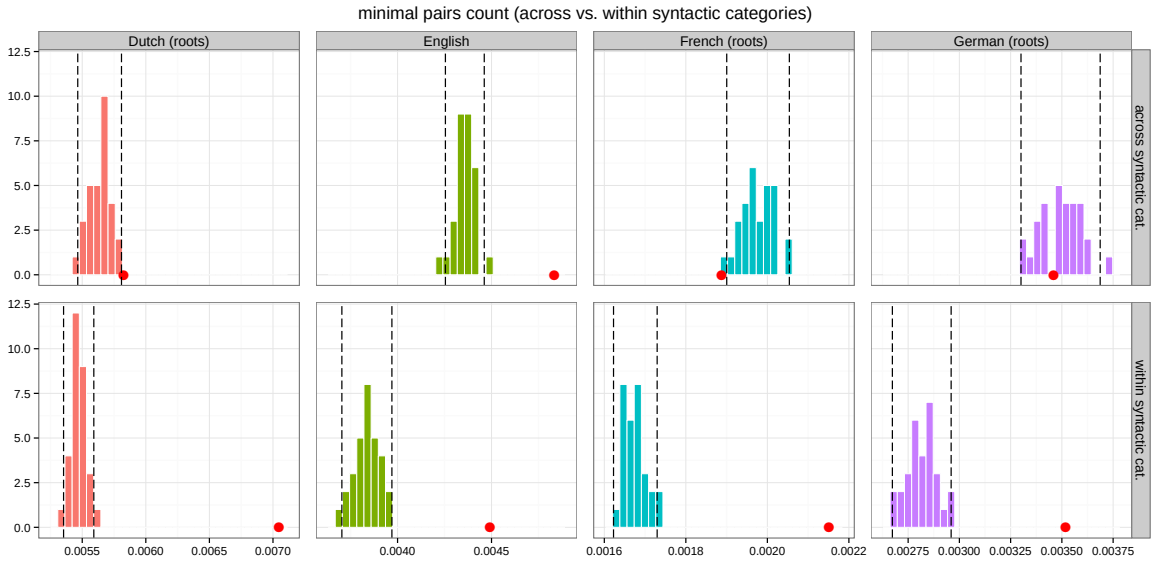

Figure 12: As in Figure 11, these histograms show the distribution of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon, but without infinitive endings on verbs in Dutch, French and German.

Note that the probability of getting a minimal pair within the same syntactic category is greater than the probability of getting a minimal pair across different syntactic categories for Dutch, French and German but not for English (compare the position of the red dots in the graphs from the upper and the lower row, language by language). A possible explanation for this difference is that there is still some verbal morphology present in the lemmas for Dutch, French and German that we could not capture, and this morphology artificially inflates the number of within-category minimal pairs compared to the number of across-category minimal pairs. For instance, in Dutch, verbs of motion systematically display phonaesthemes (typically a schwa followed by a sonorant) that are not analyzed as suffixes. Another possibility for this difference is that the probability of getting a minimal pair across and within syntactic categories may not be directly comparable because the length distributions for within-category words and across-categories words are different and may thus drive part of the difference found here. As a result we prefer to concentrate on the comparison of the real lexicon with

the simulated lexicons, since in this comparison these potential confounds are controlled for.

4.3 Interim summary

To sum up, we did not find evidence that clumpiness is more likely among perceptible than confusable phonological contrasts. That is, it seems that confusable phoneme pairs like 'm' and 'p' are just as likely to be the basis of minimal pairs as less confusable pairs. One possible explanation for this null result is that even highly confusable phoneme pairs like 'b' and 'p' are only confusable in certain specific contexts, such as after vowels at the end of words as in 'cab' and 'cap' (Steriade, 1997). Even then, though, context might be enough to disambiguate the words such that the confusability is not an issue.

We found evidence for more clumpiness within syntactic category than across syntactic categories. Again, this cannot be driven by morphology here as we focused on monomorphemic words. Yet, this may potentially be the consequence of a more general pattern: Words of the same syntactic category may share more phonological properties than words of different classes (Fitneva et al., 2009; Kelly, 1992). For English words, it is also the case that we see more clustering across categories than expected by chance. But that is not the case for French, German, or Dutch when we control for the presence of infinitival markers. Therefore, at least for these languages, it may even be the case that this syntactic category effect drives the larger clumpiness effect observed across the lexicon. This would be consistent with the findings of Monaghan et al. (2014), Tamariz (2008) and Dautriche et al. (submitted), who show a relationship between semantic and phonological similarity across many languages. This is also consistent with the Functional Load Hypothesis (e.g., Hockett, 1967; Martinet, 1952): The Functional Load Hypothesis states that the likelihood of diachronic merger between two sounds depends on the amount of 'work' that the pair does in distinguishing words in the lexicon, i.e., the more minimal pairs distinguished by a phoneme contrast, the less likely that contrast is to merge (Wedel, Kaplan, & Jackson, 2013). Interestingly, the more within-syntactic category minimal pairs distinguished by a phoneme contrast, the less likely that contrast is to merge (Wedel, Jackson, & Kaplan, 2013). This may explain why we found more minimal pairs within syntactic categories than across.

5 General Discussion

We have shown that lexicons use their degrees of freedom in a systematic and interesting way. While we can still characterize the relationship between wordforms and meanings as arbitrary, structure emerges when one considers the relationships within the space of possible wordforms. Across a wide variety of measures of phonological similarity, the real lexicons of natural languages show significantly more clustering than lexicons produced by the best generative model selected by our model comparison procedure (see point 2.3).

Because we focused on monomorphemic words, this effect cannot be a result of words sharing prefixes and suffixes. It is also not a product of any structure captured by sound-to-sound transition probabilities such as phonotactic regularities, since our models capture these patterns.

We started this paper by asking whether lexicons of natural languages show evidence for clumpiness or sparsity above and beyond phonotactics. Yet, phonotactics may be viewed as being itself a source of clumpiness—as a way of constraining the space of words. Because of the phonotactic structure that exists in every language, we thus expect a baseline level of clumpiness. And our models surely capture this clumpiness baseline as a 5-phone model is more constraining than a 4-phone model

and thus allows for fewer possible words. An important question that follows is whether the effect of clumpiness we report here is a by-product of the reduction of the space of possible words caused by phonotactics (but that our null lexicon model is not constrained enough to capture) or whether this reflects phonological regularity beyond the baseline imposed by phonotactics. While separating out the contribution of phonological regularity from the reduction of the space of possible words caused by phonotactics is beyond the scope of this article, it is important to note that it is not the case that a model with increased constraints on the space of possible words *mechanically* predicts that words are more clustered together. Indeed, the 4-phone model is closer to the real lexicon on most wordform similarity measures (see Figure 13) and thus is more clumpy than the 5-phone model yet allows for *more* possible words. Our focus in this paper, using this method, has been thus to show that phonological clustering exists above and beyond the clumpiness effects already inherent to a model that captures sound-to-sound transition probabilities.

One explanation for the clumpiness in the lexicon is shared phonetic properties of semantically related words. Like 'skirt' and 'shirt', many words in the language share deep etymological roots. Moreover, the presence of sound symbolism in the lexicon is another source of structure in the lexicon not captured by our models. For instance, there is a tendency in English for *gl-* words to be associated with light reflectance as in 'glimmer', 'gleam' and 'glisten' (Bergen, 2004; Bloomfield, 1933). There are additionally cross-linguistic correspondences between form and meaning, such as a tendency for words referring to smallness to contain high vowels (Hinton et al., 2006; Sapir, 1929). Interestingly, recent studies show that phonologically similar words tend to be more semantically similar across measures of wordform similarity over many typologically different languages (Dautriche et al., submitted; Monaghan et al., 2014; Tamariz, 2008). This suggests that clumpiness in the lexicon cannot be attributed to small islands of sound symbolism (see in particular Monaghan et al. 2014). Rather, it reveals a fundamental drive for regularity in the lexicon, a drive that conflicts with the pressure for words to be as phonologically distinct as possible.

One other possible source of the lexicon's clumpiness is that speakers may preferentially re-use common articulatory sequences. That is, beyond just phonotactics and physical constraints, speakers may find it easier to articulate sounds that they already know. Recall our example of the language in which there is only one word for a speaker to learn. She would quickly become an expert. Along those lines, the presence of any given sound sequence in the language makes it more likely that the sequence will be re-used in a new word or a new pronunciation of an existing word. In that sense, the lexicon 'overfits': any new word is deeply dependent on the existing words in the lexicon. Note that because our baseline used a lexical generation model, any pressure for re-use must occur over and above the observed statistical trends (e.g., 5-phone sequences) in the language.

Relatedly, lexical clumpiness may be advantageous for some aspects of word production. While words having many neighbors are challenging for word recognition (Luce, 1986; Luce & Pisoni, 1998), they may be easy words to produce (Gahl et al., 2012; Vitevitch, 2002; Vitevitch & Sommers, 2003). Previous studies suggest that listener-oriented models of speech production– where speakers adjust their speech to ensure intelligibility of words that might otherwise be difficult to understand (as could be words with many neighbors)– are limited by attentional demands and working memory in conversational speech (Arnold, 2008; Lane et al., 2006). However, speakers may produce words with many neighbors faster, because they are easier to access and retrieve (Dell & Gordon, 2003; Gahl et al., 2012). Hence a clumpy lexicon would be beneficial for a speaker-oriented model of speech production associated with rapid lexical access and retrieval.

A clumpy lexicon also may allow for easier compression of lexical knowledge. By having words that share many parts, it may be possible to store words more easily. Though we concentrate here on monomorphemic lemmas, these account only for one third of all the lemmas in the lexicon. The

fact that languages re-use words or parts of words in the remaining two thirds of the lemmas shows that re-use of existing phonological material must be important (though in those cases languages are re-using part of the semantic material as well). It may even be the case that, much as morphology allows the productive combination of word parts into novel words, there exist sound sequences below the level of the morpheme that *also* act as productive units of sound.

Phonological proximity may also display some functional advantages in the context of word learning. To form a novel lexical entry in their lexicon, children must be able to extract a word form and associate it to a meaning. In theory, a clumpy lexicon may be advantageous for learning as it reduces the amount of new information that must be represented in the lexicon. For instance, to learn a novel word such as 'blick', children need to create a novel phonological representation /blɪk/ that needs to be associated to a novel semantic representation. Re-using parts of existing phonological forms may be more efficient because it allows children to minimize the amount of phonological information that must be learnt and remembered (see also Storkel & Maekawa, 2005; Storkel et al., 2012).

Despite the fact that one might expect the lexicon to be maximally dispersed for communicative efficiency, these results strongly suggest that the lexicon is not nearly as sparse as it could be–even given various phonetic constraints. Thus, why does communicative efficiency not conflict with clumpiness in the lexicon?

One possibility is that clumpiness does not appear randomly in the lexicon but is organized along dimensions that maximize wordform recoverability. We hypothesized that recoverability could be enhanced if similar wordforms such as minimal pairs were disambiguated by minimally confusable sounds. Our results provide no evidence that the lexicon is less clumpy for confusable sounds than for non-confusable sounds. Relatedly, lexical access might be faster in a lexicon where confusable wordforms span different syntactic categories. Yet we find that, if anything, wordforms are more similar *within* the same syntactic category than what would be expected by chance for all four languages despite the absence of morphology.

Another possibility that would explain why communicative efficiency does not conflict with clumpiness in the lexicon is that contextual information is usually enough to disambiguate words, even when their phonological forms are similar. Therefore, it simply does not matter whether certain words are closer together in phonetic space than they might otherwise be. Piantadosi et al. (2012) showed that lexical ambiguity, such as dozens of meanings for short words like *run*, does not impede communication and in fact promotes it by allowing the re-use of short words. In a similar way, there may be a communicative advantage from having not just identical words re-used, but from re-using words that are merely similar. In all cases, context may be enough to disambiguate the intended meaning and avoid confusion–whether it be confusion between two competing meanings for the same word or confusion between two similar-sounding words.

Likewise, our analysis here concentrated on the phonemic representation of words, ignoring the fact that speech contains a lot of fine phonetic details that listeners could use to disambiguate between words. For instance, pairs of homophones such as 'thyme'/'time' in English can be differentiated based on their duration (Gahl, 2008). Kemps et al. (2005) show that English and Dutch listeners are sensitive to fine-grained durational differences between a base word ('run') and the base word as it occurs in an inflected or derived word ('runner'). Being sensitive to these cues may also be useful to disambiguate between words that sound similar such as minimal pairs.

Clumpy lexicons seem to be advantageous for word production, word learning, and memory, but detrimental for word perception. Yet the interaction of these cognitive and articulatory constraints with a pressure for clumpiness or a pressure for dispersion is complex. Clearly there are many functional pressures at play for the listener, the speaker, and the learner, and they do not individually point towards either clumpiness or distinctiveness of wordforms. In the context of word learning,

wordform similarity may be both advantageous and disadvantageous: Similar-sounding words (1) minimize the amount of information that needs to be stored (e.g., Storkel & Maekawa, 2005); (2) help for word segmentation (Altvater-Mackensen & Mani, 2013); (3) are easier to recognize because they are composed of highly-probable sequences of sounds (e.g., Jusczyk & Luce, 1994); and (4) help children group words into categories (i.e., nouns, verbs) when phonological proximity is aligned with semantic or syntactic classes (Monaghan et al., 2011). Yet when it comes to individual word learning, learners have a hard time learning a novel meaning for a sound string similar to a word they know (e.g., 'tog' a phonological neighbor of the familiar word 'dog'; e.g., Swingley & Aslin 2007) and this disadvantage is even greater when phonological similarity is aligned with syntactic or semantic similarity (Dautriche, Swingley, & Christophe, 2015). The situation is similar in the context of word production: while clumpy lexicons may be easier to produce, they may also give rise to a greater number of speech errors when the relationship between phonological proximity and semantic proximity is high (e.g., Dell & Reich, 1981). Importantly, our results suggest that the functional challenges associated with wordform similarity weigh less than its functional advantages. In other words, the sum of all these functional pressures (for the listener, the speaker, the learner) pushes towards a clumpy lexicon.

Certainly, while we can probably get an idea of the weight of different functional pressures from observing the structure of the lexicon (and of languages more generally), we cannot tell whether they actually explain why lexicons look the way they do. Research looking at language evolution offers a promising venue to understand *how* functional pressures from both language usage and language learning combine to produce the particular pattern of clumpiness observed in human languages. By observing how language is transmitted culturally from one generation to the next, using either computational models or experiments with human participants in the lab, it is possible to isolate how languages are shaped by the processes of both cross-generation transmission (language learning) and within-generation communication (language use) (Kirby et al., 2008; Kirby & Hurford, 2002; Kirby et al., 2015; Smith et al., 2003). The methodology used here, whereby the real lexicon is compared to a distribution of statistically plausible 'null' lexicons, could be used to generate hypotheses about the lexicon and human language more generally, that could be tested experimentally using such language evolution techniques. While much previous work has focused on simply measuring statistical properties of natural language, modern computing power makes it possible to simulate thousands of different languages with different constraints, structures, and biases. By comparing real natural language to a range of simulated possibilities, it is possible to assess which aspects of natural language occur by chance and which exist for a reason.

Of course, we must keep in mind that the present work examines only a small number of European languages. Estimating to what extent this effect generalizes would require a larger number of languages, and we undertake exactly such a project in other studies (Dautriche et al., submitted; Mahowald et al., *submitted*). Specifically, we used a corpus of 100 languages from Wikipedia to show large-scale evidence that a) more frequent words are more orthographically probable and have more minimal pairs than less frequent words; and b) semantically related words are more phonetically similar than less related words. While the Wikipedia corpus does not focus on monomorphemes and is therefore less controlled than the results presented here, it suggests that the clumpiness we observe in the lexicons of Dutch, English, German, and French likely generalizes to other languages as well.

In future work, it may be possible to test increasingly sophisticated models of phonotactics using this methodology. One possibility is that our models of phonotactics are simply not good enough yet to capture the rich structure of natural language. But the results here suggest that any "null" model that can approximate natural languages will need to account not just for the preferred sounds of a language but for the entire space of existing words. That is, the goodness of 'dax' as an English

word depends not only on an underlying model of English sound structure but also on the fact that 'lax' and 'wax' are words, that 'bax' is not, and on countless other properties of the existing lexicon. Another possibility is that our models of phonotactics capture more than the phonotactic constraints of languages (see above for others possible sources of clumpiness). It would be thus informative for future work to separate clearly phonotactic constraints from other sources of regularity to have a more thorough picture on how clumpiness patterns can be interpreted.

Finally, the present work has focused on quantifying wordform similarity in natural language *synchronically*, but it is likely that we may have a lot to learn from *diachronic* data to observe how clumpiness evolve in the lexicon as new words appear in the language. While we discussed the possibility that there are pressures for clumpiness exerting on the lexicon, another possibility is that there are only pressures for dispersion and not clumpiness, but that word coining leads to clumpy initial states.[16] For example, 'flour'/'flower' were originally two senses of a single word, and so pronounced identically. Now that English speakers perceive them as entirely different words, it is plausible that processes of dispersion could act to bring their pronunciations apart, as happened with 'one'/'a(n)' and may be happening in 'thyme'/'time'. This would lead to a lexicon that prefers minimal pairs to avoid lexical ambiguities. Diachronic data may thus shed light onto the mechanism by which clumpiness arises in the lexicon.

Overall, we have shown that lexicons are more richly structured than previously thought. The space of wordforms for Dutch, English, German and French is clumpier than what would be expected by the best chance model, across a wide variety of measures: minimal pairs, average Levenshtein distance and several network properties. The strongest evidence comes from minimal pairs, for which the effect size was quite large. From this, we propose that the clustered nature of the lexicon holds over and above the patterns that are captured by a phonotactic model, suggesting that the pressure for dispersion in lexical systems is a deep drive for regularity and re-use, beyond standard levels of lexical and morphological analysis.

Acknowledgements

We thank Benoit Crabbé, Emmanuel Dupoux and all members of Tedlab, the audience at AMLaP 2014, and the audience at CUNY 2013 for helpful comments. Research reported in this publication was supported by ANR-10-LABX-0087, ANR-10-IDEX-0001-02, ANR-13-APPR-0012, the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number F32HD070544 as well as the post-graduate fellowship from the Fyssen to ID and an NDSEG graduate fellowship and an NSF Doctoral Dissertation Improvement Grant in linguistics to KM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[16] We thank an anonymous reviewer for this suggestion.

References

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(01), 9–41.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, n/a–n/a.

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, *20*(03), 679–685. doi: 10.1142/S021812741002596X

Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, *23*(4), 495–527.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058.

Baayen, R. (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th annual meeting on association for computational linguistics* (pp. 271–278).

Baayen, R. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Baayen, R., Piepenbrock, R., & van H, R. (1993). The celex lexical data base on cd-rom. *n.s*.

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*(3), 339–362.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, *113*, 1001.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311. doi: 10.1353/lan.2004.0056

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417–436. doi: 10.1016/j.jecp.2012.01.005

Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in french 18-month-olds. *Language Learning and Development*, *10*(1), 1–18.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*, 97-138.

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. doi: 10.1016/j.jecp.2004.07.004

Cohen Priva, U. (2008). Using Information Content to Predict Phone Deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics* (p. 90).

Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, *64*(1), 44–56.

Dautriche, I., Fibla, L., & Christophe, A. (2015). *Learning homophones: syntactic and semantic contexts matter.* (40th Boston University Conference on Language Development)

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (submitted). Wordform similarity increases with semantic similarity: an analysis of 101 languages.

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

de Carvlaho, A., He, A., Lidz, J., & Christophe, A. (2015). *18-month-olds use the relationship between prosodic and syntactic structures to constrain the meaning of novel words.* (40th Boston University Conference on Language Development)

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, *6*, 9.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of verbal learning and verbal behavior*, *20*(6), 611–629.

de Saussure, F. (1916). *Course in general linguistics*. Open Court Publishing Company.

Ernestus, M., & Mak, W. M. (2004). Distinctive phonological features differ in relevance for both spoken and written word recognition. *Brain and Language*, *90*(1), 378–392.

Ferrer-i Cancho, R., & Moscoso del Prado Martín, F. (2011, December). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, *2011*(12), L12002. doi: 10.1088/1742-5468/2011/12/L12002

Fitneva, S. A., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, *36*(5), 967–997.

Flemming, E. (2002). *Auditory representations in phonology*. Routledge.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In *Phonetically based phonology (eds. bruce hayes, robert kirchner, donca steriade).* Cambridge: Cambridge University Press.

Gafos, A. I. (2014). *The articulatory basis of locality in phonology*. Routledge.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S. (2015). Lexical competition in vowel articulation revisited: Vowel dispersion in the easy/hard database. *Journal of Phonetics*, *49*, 96–116.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178. Retrieved 2016-11-28, from `http://www.sciencedirect.com/science/article/pii/S0749596X15001503`

Gahl, S., Yao, Y., & Johnson, K. (2012, May). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. doi: 10.1016/j.jml.2011.11.006

Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1216438110

Goldrick, M., & Rapp, B. (2002). A restricted interaction account (ria) of spoken word production: The best of both worlds. *Aphasiology*, *16*(1-2), 20–55.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of finnish vowel harmony. *Natural Language & Linguistic Theory*, *30*(3), 859–896. doi: 10.1007/s11049-012-9169-1

Graff, P. (2012). *Communicative efficiency in the lexicon* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Hayes, B. (2012). *BLICK - a phonotactic probability calculator.*

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440. doi: 10.1162/ling.2008.39.3.379

Hinton, L., Nichols, J., & Ohala, J. J. (2006). *Sound symbolism*. Cambridge University Press.

Hockett, C. (1960). The origin of language. *Scientific American*, *203*(3), 88–96.

Hockett, C. (1967). The quantification of functional load. *Word*, *23*(1-3), 300–320.

Hockett, C., & Voegelin, C. (1955). *A manual of phonology* (Vol. 21) (No. 4). Waverly Press Baltimore, MD.

Howes, D. (1968). Zipf's law and miller's random-monkey model. *The American Journal of Psychology*, *81*(2), 269–272.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298–20130298. doi: 10.1098/rstb.2013.0298

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65. doi: 10.1016/j.cognition.2008.07.015

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Bayesian inference for PCFGs via markov chain monte carlo. In *HLT-NAACL* (pp. 139–146).

Jusczyk, P., & Luce, P. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645. doi: 10.1006/jmla.1994.1030

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364. doi: 10.1037/0033-295X.99.2.349

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

Kemps, R. J., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, H. (2005). Prosodic cues for morphological complexity in dutch and english. *Language and Cognitive Processes*, *20*(1-2), 43–73.

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. Retrieved 2014-08-30, from `http://www.pnas.org/content/105/31/10681.short`

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121–147). Springer.

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

Lane, L. W., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! speakers' control over leaking private information during language production. *Psychological science*, *17*(4), 273–277.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady* (Vol. 10, p. 707).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 839–862. Retrieved 2015-05-07, from `http://www.jstor.org/stable/411991`

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception. Technical Report No. 6.*.

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

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. doi: 10.1080/03640210709336987

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. (*submitted*). Word forms are structured for efficient use.

Mandelbrot, B. (1958). An informational theory of the statistical structure of language. *Communication theory*, 486–502.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge, MA: MIT Press.

Martin, A., & Peperkamp, S. (2015). Asymmetries in the exploitation of phonetic features for word recognition. *The Journal of the Acoustical Society of America*, *137*(4), EL307–EL313.

Martinet, A. (1952). Function, structure, and sound change. *Word*, *8*(1), 1–32.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 311–314.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347. doi: 10.1037/a0022924

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B*.

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

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

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.

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, *4*(02), 115–125. doi: 10.1515/langcog-2012-0007

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*(1), 181–186. doi: 10.1016/j.cognition.2009.04.001

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. doi: 10.1016/j.cognition.2011.10.004

Piantadosi, S., Tily, H., & Gibson, E. (2013). Information content versus word length in natural language: A reply to ferrer-i-cancho and moscoso del prado martin [arXiv: 1209.1751]. *arXiv preprint arXiv:1307.6726*.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Raymond, W. D., Dautricourt, R., & Hume, E. (2006). Word-internal/t, d/deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, *18*(01), 55–97.

Sadat, J., Martin, C. D., Costa, A., & Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive psychology*, *68*, 33–58.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of experimental psychology*, *12*(3), 225.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, *54*(2), 228–264.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 623–656.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 425–440.

Smith, K., Kirby, S., & Brighton, H. (2003, October). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386. Retrieved 2014-06-27, from `http://www.mitpressjournals.org/doi/abs/10.1162/106454603322694825` doi: 10.1162/106454603322694825

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

Steriade, D. (1997). *Phonetics in phonology: The case of laryngeal neutralization.*

Steriade, D. (2001). Directional asymmetries in place assimilation: a perceptual account. In *In hume and johnson.*

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(02). doi: 10.1017/S0142716404001109

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*(02), 291. doi: 10.1017/S030500090800891X

Storkel, H. L., Armbruster, 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., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken american english. *Behavior Research Methods*, *42*(2), 497–506. doi: 10.3758/BRM.42.2.497

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.

Storkel, H. L., Maekawa, J., & Aschenbrenner, A. J. (2012). The Effect of Homonymy on Learning Correctly Articulated Versus Misarticulated Words. *Journal of Speech, Language, and Hearing Research*, *56*(2), 694–707.

Strand, J., Simenstad, A., Cooperman, A., & Rowe, J. (2014). Grammatical context constrains lexical competition in spoken word recognition. *Memory & cognition*, *42*(4), 676–687.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive psychology*, *54*(2), 99. Retrieved 2013-01-26, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613642/

Tamariz, M. (2008, September). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, *3*(2), 259–278. Retrieved 2013-04-16, from http://openurl.ingenta.com/content/xref?genre=article&issn=1871-1340&volume=3&issue=2&spage=259 doi: 10.1075/ml.3.2.05tam

Van Son, R. J., & Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*(1), 100–123.

Viebahn, M. C., Ernestus, M., & McQueen, J. M. (2015). Syntactic predictability in the recognition of carefully and casually produced speech.

Vitevitch, M. S. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1-2), 306–311. doi: 10.1006/brln.1999.2116

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

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of memory and language*, *67*(1), 30–44.

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

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and speech*, 0023830913489096.

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

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n3. *Information and control*, *10*(2), 189–208.

6 Supplemental material

6.1 Robustness of the results

We chose as our baseline a 5-phone model because it performed best on the cross-validation test. But any pattern of clumpiness or dispersion that we find should occur independently of this specific lexical generation model. To check whether our results were robust across the different measures of wordform similarity, we compared the same measures (minimal pairs count, average Levenshtein distance and network measures) obtained in the 3 best models according to our evaluation (see Figure 1): the 5-phone model, the 6-phone model and the 4-phone model.

As shown in Figure 13, we find qualitatively similar results with the 3 best models across all the measures of wordform similarity previously introduced. The 3-phone model behaves somewhat differently and in fact shows more clustering than the 5-phone model. But, because its performance on the held-out data set is poor compared to the models shown here, we do not focus on this model. In general, there were more minimal pairs and lower average Levenshtein distance in the real lexicons than across the three best models. As for the 5-phone model, no conclusive results were obtained for the average clustering coefficient and the giant component measures but the transitivity was higher in the real lexicons than in the three best models of lexicons.

This is evidence that the pattern of clumpiness we found with the 5-phone model is robust across lexical generation models. A pressure for clumpiness is thus visible beyond the particular model of phonotactic probability adopted by the best models produced here.

We also tested whether the German, Dutch, and French infinitival verb endings could be driving clumpiness effects by redoing the analyses above using just root forms (i.e., by removing the infinitival ending from the verbs). One might imagine that, because most verbs end in *-er* in French, for instance, these words have fewer degrees of freedom and thus edit distances will be smaller across the lexicon. In our analysis using just root forms, however, the results were qualitatively the same as when we used lemmas in their infinitive form, likely because the generative models already capture this regularity. That is, our baseline models too have a disproportionate number of words ending in *-er* in French and *-en* in German and Dutch. Because the presence of these infinitival stems does not substantially alter the result, we chose to keep them in the main analysis so as to be consistent with the standard databases we used (CELEX and Lexique).

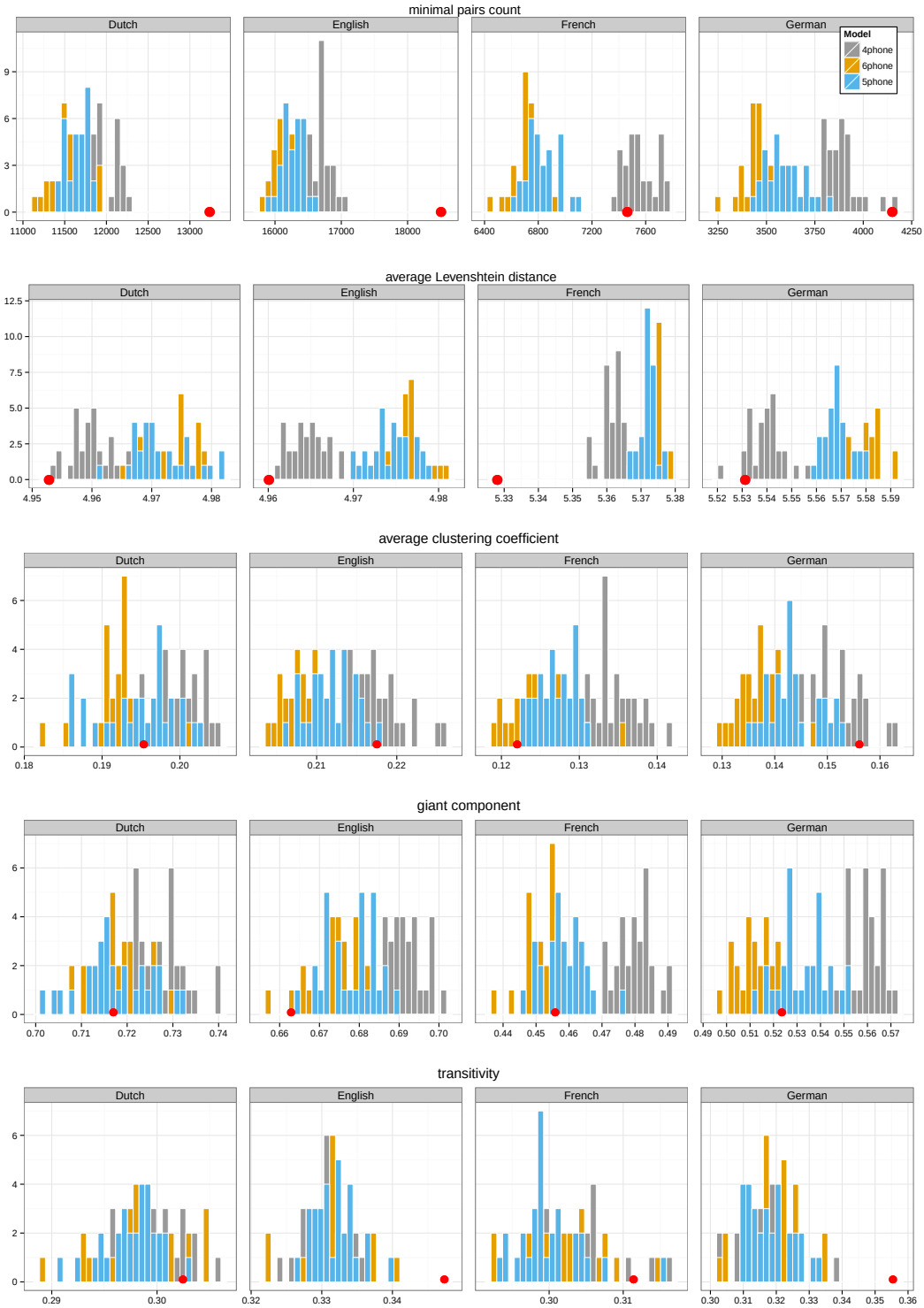

Figure 13: Distributions of a given measure for our best model of word generation (5-phone in blue), our second best model (6-phone in yellow) and our third best model (4-phone in gray) compared to the measure in the real lexicons (the red dots) for the four languages and all the measures reviewed so far.